# Forecasting the Business Cycle using Partial Least Squares

Fredrik Lannsjö

Department of Mathematics,
KTH, Stockholm, Sweden

September 15, 2014

## Abstract

Partial Least Squares is both a regression method and a tool for variable selection, that is especially appropriate for models based on numerous (possibly correlated) variables. While being a well established modeling tool in chemometrics, this thesis adapts PLS to financial data to predict the movements of the business cycle represented by the OECD Composite Leading Indicators. High-dimensional data is used, and a model with automated variable selection through a genetic algorithm is developed to forecast different economic regions with good results in out-of-sample tests.

*Keywords:* Quantitative Forecast, Partial Least Squares, Variable Selection, High-dimensional Regression, Big Data, Business Cycle, Leading Indicators

## Acknowledgements

I would like to thank my mentor Niclas Röken at Consafe Capital Advisors AB for recommending the subject of this thesis and giving me much feedback and valuable guidance. I would also like to thank my supervisor Boualem Djehiche for support and encouragement.

Stockholm, September 2014

Fredrik Lannsjö

# Contents

# Chapter 1

# Introduction

Forecasting the business cycle is a well studied field in finance, and many qualitative and quantitative methods are commonly used. Most techniques frequently use historical data and the study of leading indicators. According to Stock and Watson (2004), academic work of macroeconomic forecasting historically focuses on models with only a handful of indicators, while analysts in business and government often use numerous indicators. Stock and Watson argues that this suggests there is information content in economic data not being fully utilized in the major economic forecasts of today.

One of the major economic forecasts is the Composite Leading Indicator (CLI) published by the Organisation for Economic Co-operation and Development (OECD). It has historically shown a good forecast performance with a lead of 6-9 months of the business cycle. However, a study by Fichtner *et al.* (2011) indicates that the lead for the CLI has decreased in later years. They argue that this is due to the CLI being solely based on domestic indicators, while in an increasingly globalized market, much information about a region's economy can be found in the external environment.

Partial Least Squares (PLS) was developed in the 1970s as a regression tool for analyzing quantitative collinear data in the field of chemometrics. In recent years an extension of the method has been developed for variable selection and statistical classification. This method has shown to often outperform established methods in finding relevant variables for prediction models (Barker *et al.* (2003)).

This thesis aims to take advantage of the general unused quantitative information in financial forecasting (Stock and Watson (2004)), and specifically unused inter-regional information for the CLI (Fichtner *et al.* (2011)), by adapting the latest advancements in PLS to financial analysis. The concrete goal is to use this information to forecast the CLI itself, by using high-dimensional economic data and a PLS regression model with automated variable selection.

Forecasting the CLI means predicting a prediction of the business cycle, which may seem a bit unorthodox. An alternative approach would be to recreate the methods of the CLI with an increased lead, however, this requires many techniques and assumptions that fall outside of the scope of this thesis. The end product with our approach is the same, and a good result would imply the proficiency of PLS for finding and employing currently unused information. In addition, a good prediction of the CLI, say six months ahead, will give us an even greater lead on the business cycle of 12-15 months, via the proven accuracy of the CLI [5].

In the subsequent parts of this thesis, Chapter 2 introduces the data being used while Chapters 3 explains the main regression method under study. Chapter 4 looks at alternative approaches and motivates our choice of regression method. Chapter 5 introduces the discriminant analysis techniques to be examined. Chapter 6 constructs the basics of our modeling methods while Chapter 7 introduces the full model and its evaluation design, as well as the examination of the discrimination techniques. The results of the forecast performance are given in Chapter 8 and their accuracy is discussed in Chapter 9, along with a discussion of the methods and assumptions of the thesis in general. In Chapter 10 the fulfillment of our objectives is evaluated and the validation and further applications of our methods are briefly discussed.

# Chapter 2

# Time series under study

## 2.1 Composite Leading Indicators

The Organisation of Economic Co-operation and Development (OECD) has been publishing the Composite Leading Indicators (CLIs) since 1981. It has been proven to have a good prediction power of the movements of the economy [5]. To quote OECD, the CLI is designed to give "early signals of turning-points in economic activity". As a proxy of economic activity, the monthly Index of Industrial Production (IIP) is used, and the business cycle is defined as the difference between the smoothed IIP data and its long term trend. Since the goal of the CLI is to detect turning-points in the business cycle, about six to nine months ahead, it is not aimed at forecasting certain levels or numerical values, but is rather a dimensionless event forecast, with the turning-points as events. It is based on a selected set of economic indicators and solely on historical data, not including expert judgement. Individual CLIs are available for the member countries of the OECD as well as some non-member economies and zone aggregates.

The regions chosen for this study are the ones with CLIs showing relatively low inter-collinearity, see Table 2.1. Since some of the economic regions featured in OECD's database have strongly collinear CLIs, applying the model on these data can not be seen as independent model validations. Therefore we do not consider regions such as G7 and United States, for example, showing correlations of 0.99 and 0.95 respectively with the OECD's CLI.

| | **Australia** | **Austria** | **Finland** | **Italy** | **OECD** | **OMSNME** | **Four Big Euro** | **Japan** |
|---|---|---|---|---|---|---|---|---|
| **Australia** | 1 | 0.42632 | 0.65498 | 0.36106 | 0.65211 | 0.55862 | 0.49556 | 0.034314 |
| **Austria** | ... | 1 | 0.68253 | 0.85366 | 0.82891 | 0.70106 | 0.93167 | 0.5443 |
| **Finland** | ... | ... | 1 | 0.52888 | 0.7493 | 0.71423 | 0.72833 | 0.25752 |
| **Italy** | ... | ... | ... | 1 | 0.72069 | 0.52619 | 0.89878 | 0.39417 |
| **OECD** | ... | ... | ... | ... | 1 | 0.87361 | 0.90063 | 0.62585 |
| **OMSNME** | ... | ... | ... | ... | ... | 1 | 0.77879 | 0.57956 |
| **Four Big Euro** | ... | ... | ... | ... | ... | ... | 1 | 0.53425 |
| **Japan** | ... | ... | ... | ... | ... | ... | ... | 1 |

Table 2.1: Correlation matrix of the Composite Leading Indicators of the eight economic regions studied between 1990:01 and 2013:12. OECD is the combined economy of the OECD member countries, OMSNME stands for OECD plus Major Six Non-Member Economies, and consist of the 30 OECD countries plus Brazil, China, India, Indonesia, the Russian Federation and South Africa. Four Big Euro is the combined economies of France, Germany, Italy and United Kingdom

## 2.2 Acquiring Data

The data is taken from the package named *Main Economic Indicators - complete database* available at the OECD's iLibrary. This includes national accounts, business surveys, retail sales, production and employment data, interest rates etc., as well as various CLI series. Most indicators are represented in different subsets, e.g. unemployment rates are partitioned into different ages as well as aggregates of these. Further, many are also represented in different measures, e.g. indexed series or growth rate previous year, with or without seasonal adjustments. This data is available for 58 different countries and zone aggregates, including all OECD members as well as six non-members. The complete dataset includes 7882 time series of international economic indicators, although the dates of available data differs between subjects and regions. When selecting the data for our model the time series would preferably not have blank entries. This gives us a trade-off between number of observations of monthly data $N$ and the number of predictors, i.e. economic indicators, $M$. Figure 2.1 shows the available data for the time series by date, from 1980:01 to 2013:12, with blank entries left white.

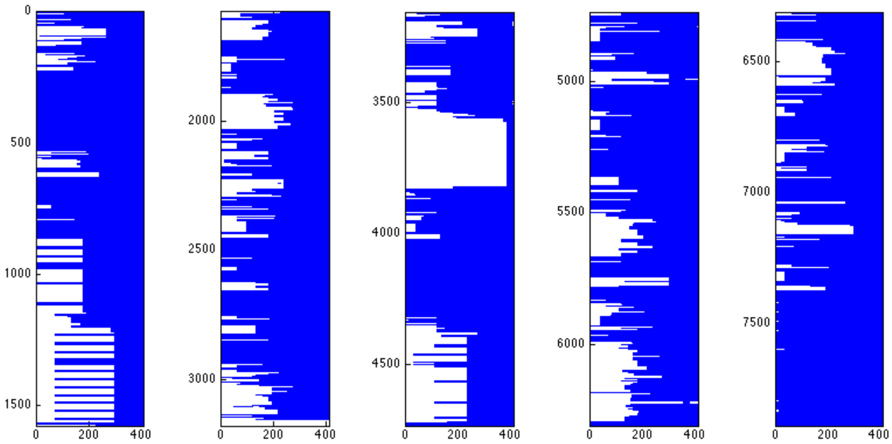


Figure 2.1: Available monthly data from the OECD's Main Economic Indicators dataset marked with blue. The time series of the predictors are arbitrarily lined up along the split y-axis. The x-axes are the monthly time steps from 1980:01 to 2013:12.

By inspection there is an influx of data available from 1990:01, at month 120, for many of the indicators, while some are blank until 2000:01 or even later. Thus we choose to use the indicators with time series containing no blank entries between 1990:01 to 2013:12. This gives us 288 observations of monthly data for 5012 time series, to be evaluated for inclusion in the modeling. The predictor series are not discriminated manually any further, this discrimination is left for the model.

For the CLI we choose the measure named "12 month rate of change of the trend restored CLI". Here trend refers to the long time growth of the economy, and the Rate Of Change at time $t_0$ is given by

$$\text{ROC} = (CLI_{t_0} - CLI_{t_{-12}})/CLI_{t_{-12}},$$

where $t_{-12}$ is the time step twelve months prior to $t_0$. According to OECD the fluctuations of this series are comparable with the growth rate of the turning points of the real Gross Domestic Product (GDP). In short, the interpretation is a positive value means the economy is expanding, while a negative value means it is contracting, and the slope measures at which rate. This version of the CLI is the one most sensitive to the movements of the business cycles, and ideal for a forecast model.

## 2.3 Notation

We will use the convention of writing column vectors in bold-face lowercase form and matrices in bold-face capital letters. To denote a transpose we use "$'$", so that e.g. $\boldsymbol{x'}$ is the row vector of $\boldsymbol{x}$. Scalars resulting from vector multiplication will be included in parenthesis, $(\boldsymbol{x}'\boldsymbol{x})$. The number of elements of a vector or dimensions of a matrix are written in capital letters, while the subscript indices for the corresponding elements have the lowercase version of the same letter. If $\boldsymbol{X}$ is our data matrix with elements $x_{nm}$ for the independent variables as column vectors with observations as rows, this means we have $N$ number of observations and $M$ number of variables.

# Chapter 3

# Partial Least Squares Regression

## 3.1 Historical Background

Partial Least Squares on latent variables (PLS) is a method originally developed for multivariate regression in high dimensional and collinear data by Herman Wold around 1975. It was later improved to better apply to science and technology by Sven Wold and Harald Martens, around 1980 [17]. It has since been a popular regression tool among scientists, particularly in the field of chemometrics (*cf.* [2][4][7][10][12]). Its main features are the ability to deal with strongly collinear data, and using numerous amounts of input variables.

In recent years PLS has found additional applications as a tool for variable selection in statistical discrimination [1][3]. This technique is usually referred to as PLS-DA, for Discriminant Analysis, to distinct from the regression method, referred to as PLSR. Today, PLSR and PLS-DA have been applied to a broad variety of fields, including classifying wastewater pollution, to distinguish coffee beans, classify soy sauce, tumor classification for breast cancer and distinguishing between diagnoses of mental disorder [1][12]. For a longer list and references to these studies see Pérez and Tenenhaus (2003).

The applications of PLS in finance are rare and there is, as far as the author is aware of, no similar published work of PLS for forecasting the business cycle. Therefore some differences between financial data and data from the fields mentioned above, needs be taken into consideration, and these will be discussed throughout this paper.

## 3.2 The data - X and Y

There are some variations on the PLS algorithm for making the multivariate regression of $\boldsymbol{X}$ onto $\boldsymbol{Y}$, and we will focus on the one given by Wold *et al.* (2001) called NIPALS. In this algorithm there can be several dependent variables, i.e. the Y-matrix may consist of any number of column vectors. We will state the general algorithm, see next section, although our model will only be using a single column in the Y-matrix. The X-matrix and the Y-vector will consist of the previously mentioned predictors of the Main Economic Indicators (MEI) and the Composite Leading Indicator (CLI), respectively. Before applying the PLS to a dataset the X-matrix might be scaled and centered, i.e. using the z-score of the column vectors of $\boldsymbol{X}$. The z-score, $z$, of a vector, $v$, is defined by subtracting the mean, $\mu$, of the vector and dividing it by its standard deviation, $\sigma$, as

$$z = \frac{v - \mu}{\sigma}. \tag{3.1}$$

This is not obligatory for PLS to function, and not always desired in chemometric applications [17], thus not part of the algorithm. In our case however it is crucial since the data used includes, for example, GPD measured in trillions of dollars as well as rate of change indicators measured in percent. These different orders of magnitudes will affect the PLS-weights and beta coefficients, to be defined. We are not interested in the magnitude of the variables, but simply their variance and covariance with the Y-matrix. The PLS algorithm constructs a series of variables (weights, scores, loadings etc) as linear combinations of the datasets, this is the latent variables giving the PLS its full name. Similar to a Principal Component Regression, see Section 4.3, the PLS decomposes the X- and Y- matrix to these latent variables as

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}' + \boldsymbol{E}, \tag{3.2}$$

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{C}' + \boldsymbol{G}, \tag{3.3}$$

where $\boldsymbol{T}, \boldsymbol{U}, \boldsymbol{P}$ and $\boldsymbol{C}$ are matrices consisting of the the scores and loadings to be explained in the following section. The remaining terms $\boldsymbol{E}$ and $\boldsymbol{G}$ are error terms, assumed to be independent and identically distributed random variables.

## 3.3 The PLS Algorithm

The zeroth step of the algorithm is finding a first representative for $\boldsymbol{Y}$, a preliminary Y-score vector, $\boldsymbol{u}$. This parameter is later updated, and as a start we use any column of $\boldsymbol{Y}$, in our case the only column, $\boldsymbol{u} := \boldsymbol{y}$. The latent variable matrices are then built one vector at the time through the following projections.

| | | |
|---|---|---|
| (i) | $\boldsymbol{w} = \boldsymbol{X}'\boldsymbol{u}/(\boldsymbol{u}'\boldsymbol{u})$ | X-weights |
| (ii) | $\boldsymbol{t} = \boldsymbol{X}\boldsymbol{w}$ | X-scores |
| (iii) | $\boldsymbol{c} = \boldsymbol{Y}'\boldsymbol{t}/(\boldsymbol{t}'\boldsymbol{t})$ | Y-weights |
| (iv) | $\boldsymbol{u} = \boldsymbol{Y}\boldsymbol{c}/(\boldsymbol{c}'\boldsymbol{c})$ | Y-scores |
| (v) | $\boldsymbol{p} = \boldsymbol{X}'\boldsymbol{t}/(\boldsymbol{t}'\boldsymbol{t})$ | X-loadings |
| (vi) | $\boldsymbol{X} := \boldsymbol{X} - \boldsymbol{t}\boldsymbol{p}'$ | peel off component info. |

This procedure is repeated an arbitrary number of times, $A$, until the desired number of *components*, represented by the vectors $\boldsymbol{t}\boldsymbol{p}'$, have been obtained. These components are (together) approximations of $\boldsymbol{X}$, orthogonal to each other, and contain as much unique variance of $\boldsymbol{X}$ as possible, in descending order of $a = 1, \ldots, A$. After the first component is created, the info it has is "peeled off" from the original X-matrix in step (vi), i.e. the variance it has been given from the data is subtracted from the X-matrix. The matrix with the remaining values are set as the updated X-matrix as the steps (i - vi) are repeated, this time with the Y-scores from step (iv) as $\boldsymbol{u}$.

### 3.3.1 Interpretation

The PLS algorithm gives $A$ vectors for the X-scores, $\boldsymbol{t}_a$. They are estimates of the original X-vectors, by linear combinations with the X-loadings $\boldsymbol{p}_a$, they model $\boldsymbol{X}$ (in element form) as

$$x_{nm} = \sum_a t_{na} p_{nm} + e_{nm}, \tag{3.4}$$

where the $e_{nk}$ are the X-residuals, $\boldsymbol{E}$. The X-scores are also estimates of $\boldsymbol{Y}$ when multiplied with the Y-weights $\boldsymbol{c}_a$

$$y_n = \sum_a c_a t_{na} + f_n, \tag{3.5}$$

with the $f_n$ being the Y-residuals. The scores $\boldsymbol{t}_a$ and $\boldsymbol{u}_a$ contain information about the predictors, how they relate to each other with respect to the model. The weights $\boldsymbol{w}_a$ and $\boldsymbol{c}_a$ can give information about how the scores should be interpreted. They tell us about how the variables combine to form the quantitative relation between the $\boldsymbol{X}$ and $\boldsymbol{Y}$ [17].

Most multivariate regression methods gives the dependent variable as a linear model of the predictors in the form $y = \sum \beta x + e$. To arrive at this representation of the PLS regression, we include the element form of the construction of X-scores from the X-weights

$$t_{na} = \sum_k w_{ka} x_{nk}. \tag{3.6}$$

Now, the PLS representation of $\boldsymbol{Y}$ can be described as

$$y_{nm} = \sum_a c_{ma} \sum_m w_{ma} x_{nk} + f_{nm} \tag{3.7}$$
$$(\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}\boldsymbol{C}' + \boldsymbol{F}). \tag{3.8}$$

Renaming the matrix $\boldsymbol{W}\boldsymbol{C}' = \boldsymbol{B}$ the beta coefficients of the linear model is obtained with the same number of columns as the Y-matrix, giving us the vector $\boldsymbol{\beta}$ in our case. Conclusively, the scores and loading are good at describing the structures and relations in $\boldsymbol{X}$ and $\boldsymbol{Y}$, while the weights combined as the estimated $\hat{\boldsymbol{B}}$ is the basis for predicting new Y-values, $\hat{\boldsymbol{Y}}$, from new X-data, $\boldsymbol{X}_{\text{new}}$, as

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}_{\text{new}} \hat{\boldsymbol{B}}. \tag{3.9}$$

These are the main parts of the method of PLSR modeling, and shows how a dataset can be projected onto an arbitrary number of synthetic components.

The number of components can be chosen as any number between one and the rank of the matrix. There is little reason, or even a hindrance to use more components than necessary. Usually the algorithm is repeated until there is no more information about $Y$ left in $X$ [17]. Wold *et al.* (2001) gives a good example of how to use cross-validation to see when this number is reached, but simply calculating the percentage of variance in $\boldsymbol{X}$ made up by each component is an adequate method [16], see section 6.3

In the following section we will state the mathematical proof of the PLS weights being optimized to include as much of the information of $\boldsymbol{X}$ and $\boldsymbol{Y}$ in as few components as possible. In the current notations this is explained by the first X-weight vector $\boldsymbol{w}_1$ being the first eigenvector of the matrix $\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{Y}'\boldsymbol{X}$, by Wold *et al.* (2001) referred to as the "variance-covariance matrix". For the later components, the $\boldsymbol{w}_a$ is the first eigenvector of the deflated variance-covariance matrix, $\boldsymbol{Z}_a'\boldsymbol{Y}\boldsymbol{Y}'\boldsymbol{Z}_a$ [17]. Thus we get an alternative interpretation of the weights as

$$\boldsymbol{w}_a = \operatorname{eig}\left(\boldsymbol{Z}_a'\boldsymbol{Y}\boldsymbol{Y}'\boldsymbol{Z}_a\right), \qquad \boldsymbol{Z}_a = \boldsymbol{Z}_{a-1} - \boldsymbol{T}_{a-1}\boldsymbol{P}_{a-1}'. \tag{3.10}$$

The eigenvector relationship among the weights grants them the linear independence property, as their span forms an orthogonal set. This is the original algorithmically defined PLS, and the interpretations of its properties, as formulated by Wold *et al.* (1985). The weights obtained can be seen as the principal components of the empirical covariance matrix between $\boldsymbol{X}$ and $\boldsymbol{Y}$. In regular principal component analysis, the eigenvalues of the components corresponds to the variation of the $X$-matrix, while for PLS, the $\boldsymbol{w}$ instead corresponds to the maximum covariance of $X$ and $Y$. This is not a formal proof of the properties of PLS, but it gives a hint of their origin, and will serve as a link between the following mathematically defined theorem, equation (3.13), and the above mentioned algorithm.

## 3.4 Mathematical Background

Before recent years, PLS has not been given rigorous a mathematical definition, and up until the work of Delaigle and Hall (2011), PLS has never been given an analytical proof of its properties. This definitions and the main theorem will be presented here in relation to our algorithmically defined PLS.

Let $\{(\boldsymbol{X}_1, \boldsymbol{Y}_2), \dots, (\boldsymbol{X}_n, \boldsymbol{Y}_n)\}$ be a set of samples of independent data pairs, distributed as $(\boldsymbol{X}, Y)$. Here $Y$ is a real-valued random variable and $\boldsymbol{X} = (X_t)_{t\in[0,T]}$ is a random function that takes values in the Hilbert space of square integrable functions, say $H = L^2([0,T])$, where $[0,T]$ is a compact interval of $\mathbb{R}$. In the following, we denote by $\langle \phi, \vartheta \ \rangle = \int_0^T \phi(t)\vartheta(t)\,\mathrm{d}t$ the usual inner product in $H$ and by $\|\phi\|$ the induced norm. Further, $X_t$ satisfies $\int_0^T \mathrm{E}[X_t^2]\,\mathrm{d}t < \infty$, and $Y$ is generated by the linear model

$$Y = a + \int_0^T b(t)X_t\,\mathrm{d}t + \epsilon,$$

where $a$ is a scalar parameter, $\epsilon$ is a scalar random variable with finite mean square and $\mathrm{E}[\epsilon|X_t] = 0$, while $b(t)$ is a deterministic square integrable function on $[0, T]$. Stated this way, predicting $Y$ given $X$ means estimating

$$g(x) = \mathrm{E}[Y|X_t = x] = a + \int_0^T b(t)x \,\mathrm{d}t,$$

by estimating $a$ and $b(t)$ from observed data. A general approach is to express $X_t$ and $b(t)$ in terms of an orthogonal basis $\psi_1(t), \psi_2(t), \ldots$ defined on $[0, T]$. Expansions for $X_t$ and $b(t)$ in this basis can be written as

$$\begin{aligned} X_t &= \sum_j \left( \int_0^T X_t \psi_j(t) \,\mathrm{d}t \right) \psi_j(t), \\ b(t) &= \sum_j v_j \psi_j(t), \qquad\qquad v_j = \int_0^T b(t)\psi_j(t) \,\mathrm{d}t. \end{aligned}$$

In practice, we have to use a finite integer number of terms $p \geq 1$, and $b(t)$ is approximated by the sum of $p$ terms, estimated from the data. Note that $\int_0^T b(t)X_t \,\mathrm{d}t = \sum_j v_j \int_0^T X_t\psi_j(t) \,\mathrm{d}t$ for, possibly, an infinite amount of terms, which motivates us to take

$$a = \mathrm{E}[Y] - \int_0^T b(t)\mathrm{E}[X_t] \,\mathrm{d}t,$$

and define $\beta_1, \ldots, \beta_p$ to be the finite sequence $v_1, \ldots, v_p$ that minimizes

$$s_p(v_1, \ldots, v_p) = \mathrm{E}\left\{ \int_0^T b(t)(X_t - \mathrm{E}X_t) \,\mathrm{d}t - \sum_{j=1}^p v_j \int_0^T (X_t - \mathrm{E}X_t)\psi_j(t) \,\mathrm{d}t \right\}^2.$$

In terms of the algorithmically defined PLS, or indeed other multivariate regression models, this step represents finding the beta coefficients $\boldsymbol{B}$ that minimizes the residuals $f_i$ of the prediction in (3.7), with $m = 1$, and (3.8). That is, the residual sum of squares

$$S(\boldsymbol{B}) = \sum_n f_n^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}).$$

The functions

$$b_p(t) = \sum_{j}^{p} \beta_j \psi_j(t),$$

$$g_p(x) = \mathrm{E}[Y] + \int_0^T b_p(t)(x - \mathrm{E}X_t)\,\mathrm{d}t = \mathrm{E}[Y] + \sum_{j=1}^{p} \beta_j \int_0^T (x - \mathrm{E}X_t)\psi_j(t)\,\mathrm{d}t,$$

are approximations to $b(t)$ and $g(x)$ and their accuracy depends on how the sequence $\psi_1(t), \psi_2(t), \ldots$ is chosen. These can be chosen as explicit functions, e.g. polynomial or trigonometric, which have the advantage that the functions are known, but there is no reason why they should, in our case, capture the most variance possible in the first $p$ terms.

PLS was invented to not only capture the most variance in the first $p$ terms, but also to capture the most covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ [17]. This is achieved by choosing $\psi_p(t)$ (corresponding to $\boldsymbol{w}_a$ in Section 3.3) in a sequential manner to maximize the covariance functional, $f_p(\psi_p)$. A rigorous mathematical definition of this objective can now be given as

$$\text{maximize } f_p(\psi_p(t)) = \mathrm{cov}\left\{Y - g_{p-1}(X_t), \int_0^T X_t \psi_p(t)\,\mathrm{d}t\right\} \tag{3.11}$$

$$\text{subject to } \int_0^T \int_0^T \psi_j(s) K(s,t) \psi_p(t)\,\mathrm{d}s\,\mathrm{d}t = 0, \qquad 1 \le j \le p-1 \tag{3.12}$$

$$\|\psi_p\| = 1,$$

given that $\psi_1(t), \ldots, \psi_{p-1}(t)$ have already been chosen, and where $K(s,t) = \mathrm{cov}\{X_s, X_t\}$ is the covariance function for the random function $X_t$. In short, the covariance is maximized by using the basis constructed by the eigenfunctions of the linear transformation that takes $\psi$ to $K(\psi)$, given by $K(\psi)(t) = \int_0^T \psi(s)K(s,t)\,\mathrm{d}s$. For further details on the specifics of this basis and its empirical representation, see Delaigle and Hall (2012). To show that PLS in fact optimizes the covariance between predictors and the modeled variable $Y$, Delaigle and Hall (2012) gives the following theorem, along with the proof

**Theorem 3.4** *If* $\int_0^T \mathrm{E}[X_t^2]\,\mathrm{d}t < \infty\,\mathrm{d}t$ *then the function* $\psi_p(t)$ *that maximizes* $f$ *at (3.11), given* $\psi_1(t), \ldots, \psi_{p-1}(t)$ *and subject to (3.12), is determined by*

$$\psi_p(t) = c_0 \left[ K \left\{ b(t) - \sum_{j=1}^{p-1} \left( \int_0^T b(s)\psi_j(s)\,\mathrm{d}s \right) \psi_j(t) \right\} + \sum_{k=1}^{p-1} c_k \psi_k(t) \right] \tag{3.13}$$

*where, for* $1 \le k \le p-1$ *the constants* $c_k$ *are obtained by solving the linear system of* $p-1$ *equations*

$$\int_0^T \int_0^T \psi_j(s)\psi_p(t)K(s,t)\,\mathrm{d}s\,\mathrm{d}t = 0, \qquad j = 1, \ldots, p-1, \tag{3.14}$$

*and where* $c_0$ *is defined uniquely, up to a sign change, by the property*

$$\|\psi_p\| = 1.$$

To prove this theorem, we recall the property of the covariance, $\sigma$ for constants $a, b, c, d$ and random variables $x, y, z, w$

$$\sigma(ax + by, cz + dw) = ac\sigma(x, z) + ad\sigma(x, w) + bc\sigma(y, z) + bd\sigma(y, w) \tag{3.15}$$

*Proof.* The right hand side of (3.11) can along with (3.15) be written as

$$\begin{aligned} &\operatorname{cov}\left\{ \left( \int_0^T b(t)X_t\,\mathrm{d}t \right) - \sum_{j=1}^{p-1} \left( \int_0^T b(t)\psi_j(t)\,\mathrm{d}t \right) \left( \int_0^T X_t\psi_j(t)\,\mathrm{d}t \right), \int_0^T X_t\psi_p(t)\,\mathrm{d}t \right\} \\ &= \int_0^T \int_0^T b(s)\psi_p(t)K\,\mathrm{d}s\,\mathrm{d}t - \sum_{j=1}^{p-1} \left( \int_0^T b(t)\psi_j(t)\,\mathrm{d}t \right) \left( \int_0^T \int_0^T \psi_j(s)\psi_p(t)K\,\mathrm{d}s\,\mathrm{d}t \right). \end{aligned}$$

Taking the partial derivative of this expression with respect to $\psi_p$, yields

$$K\left\{b(t)-\sum_{j=1}^{p-1}\left(\int_0^T b(s)\psi_j(s)\,\mathrm{d}s\right)\psi_j(t)\right\}.$$

The equation in $c_k$ at (3.14) is the result of adjoining Lagrange multipliers on the right-hand side so as to accommodate the first $p-1$ constraints in (3.12). The factors $c_0$ on the right-hand side of (3.13) accommodates the last constraint in (3.12). □

# Chapter 4

# Alternative Methods

Many techniques are available for a multivariate regression model, and the arguably most generic one is the Ordinary Least Squares regression. Although used in similar fields, it is not as closely related to Partial Least Squares as their names might suggest. A very close relative to PLS is rather the technique called Principal Component Regression. This is often considered as an alternative to PLS by scientists. We will present the basic concepts of these two alternatives and argue for our technique of choice.

## 4.1 Ordinary Least Squares

A benchmark method to model dependency structures in science and technology is with the multivariate linear regression approach known as Ordinary Least Squares (OLS) [14]. OLS aims to minimize the error terms $\epsilon$ in the equation

$$y = \beta_0 + \sum_i \beta_i x_i + \epsilon, \tag{4.1}$$

where $y$ is the variable to be modeled as linearly dependent of the $x$-variables the predictors. The $\beta_0$ is the intercept and the $\beta_i, i \geq 1$ are the coefficients to be estimated for each independent $x$-variable. If these variables are stored as before in an $N \times M$ matrix, $\boldsymbol{X}$, the estimated beta coefficients are given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{4.2}$$

These beta coefficients are then used to estimate $y$ for new $x$-data with the model

$$\hat{y} = \hat{\beta}_0 + \sum_n \hat{\beta}_n x_n^{\text{new}}. \tag{4.3}$$

The OLS and PLS both end up at similar expressions for predicting new y-values, but from rather different ways. In order to relate them to each other, it can be worth noting that in the special case when the $\boldsymbol{X}$-matrix is diagonal, and the $\boldsymbol{Y}$-matrix consist of a single column vector, the PLS algorithm for just one component gives the same beta coefficients as the OLS. This have led to some calling PLS a more general version than OLS [17]. This special case is however not realistic to be found in a real world application, since a diagonal X-matrix would only have one non-zero observation per predictor.

Although OLS is a very common way to create statistical models it faces several difficulties and puts high demands on the dataset. When numerous amounts of predictors are used, especially in relation to the number of observations, an OLS model will likely suffer from overfitting. This occurs when the model is too adjusted to the training data (and its noise) to give good predictions.

When the predictors are strongly correlated, OLS faces a problem called multicollinearity. From linear algebra we know that for a matrix to be invertible, the column vectors cannot be linearly dependent. Since the OLS algorithm includes the inverse of the $\boldsymbol{X}'\boldsymbol{X}$-matrix, multicollinearity can make the computations difficult and inaccurate in the sense of unreliable regression coefficients. Since collinear predictors have similar slope, the regression does not "know" how to value them individually. This leads to bad predictability, especially in combination with other obstacles such as over fitting.

## 4.2 PLS vs OLS

Two of the biggest obstacles for a model based on OLS regression are very present in our case, namely multicollinearity and overfitting. The MEI dataset offers a number of predictors in the thousands, while monthly data observations are in the hundreds. In addition, the predictors of the MEI dataset are not only strongly correlated, but in some cases actual aggregations of each other, e.g. the total GDP of the combined G7 region is one of the available predictors, as are each of the included G7 countries individual GDPs. With such a dataset OLS can not be considered as a modeling method.

The PLS was originally developed to counter these problems in quantitative strongly collinear data [8], and deals with them promptly. Since the information from the predictors are regressed onto an arbitrarily chosen number of synthetic components, overfitting need not be a problem, and since these components themselves are orthogonal, there cannot be any multicollinearity. These properties extends to the regression coefficients via the regression in Equation 3.8.

The singularity problem in OLS arising from an non-invertible $\boldsymbol{X}'\boldsymbol{X}$ matrix in Equation 4.2 is also bypassed with PLS. In PLS there is no need for the inverse of $\boldsymbol{X}'\boldsymbol{X}$ since $\boldsymbol{Y}$ is regressed on the X-scores instead of $\boldsymbol{X}$ itself [8][7].

## 4.3 Principal Component Regression

Principal Component Regression (PCR) is closely related to PLS [17]. It also uses latent variables constructed to have maximum information from the dataset in the first $A$ components, rather than creating one component per X-column, as in OLS. Thus PCR can overcome similar obstacles as the PLS, namely multicollinearity and overfitting. Many papers discuss the difference between PLS and PCR in detail. In PCR, the scores are created from $\boldsymbol{X}$ alone, while in PLS the scores use both information from the X- and the Y-matrix. While PCR focus on describing the variance in $\boldsymbol{X}$, PLS focus on describing the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ [8].

For a good summary of literature comparing PLS with PCR, Wentzell and Vega (2003) lists the conclusions of 27 articles. The general conclusions are either in favor of PLS as a prediction method, or that they are both potent with no clear advantage for either method. A study showing PCR as strictly advantageous does, to the knowledge of the author, not exist, which is enough to only consider PLS methods for our model.

Wentzell and Vega (2003) completes their research by making their own study of the subject, in the form of a spectroscopy experiment, and sides with the view that none of the methods is more advantageous. However, a factor not as thoroughly considered in their study is a very large number of predictors in relation to observations. A later study by Boulesteix and Stimmer (2006) focus on datasets with very numerous predictors, and find the PLS superior.

Further our goal is not only to use PLS as a regression tool, but as a discrimination method as well. For this purpose [1] shows that the PLS is the preferred choice.

# Chapter 5

# Variable Selection

Variable selection and statistical classification is a very important issue in scientific engineering and statistical modeling. Identifying and discarding redundant variables from the observation dataset is essential for making accurate predictions. Although PLS was originally designed as a regression method, applied scientists have recently started using it as a tool for statistical classification and Discrimination Analysis (DA) ([1][3]). The advantage of PLS-DA for variable selection is that the beneficial properties of the regression can be transferred to validation methods of the variables, or predictors. While less sophisticated methods may only look at one variable's correlation to $Y$, PLS-DA uses information about between-group variation as well, i.e. their covariance with both $Y$-data and other $X$-variables. Chong *et al.* (2005) studies four different methods for discrimination and conclude that the ones named PLS-VIP and PLS-Beta are the most prominent. Since neither has a clear advantage over the other in every situation [3], we will make one version of the automated variable selection for each method.

The basic setup for creating a model using PLS-DA is two fold, including one regression step for the variable selection, and one regression step for the actual model creation. Initially a PLS regression is made on a given dataset, $\boldsymbol{X}$ and $\boldsymbol{Y}$, followed by calculating a *discriminant function* for each predictor of $\boldsymbol{X}$, constructed from the latent variables of the regression. The discriminant function, or Selection Parameter, is then used to classify the predictors as significant or redundant for modeling $\boldsymbol{Y}$. Once every predictor has been classified, the redundant predictors are discarded and a reduced version of $\boldsymbol{X}$ is created using exclusively the significant predictors. Now the regression can be repeated to create the actual model of $\boldsymbol{Y}$.

## 5.1 PLS-VIP

The Variable Influence on Projection (VIP) was originally defined by Wold in 1993 [3]. This parameter is the discriminant function of the PLS-VIP method for selecting and classifying variables. The $\text{VIP}_m$ for the $m$-th predictor is calculated as

$$\text{VIP}_m = \sqrt{M\sum_{k=1}^{A}\Big(SS(\beta_k \boldsymbol{t}_k)\big(\boldsymbol{w}_{jk}/\|\boldsymbol{w}_k\|\big)^2\Big)/\sum_{k=1}^{A}SS(\beta_k \boldsymbol{t}_k)}.$$

The $SS(\beta_k \boldsymbol{t}_k) = (\beta_k^2 \boldsymbol{t}_k' \boldsymbol{t}_k)$ is the regression sum of squares, and is proportional to the squared correlation between $\beta_k$ and $\boldsymbol{t}_k$, thus it explains the amount of covariance modeled in these variables. The $w_{jk}$ measures the contribution of each variable $m$ to the $k$-th component. Thus, VIP quantifies the influence on the response of each variable summed over all components, relative to the models total sum of squares [12].

For a predictor to be significant the "greater than one rule" is generally used as a criterion, since the mean of all VIPs is one [3][6]. That is, if $\text{VIP}_j > \eta = 1$, the $j$-th predictor is significant for modeling Y, and if not, it is removed from the X-matrix and not used in the regression model. However others argue that $\eta = 0.8$ is a good general rule and $\eta = 2$ for large $K$ [12], and [6] studies PLS-VIP down to $\eta = 0.6$. We will let our model find the optimal cutoff value $\eta$, se Section 7.2.

## 5.2 PLS-Beta

This variable selection method is based on the study of the beta coefficients, $\beta$, obtained from the regression. In its simplest form, if the absolute value of the beta coefficient corresponding to a certain predictor is large enough, $|\beta_m| > \mu$, the predictor is selected for the model. The cutoff value $\mu$, however, has not been as researched as the $\eta$ of the PLS-VIP, and features no general rules or representations. Some versions of PLS-Beta does not even feature this cutoff value, but use general methods from statistical discrimination such as Mallow's $C_p$ [3] or running numerous simulations and choosing the optimal value from the plot [6]. This is motivated by the beta coefficients being interpreted as similar to the regression coefficients of an OLS.

One way to implement the PLS-Beta method given by Fujiwara *et al.* (2012) employs the vector $\beta_{\text{select}}$, which consists of the beta coefficients of the selected variables. The input variables for the vector are selected in descending order until a certain threshold is met

$$\frac{\|\boldsymbol{\beta}_{\text{select}}\|}{\|\boldsymbol{\beta}_{\text{all}}\|} > \mu_f, \qquad\qquad 0 < \mu_f \leq 1,$$

where $\boldsymbol{\beta}_{\text{all}}$ is the full beta vector, and $\mu_f$ is the threshold parameter.

Inspired by Fujiwara *et al.* (2012) we construct a proportional, but perhaps more illustrative, implementation of our own. As a cutoff we choose a proportion of the average value of the magnitudes of the coefficients, to give the significance criteria

$$|\beta_m| > \mu \frac{1}{M} \sum_{i=1}^{M} |\beta_i|.$$

The parameter $\mu$ will be referred to as the cutoff value, and the numerical values will be examined in Section 7.2. Reasons for not employing the exact PLS-Beta method of Fujiwara *et al.* (2012) is that we want something that can be related to the PLS-VIP method, and more importantly, combined with this method in order to create an alternative PLS-DA approach, see next section. An example of the selection parameters can be seen in Figure 5.1.

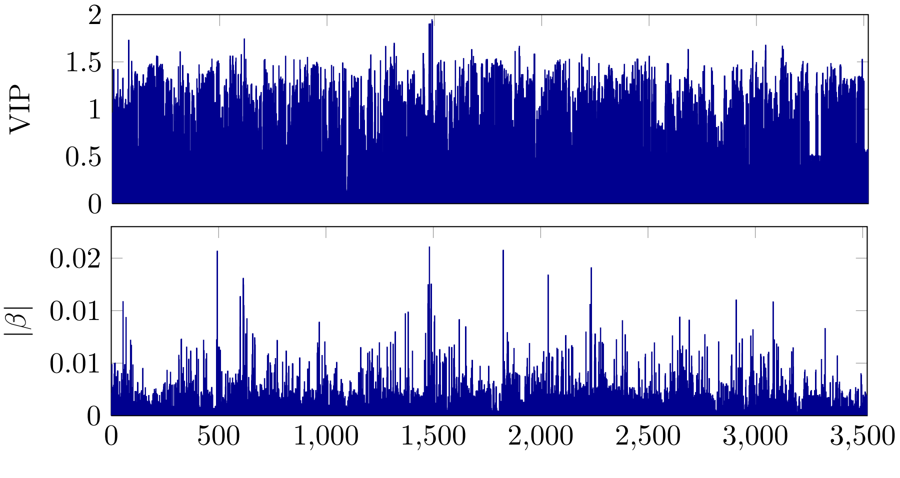


Figure 5.1: Selection parameter values, y-axis, from a typical regression of the basic model. The predictors are arbitrarily indexed along the x-axis. *Top:* The values of the VIP. *Bottom:* The absolute values of the $\beta$.

## 5.3 PLS-VIP-Beta

In their conclusion, Chong *et al.* (2005) suggest that the two above mentioned PLS-DA methods might be combined in order to create an even more advantageous method, which is something early works of Wold have suggested as well. This is never examined by neither author, and we will therefore, without further academic guidance, include this combined variable selection method, referred to as PLS-VIP-Beta, in a straight forward interpretation

$$\boldsymbol{X}_{\text{select}} = \{\boldsymbol{x}_m : \text{VIP}_m > \eta\} \cap \{\boldsymbol{x}_m : \beta_m > \mu\}.$$

If the predictor's VIP value and $\beta$ value are both larger than their respective cutoff value, the predictor is included in the dataset $\boldsymbol{X}_{\text{select}}$ for the model.

# Chapter 6

# The Model

To be able to predict the future movements of the CLI, the model will use a regression onto time lagged predictors. Modeling the CLI with data available some time ago will let us forecast the future CLI when using the latest data available today. The trick is to find Leading Indicators, and this is the initial step of creating our model. A later step is to select only the most relevant predictors for the model, and for this goal we will employ the three different PLS-DA methods of Chapter 5.

## 6.1 Leading Indicators

A leading economic indicator is a variable whose movement is correlated to the business cycle, but changes prior to the business cycle itself, and can therefore be used to predict the economy. In our case we want to find the indicators, i.e. predictors of the X-matrix, $X(t)$ for time step $t$, that are leading indicators to the CLI, our $Y(t)$. For this reason we introduce the correlation function $C(s,t) = \mathrm{corr}(Y(t), X(s))$, for time steps $s$ and $t$. We calculate $C(s,t)$ by measuring the correlation of $Y$ with each predictor $X$ for different time misplacements, or lag, of the $X(t)$. This is done for the full set of observations for each column vector $\boldsymbol{x}_m$, representing $X(s)$, by adding different lag $s$ up to 24 months, as $X(t+s), s = 0, -1, \ldots, -24$, while keeping $Y(t)$ fixed at $t = 0$. We choose to specify this study to forecasting six month in advance, and will thus make the regression model on data available in the MEI dataset six or more months prior to the forecasted CLI.

The empirical correlation function for fixed $Y$ is then given by

$$\hat{C}(s) = \operatorname{corr}\{(y_{25}, y_{26}, \ldots, y_n), (x_{25+s}, x_{26+s}, \ldots, x_{n+s})\}, \quad s = -24, -23, \ldots, 0$$

where the first 24 values of the Y-matrix is discarded since both time series must have the same length. A predictor is regarded as a leading indicator if its correlation with $Y$ is stronger when lagged, i.e. $\hat{C}(s) > \hat{C}(0)$ when $s < 0$.

The lead of an indicator, or predictor, $s_m$, is defined as the number of months ahead, $s$, giving the maximum absolute value of the correlation function; in our case

$$s_m = \arg\max_s\{|\hat{C}(s)|\}, \qquad\qquad -24 \leq s \leq -6.$$

Predictors with a lead less than six month are discarded, since they are not leading predictors and thus not suited for modeling the CLI. We choose the maximum possible lead to be 24 months, since more might lead to interference with a prior business cycle, as these have been known to be as short as only two years in some cases [9].

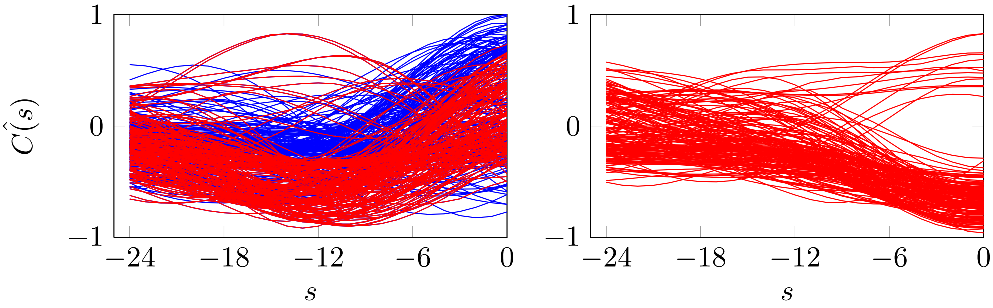


Figure 6.1: Correlation as a function of time lag, $\hat{C}(s)$, between a subset of the Main Economic Indicators and the Composite Leading Indicator. *Left:* The correlation of indicators with relevant lead on the CLI shown in red, and non-leading in blue. *Right:* The correlation of leading indicators after being lagged individually, now exhibiting their peak in correlation in sync with the present time step of the CLI.

Not all leading predictors will have their optimal value for $C(s, t)$ with exactly a six month lag, as Figure 6.1 illustrates. Therefore we re-order the $X$-matrix with each predictor lagged proportional to its lead

$$\boldsymbol{X}^{lag} = \{X^m(24 - s_m), X^m(25 - s_m), \ldots, X^m(n - s_m)\}_{m=1}^{M}.$$

As the MEI-matrix is reordered with every indicator now synced at its maximum correlation with the CLI, the last six observations are saved for making the prediction, i.e. the forecast. The preceding observations are used to make up the updated, now leading, X-matrix for being used in the regression model.

To the right in Figure 6.1 the correlation function for a subset of leading predictors are shown, now perfectly lined up in correlation peak after being lagged. In Figure 6.1 only about one thousand of the available MEI predictors are examined, in order to give a not too overcrowded plot. In a typical run of our model, about 3000-3500 of the original 5012 predictors meets the criteria of leading indicators for the different regions. Note that in this step no discrimination is done based on the actual magnitude of $\hat{C}(s_m)$, but rather the temporal value of $s_m$. The significance of the predictors is not evaluated through their correlation, but rather their cutoff values discussed in Section 7.1.

## 6.2 Basic Model

As the subset of lagged leading predictors of the MEI dataset, $\boldsymbol{X}$, is regressed onto the CLI, $\boldsymbol{Y}$, the beta coefficients, $\hat{\boldsymbol{B}}$, for the specific time interval are obtained through the PLS-algorithm described in Section 3.3. The previously excluded last six entries of the monthly data, $\boldsymbol{X}_{\text{new}}$, can now be projected on the coefficients to obtain the six month forecast $\hat{\boldsymbol{Y}}$ through

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}_{\text{new}}\hat{\boldsymbol{B}}.$$

This is the basic version of the forecast model, and shows how data known at the present time can be used to forecast unknown data of the future. The result of an arbitrarily chosen time interval for the CLI of the OECD region can be seen in Figure 6.2. The fitted curve in green is $\boldsymbol{X}\hat{\boldsymbol{B}}$ and shows a close fit to the given $\boldsymbol{Y}$-data, which is to be expected, see Section 6.3. The forecasted values in red is a very good prediction of the movements of the CLI and exhibits a correlation of 0.96 with the actual six month period. We are not looking to predict the exact values of the CLI, since they have little meaning of their own, but rather to catch the turning points and directions they form for the next six months. Although this particular run of the model shows very good results in this regard, this is no guarantee of a reliable forecasting method for any given time or region. Indeed, several runs of the basic model at different points in time assures us that this was one of the luckier test runs. There are many parameters and variables to be

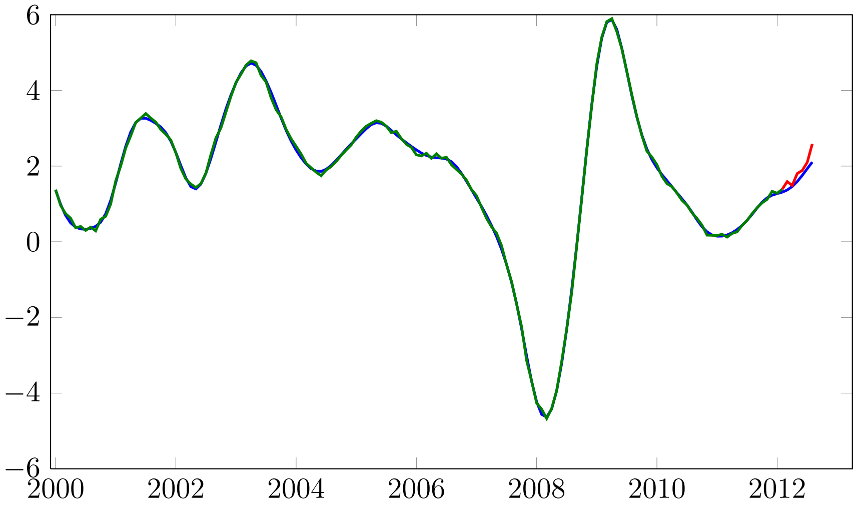


Figure 6.2: The CLI for the OECD region in blue along with the fitted regression in green. The last six entries in red constitutes the forecasted curve.

examined before a reliable model can be proposed. In the following sections all the important parameters for a forecasting model based on PLS will be presented and a more universal Full Model will be proposed, valid for many regions and various points in time.

## 6.3 Goodness of Fit

When evaluating regression models it is common to look at the goodness of fit of the regression, represented by e.g. the Root Mean Square Error (RMSE) between the fitted variable and the actual dependent variable. For PLS regression however, this is not a good measure for how well the predictions for new Y-values will be [12], which is our only concern.

With the number of predictors of the PLS algorithm, $M$, being in the thousands we can always choose our number of components $A$ large enough to get a perfect fit for the regression. But as in the case with ordinary linear regression, this would lead to over fitting, and poor forecasting properties for new data. Figure 6.3 illustrates this by showing the curve of the CLI along with the fitted model.

Instead of using goodness of fit for the regression, we will use the correlation, $\rho$, and the RMSE, of the predicted values i.e. our six month forecast, and the actual $Y$-values for this period. The RMSE of the forecast is given by

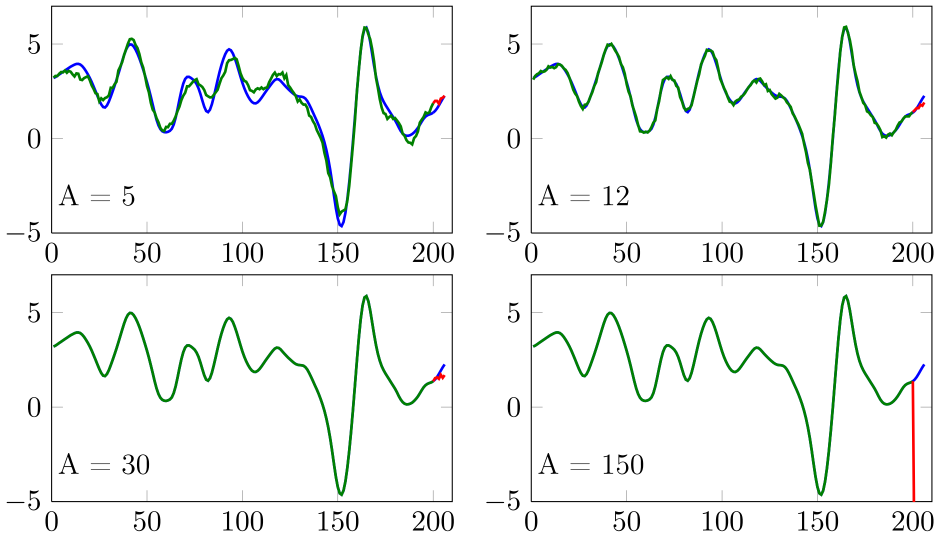


Figure 6.3: A demonstration of the trade-off between goodness of fit and accurate predictions. The basic model of the OECD CLI is applied to an arbitrary time period with different number of components, $A$, for the PLS algorithm.

$$\text{RMSE} = \sqrt{\frac{\sum_{t=n}^{n+h}(\hat{y}_t - y_t)}{h}},$$

where $\hat{y}_t$ and $y_t$ is the forecasted value and actual value at time $t$, respectively. The number of predicted monthly values is $h$ (in our case six), and the regression is made on the first $n - h$ observations of the data. Thus the forecast starts at the $n$-th time step and the RMSE evaluates the performance of the prediction or forecast only, not the fitting of the observed data.

Deciding the number of components can give a trade-off between a good fit for the training data, and an accurate forecast [13]. As can be seen in Figure 6.3, the results depend on this parameter being within "reasonable" bounds. For five components we get large deviations from the actual value in the regression, and cannot expect the forecast to be any better than this fitting. With 150 components we get a close to perfect fit, but unusable regression coefficients, giving a forecast with values three order of magnitudes above the CLI's range, and a forecast correlation of -0.2459.

To choose the right amount of components a good method is to calculate the PCTVAR, the percentage of variance of $\boldsymbol{Y}$ made up by each component added to the PLS algorithm [16].

$$\text{PCTVAR}(A) = \frac{\sum_{j=1}^{A} c_j^2}{\sum_{k=1}^{n} (y_k - \bar{y})^2},$$

where $c_j$ is the $Y$-weights and $\hat{y}$ is the mean value of $\boldsymbol{y}$. Using the minimum amount of components that still make up for most of the variance is ideal. Fig 6.4 shows the PCTVAR as a function of $A$ for the basic model regression, and it is clear that there is little use for having more than twelve components, giving us $A = 12$. By inspection of several regressions made in different time steps we conclude that this $A$ is at the same time small enough not to cause overfitting.

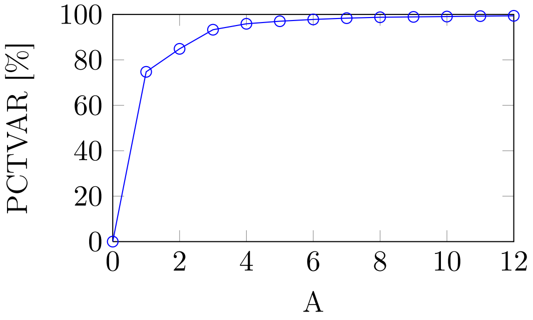


Figure 6.4: Percentage of the variation of $\boldsymbol{Y}$ explained by the components as a function of number of $A$

## 6.4 Historically Relevant Variables

With the methods for variable selection introduced in Chapter 5, the dimension of the model, i.e. number of predictors for the regression, can be reduced. A good variable selection generally gives better predictions by only selecting relevant variables for the modeling. However, in our case we will also use this method to classify which variables can be considered historically relevant for the forecast, i.e. predictors showing a time independent significance for describing the CLI.

The movements of economic variables may differ over time, and this must always be taken into consideration when creating financial models. Common methods for dealing with this fact are, for autoregressive models, to assume that the variable under study is a stationary time series, or for low dimension regression models, to assume that the relationship between e.g. the business cycle and industrial production is time independent. For models based on few variables and simple, maybe linear, relationships, these assumptions can be valid based on qualitative knowledge of the involved time series. But in our case, with a very large $M$, and a sophisticated algorithm taking inter-predictor relations into consideration, manually evaluating each predictor and its relation to every other predictor would be too tedious. Therefore, the time dependency of the variables need to be quantitatively assessed with automated variable selection.

To select the historically relevant variables, the basic model and variable selection methods will be applied and evaluated through iterations over different time steps of a training set, see Section 7.1.

# Chapter 7

# Cross-validation - Full Model

To evaluate the complete forecasting model, including the methods for variable selection, lagging predictors and the actual PLS-regression, a Cross-Validation (CV) approach is adapted. In CV the dataset is partitioned into disjoint subsets, where separate data are used for constructing and evaluating the model. A common type of CV is the 2-fold version, where two subsets of equal size are used, the training set and the validation set. The model is constructed and tuned using only data from the training set. The predictive performance of the model is then evaluated using the validation set, thus data-snooping is prevented and the validation becomes an out-of-sample test. In ordinary 2-fold CV the roles of the sets are then interchanged, and the procedure repeated, in order to further assess the prediction performance. We will use this method, with the exception of interchanging the sets, since in our case some of the predictors involved are of the kind "same period previous year". This means if we do not keep the chronological order of the observations and subsets, data from predictors showing previous year's values of certain economic indicators will be used in the training set, and then again in the validation set, via the present time version of the same economic indicators, containing the same data. This would let the model use future data. To avoid this the training set has to exclusively contain data available prior in time to the validation set, thus we let the first half of our observations make up the training set and the second half the validation set.

With our 288 observations spanning from 1990:01 to 2013:12, we get 144 data points in each partition. The training set will then span from 1990:01 to 2001:12 and the validation set from 2002:01 to 2013:12.

## 7.1 Model Training

To perform the variable selection discussed in Section 6.4, we construct an iterative method for evaluating the historical significance of the predictors. In the first step, the correlation function, $\hat{C}(s)$, is calculated for the entire training set and the predictors considered leading indicators are selected for the model and lagged accordingly.

Secondly the basic model, with its included six month forecast, is iteratively applied to shorter intervals of the training set, $\mathcal{I}_i = [t_i, t_{r+i}]$, for $i = 0, 1, \ldots, K$ where $K = T - r$, and $T$ is the number of observations in the training set while $k$ is the length of the intervals.

The interval's length, $r$, is chosen to represent an average length of a business cycle. According to Moore and Zarnowitz (1984), five years is a good estimation, i.e. we let $r = 60$. This will give every iteration of the model a chance to capture the full periodical movements of a business cycle, and include this information for variable selection.

With the original CV partition for the model training consisting of 144 monthly data, we discard 24 while lagging the matrix, see Section 6.1, and are left with $T = 120$. This gives us $K = 60$ iterations from which the predictors can be evaluated for historical significance through the variable selection methods and forecast performance.

As the basic model is applied to the time intervals, $\mathcal{I}_k$, the $M$ predictors' values for $\text{VIP}_k^m$ and $\beta_k^m$ are calculated. The predictors not meeting the respective cutoff values, $\eta$ and $\mu$, for the variable selection are excluded and the basic model is again applied to the same interval. This time the six month forecast is calculated and saved, along with the values of the selection parameters $\text{VIP}_k^m$ and $\beta_k^m$. The process is repeated for the $K$ iteration intervals, $\mathcal{I}_i, i = 0, \ldots, K$. A typical run of this process can be seen in Figure 7.1. The resulting six month forecasts are shown for every iteration of CLI for the OECD region, along with the actual values, as well as the corresponding forecast correlation in a bar diagram, $\rho_k$.

The ideal predictor, to be included in the final model, should have contributed to a correlated forecast and have selection parameter values above the cutoff, i.e. be significant, in every iterative time step of the training set. To select these predictors we can gather the ones that have fulfilled the variable selection criteria in every time step, and thus are considered significant, independent of the time frame, i.e. historically significant. However, not every iteration succeeds in casting an accurate forecast, as Figure 7.1 shows. While the majority of iterations, $I_k$, show good forecasting performance with high correlation values, shorter periods fail completely, even showing negative correlation. Therefore the information about significant predictors from

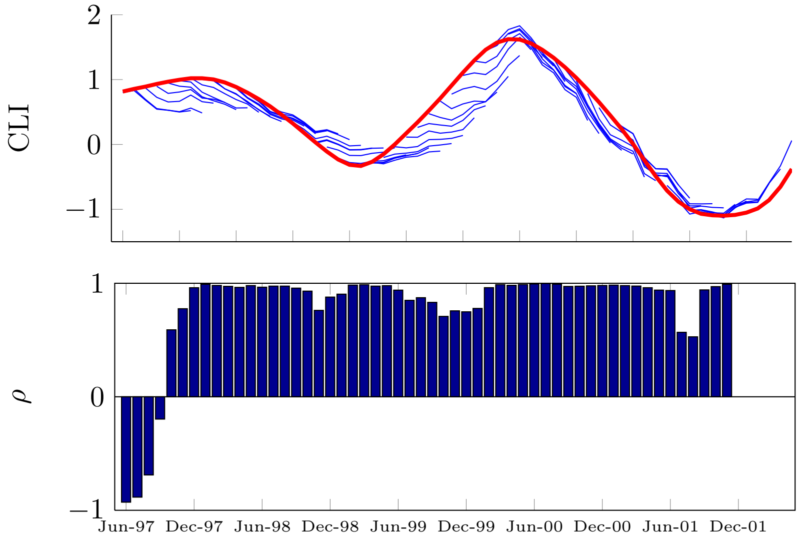


Figure 7.1: The forecast performance of the basic model with known lag in the training set. *Top:* The six month forecasts in blue along the actual CLI, red, re-modeled for each iteration. *Bottom:* The corresponding correlation between each forecast and the actual CLI .

the iterations with $\rho_k < 0$ will be discarded, since the basic model have failed in this time step, and there is no useful information for the variable selection to identify historically significant predictors. The iterations with useful information are thus defined as

$$I_k = \{k : \rho_k > 0\}, k = 1, \dots, K.$$

Reasons for the basic model failing in specific time periods, i.e. exhibiting negatively correlated forecasts in some iterations, might be explained by real world events. In the case of OECD's CLI, Figure 7.1 shows a period of negative correlation for forecasts made about the late 1997. This might be caused by irregularities in the behaviors of the economic predictors, due to events following the 1997 Asian financial crisis and the subsequent Ruble crisis. Regardless of the reason, these periods are overlooked in the automated variable selection. The number of iterations with negative $\rho$ varies for the different regions and cutoff values, and are usually between five and zero.

The predictors showing historical significance, $\boldsymbol{X}_{\text{select}}$, and thus considered as stationary time series on $\boldsymbol{X}$ are now chosen as

$$\boldsymbol{X}_{\text{select}} = \{\boldsymbol{x}_m\} \in \boldsymbol{I}_{\text{select}} \tag{7.1}$$

$$\boldsymbol{I}_{\text{select}} = \{m : \text{VIP}_m^k > \eta, \beta_m^k > \mu, \forall k \in I_k\}, \qquad m = 1, \ldots, M. \tag{7.2}$$

As the predictors showing historical significance on the training set have been selected, the "historical" beta coefficients can be calculated. These are created by a regression of $\boldsymbol{X}_{\text{select}}$ on the full training set of $T$ observations at once. These coefficients, $\hat{\boldsymbol{\beta}}$, can now be used to describe the CLI throughout the twelve years of training observations, and will likely be able to predict the corresponding values of the twelve years of validation observations.

We now have a historically optimized model, represented in selected variable indices $\boldsymbol{I}_{\text{select}}$, with corresponding beta coefficients $\hat{\boldsymbol{\beta}}$ and known lead $s_m$

$$\hat{y}_t = \sum_{i \in \boldsymbol{I}_{\text{select}}} X_i(t - s_m)\hat{\beta}_i. \tag{7.3}$$

This information will be carried over to the validation set, to examine the validity of our model and variable selection algorithm, see Section 7.3

## 7.2 Cutoff values

As mentioned in Chapter 5, there is a lack of consensus among researchers when it comes to an optimal value for the cutoff parameters $\eta$ and $\mu$. While many applications of PLS-DA in chemometrics uses cutoff values within similar ranges, it has been shown that the optimal values depend on different factors of the dataset [3]. With little guidance for PLS-DA applied in financial analysis, we will include a way to obtain the optimal cutoff values in our automated variable selection, inspired by similar work of Chong *et al.* (2005) and Fujiwara *et al.* (2012). The different economies under study will likely have different numbers of significant predictors to be found in the MEI dataset, depending on the region's size and trading partners. Therefore the cutoff values will be evaluated individually for each region, as part of the model training.

Like Fujiwara *et al.* (2012) we use trial and error to find the range of cutoff parameters to be evaluated, with the exception of not assuming a lower bound, i.e. we include zero as a possible value. This corresponds to the case of all variables being classified as significant enough for inclusion in the forecast model, which should be considered in our case since we want a method that does not need any prior knowledge of the economic indicators involved. The upper bound of the cutoff values are chosen as the highest value where the algorithm does not experience singularities. If the number of predictors selected are lower than the number of components chosen for the PLS algorithm, in our case $A = 12$, the regression can not be made. This gives the possible cutoff values for the PLS-VIP parameter $\eta = \{0, 0.1, \dots, 1.6, 1.7\}$.

For the PLS-Beta parameter, the steps between the lower values give relatively large number of discarded variables. Thus the cutoff values will be closer for the ten first steps, $\mu = \{0, 0.01, \dots, 0.09, 0.1, 0.2, 0.3, \dots, 2.4\}$.

The performance of the training model is measured in mean correlation for the iterated forecasts, $\hat{\rho}_{train}$ and average Root Mean Squared Error of the forecasts, RMSE. The results, along with the number of predictors classified as historically significant, $M$, are shown in Figures 7.2 and 7.3 as functions of of the different cutoff values. The most relevant parameter for choosing the optimal cutoff value is the RMSE, while the correlation is more of a complement.

It is clear that the variable selection improves the performance by discarding un-significant variables, since no region shows its lowest RMSE for cutoff values at zero. Similarly, the maximum correlation is often found in the mid range of the cutoff values, when a considerable amount of predictors have been discarded.

Noticeably the forecast performance for the region of Japan shows some irregularities in its dependence on the cutoff values. As will be discussed, the available data in the MEI dataset is not favorable for Japan, and the model shows limited forecast performance for this region overall.

The optimal cutoff values and corresponding performance for the PLS-VIP and PLS-Beta can be seen in Table 8.1.

As expected, the optimal cutoff values and number of variables included, differs between the regions, and the "greater than one" rule is not encouraged by the model. The goal for examining the cutoff values is not to contribute to the debate on general rules, but rather a proposed method to deal with the lack of consensus.

For the proposed method of PLS-VIP-Beta, there is hardly any academic guidelines, Chong *et al.* (2005) simply mentions that this method should use small cutoff values. Indeed, with variables being excluded by two combined criterions, the cutoff values sooner meets the limit of too few variables for the regression, especially in the $\eta$ dimension. The combined cutoff values considered is thus $\eta = \{0, 0.1, \dots, 1.4\}$, $\mu = \{0, 0.1, \dots, 2.4\}$, and the optimal values can be seen in Table 8.2.

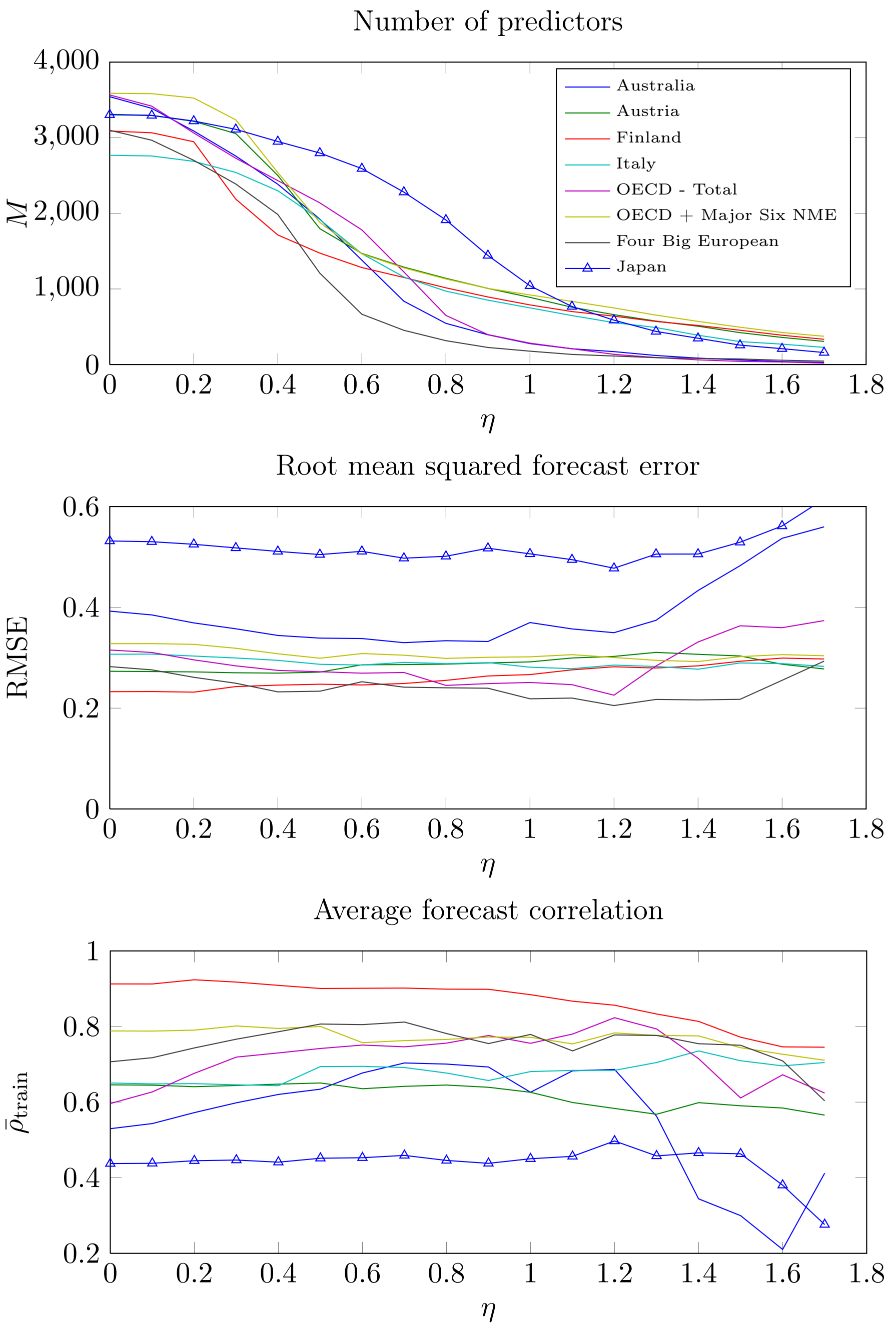


Figure 7.2: Performance of the model on the training set for different cutoff values $\mu$ along x-axis. *Top:* The number of predictors considered historically significant. *Middle:* Average RMSE of the training forecasts. *Bottom:* Average correlation of the training forecasts.

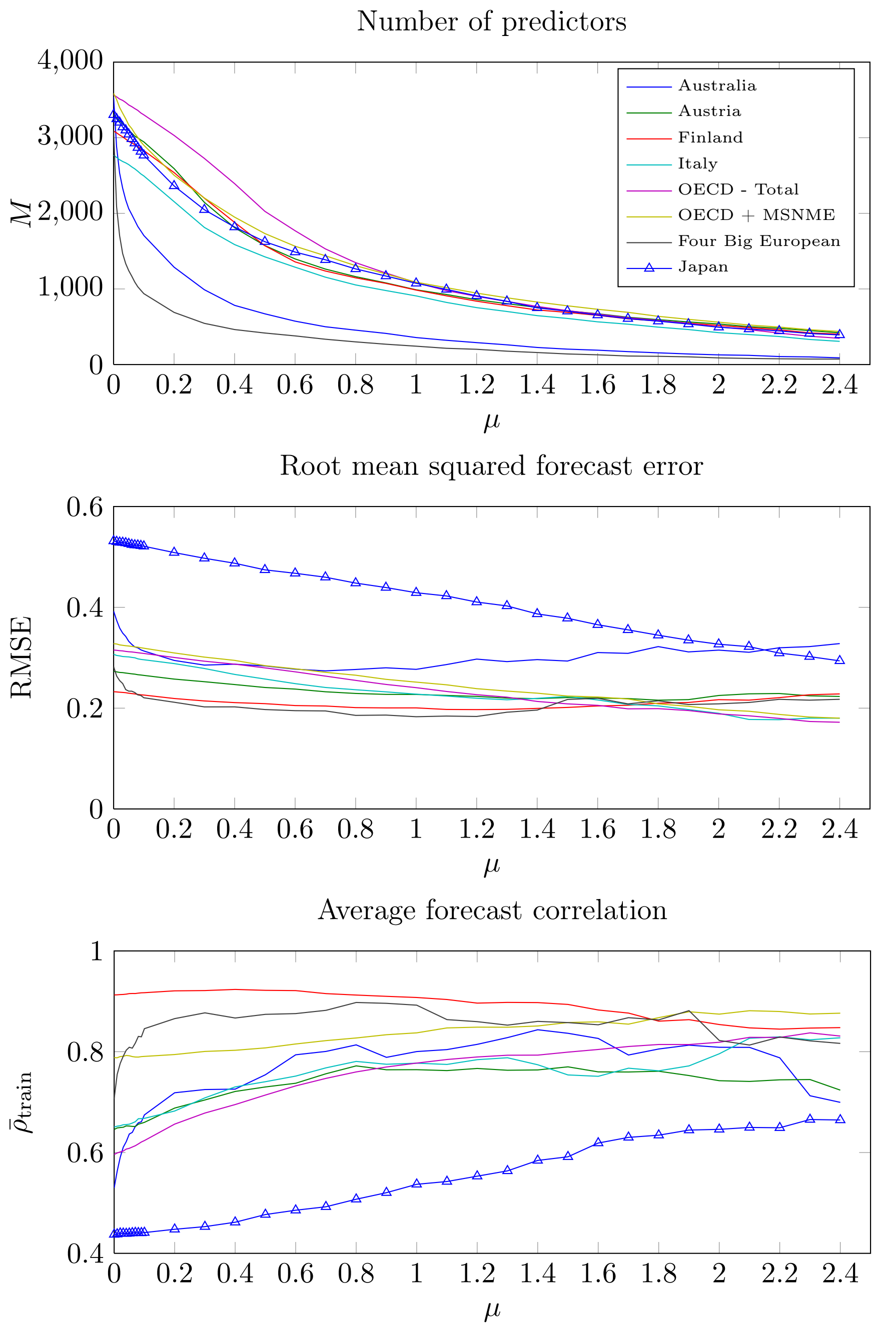


Figure 7.3: Performance of the model on the training set for different cutoff values $\mu$ along x-axis. *Top:* The number of predictors considered historically significant. *Middle:* Average RMSE of the training forecasts. *Bottom:* Average correlation of the training forecasts.

## 7.3 Model Validation

The resulting model to be validated, represented by $\boldsymbol{I}_{\text{select}}, \hat{\boldsymbol{\beta}}$ and $\boldsymbol{s}_m$, are carried over to the validation set, $\boldsymbol{X}^{\text{val}}$, consisting of the 144 observations between 2002:01 and 2013:12. To validate the forecast performance we will predict every observation of $\boldsymbol{Y}^{\text{val}}$ using X-data available six months in advance. This is done in a similar way to the forecasting of the basic model, but for all observations of the validation set at once

$$\hat{y}_t^{\text{val}} = \sum_{i \in \boldsymbol{I}_{\text{select}}} X_i^{\text{val}}(t - s_m)\hat{\beta}_i. \tag{7.4}$$

In words, the procedure uses the following steps. Firstly the predictors not showing historical significance are discarded. Secondly the remaining X-matrix, $\boldsymbol{X}_{\text{select}}^{\text{val}}$, is lagged with the known leads $\boldsymbol{s}_{\text{m}}$, as explained in Section 6.1, leaving us with V = 120 observations left. Finally the predictors are weighted with the beta coefficients $\hat{\boldsymbol{\beta}}$ and summed up to give the estimated $\boldsymbol{Y}$-time series, the CLI, for the whole validation set at once.

One measure of the forecast performance of the full model is the correlation between the out-of-sample prediction, $\hat{\boldsymbol{y}}$, and the actual CLI for the validation set, $\boldsymbol{y}$, that is, $\rho_{\text{val}} = \text{corr}(\boldsymbol{y}, \hat{\boldsymbol{y}})$. However, in a live version of the model, the forecast will be made at a given point in time for the subsequent six months, with the goal of predicting the movements and turning points of the CLI for this short period. A high correlation, $\rho_{\text{val}}$, for the complete 120 month validation interval, might not guarantee a good forecast performance in a shorter six month forecast. Thus dividing the validation set into six month subintervals, and measuring the mean of these intervals correlation with the corresponding actual CLI, $\bar{\rho}_{\text{val}}$, is a more appropriate validation measure, and will be our main parameter for valuing the forecast performance of the full model.

$$\bar{\rho}_{\text{val}} = \frac{1}{V - h} \sum_{i=1}^{V-h} \text{corr}\{(\boldsymbol{Y}_i, \ldots, \boldsymbol{Y}_{i+h}), (\hat{\boldsymbol{Y}}_i, \ldots, \hat{\boldsymbol{Y}}_{i+h})\},$$

where $V$ is the length of the validation set and $h$ the length of the forecast, i.e. 120 and 6, respectively.

# Chapter 8

# Results

The resulting prediction $\hat{\boldsymbol{y}} = \boldsymbol{X}_{\text{val}}\hat{\boldsymbol{\beta}}$ is shown along with the actual CLI for each region in Figure 8.1, for the PLS-VIP and PLS-Beta methods.

The optimal cutoff values, $\mu^*$ and $\eta^*$ are the ones giving the best results in the training set, i.e. the lowest mean RMSE for the training forecasts RMSE$^*$. The predictors being classified as historically significant using these cutoff values, are the ones included in the $\boldsymbol{I}_{\text{select}}$ vector of indices. These are the predictors carried over to the validation set for evaluating the model in Equation 7.4. The results from the model training and validation for the PLS-VIP and PLS-Beta methods can be seen in the Table 8.1. The two column to the farthest right are the results from the validation set, and measures the actual out-of-sample performance of our full model. A visualization of the results is given in Figure 8.1.

The model shows good forecasting performance for most regions, the exceptions being especially Australia and to some extent Japan, the reasons will be discussed in Chapter 9. Comparing the discrimination methods, PLS-VIP out-performs PLS-Beta in most cases. Notably, the variation of optimal number of predictors, $M^*$, chosen by the PLS-VIP model is far greater than that of the PLS-Beta model, with ranges from 113 to 2947 and 243 to 838 respectively.

The results of our proposed PLS-VIP-Beta method can be found in Table 8.2 as well as Figure 8.2. Regardless of performance measure, this method outperforms the singular methods for almost every region. While our experiment is not extensive enough to prove this method universally superior, it is still interesting that our findings are well in line with the assumptions of Fujiwara and Wold.

PLS-Beta

| | **RMSE**$^*$ | $\mu^*$ | $M^*$ | $\bar{\rho}^*_{\text{train}}$ | $\max(\bar{\rho}_{\text{train}})$ | $\min(\bar{\rho}_{\text{train}})$ | $\bar{\rho}_{\text{val}}$ | $\rho_{\text{val}}$ |
|---|---|---|---|---|---|---|---|---|
| **Australia** | 0.27394 | 0.7 | 500 | 0.80059 | 0.84361 | 0.52974 | -0.24065 | -0.0442 |
| **Austria** | 0.21581 | 1.8 | 594 | 0.76128 | 0.77192 | 0.64571 | 0.78044 | 0.85775 |
| **Finland** | 0.19711 | 1.2 | 838 | 0.89648 | 0.92359 | 0.84487 | 0.64133 | 0.81154 |
| **Italy** | 0.17707 | 2.2 | 371 | 0.82954 | 0.82954 | 0.65064 | 0.61684 | 0.85794 |
| **OECD - Total** | 0.17201 | 2.4 | 347 | 0.83108 | 0.83741 | 0.59626 | 0.60277 | 0.86392 |
| **OECD + Major Six NME** | 0.18018 | 2.4 | 438 | 0.87646 | 0.88138 | 0.78802 | 0.51524 | 0.74719 |
| **Four Big European** | 0.18298 | 1 | 243 | 0.89238 | 0.89771 | 0.70671 | 0.72111 | 0.83207 |
| **Japan** | 0.29383 | 2.4 | 392 | 0.66458 | 0.66549 | 0.4375 | 0.50296 | 0.18974 |

PLS-VIP

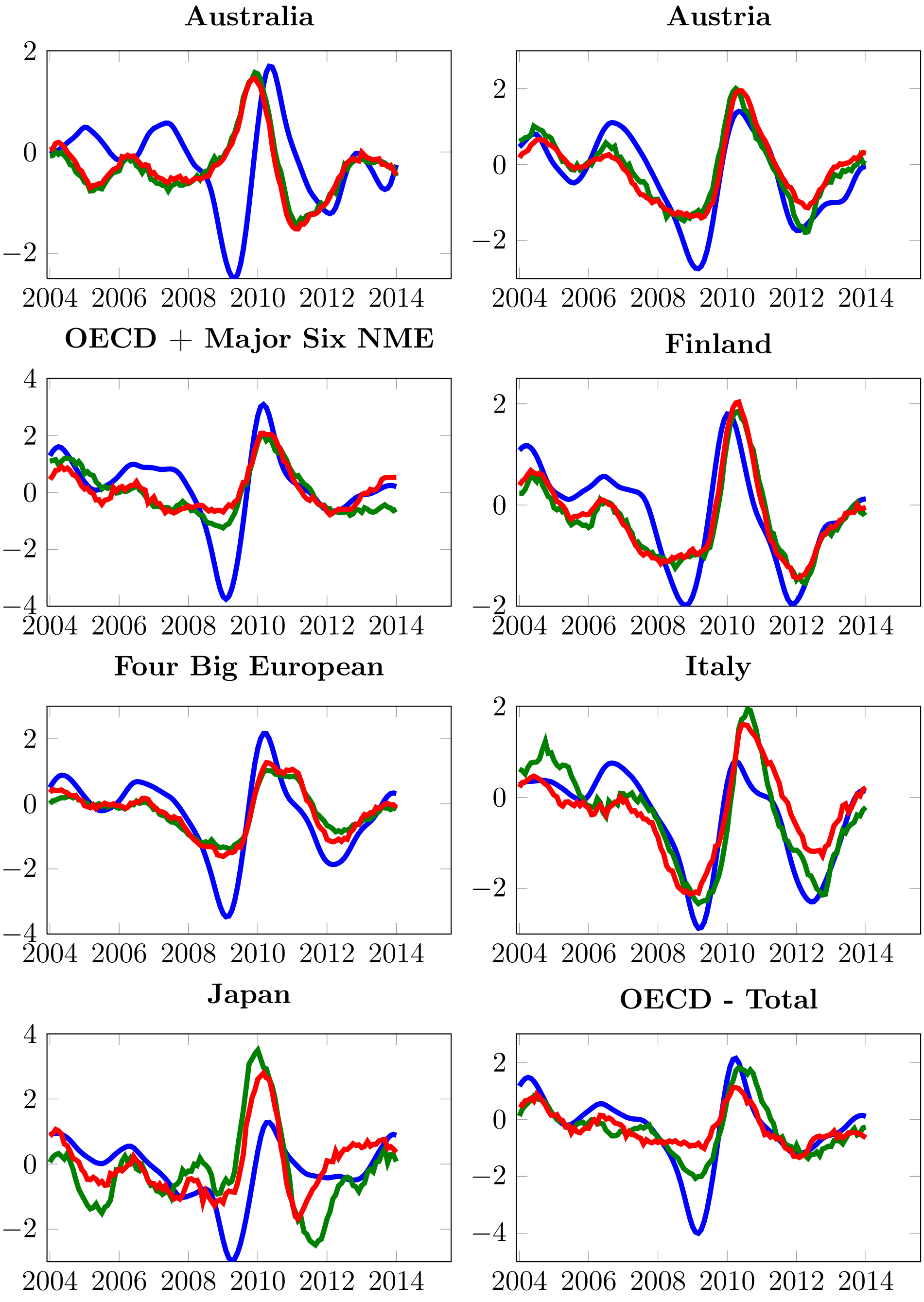
| | **RMSE**$^*$ | $\eta^*$ | $M^*$ | $\bar{\rho}^*_{\text{train}}$ | $\max(\bar{\rho}_{\text{train}})$ | $\min(\bar{\rho}_{\text{train}})$ | $\bar{\rho}_{\text{val}}$ | $\rho_{\text{val}}$ |
|---|---|---|---|---|---|---|---|---|
| **Australia** | 0.3299 | 0.7 | 838 | 0.70358 | 0.70358 | 0.2103 | -0.23314 | -0.0353 |
| **Austria** | 0.26947 | 0.4 | 2501 | 0.64762 | 0.65062 | 0.56568 | 0.82592 | 0.81284 |
| **Finland** | 0.23173 | 0.2 | 2947 | 0.92367 | 0.92367 | 0.7453 | 0.69562 | 0.83364 |
| **Italy** | 0.27725 | 1.4 | 388 | 0.73541 | 0.73541 | 0.64355 | 0.61282 | 0.77255 |
| **OECD - Total** | 0.22569 | 1.2 | 135 | 0.82328 | 0.82328 | 0.59626 | 0.64387 | 0.76861 |
| **OECD + Major Six NME** | 0.2924 | 1.4 | 571 | 0.77513 | 0.80148 | 0.71066 | 0.64326 | 0.67571 |
| **Four Big European** | 0.20514 | 1.2 | 113 | 0.77772 | 0.81175 | 0.60349 | 0.68699 | 0.87641 |
| **Japan** | 0.47758 | 1.2 | 587 | 0.49722 | 0.49722 | 0.27621 | 0.33979 | 0.48785 |

Table 8.1: The results from the cutoff-value-optimization for the PLS-Beta and PLS-VIP methods. The optimal cutoff values $\mu^*$ and $\eta^*$ are the ones where the Root Mean Squared Error is at its minimum, RMSE*. The optimal values for mean correlation $\bar{\rho}_{\text{train}}$ and number of predictors $M$, resulting from the optimal cutoff values, are marked with *. The two rightmost columns are the resulting mean and total correlation of the validation set $\rho_{\text{val}}$. OECD - Total is the combined economy of the OECD member countries, OECD + Major Six NME stands for OECD plus Major Six Non-Member Economies, and consist of the 30 OECD countries plus Brazil, China, India, Indonesia, the Russian Federation and South Africa. Four Big European is the combined economies of France, Germany, Italy and United Kingdom

PLS-VIP-Beta

| | **RMSE**$^*$ | $\eta^*$ | $\mu^*$ | $M^*$ | $\rho^*_{\text{train}}$ | $\max(\bar{\rho}_{\text{train}})$ | $\min(\bar{\rho}_{\text{train}})$ | $\bar{\rho}_{\text{val}}$ | $\rho_{\text{val}}$ |
|---|---|---|---|---|---|---|---|---|---|
| **Australia** | 0.27394 | 0.1 | 0.8 | 500 | 0.80059 | 0.84361 | 0.21063 | -0.1979 | 0.00474 |
| **Austria** | 0.13956 | 0.5 | 2.4 | 100 | 0.90565 | 0.91317 | 0.56791 | 0.8416 | 0.89195 |
| **Finland** | 0.18768 | 0.2 | 0.3 | 946 | 0.94027 | 0.95413 | 0.70594 | 0.65566 | 0.81798 |
| **Italy** | 0.15027 | 0.1 | 0.8 | 329 | 0.87681 | 0.92698 | 0.64355 | 0.65888 | 0.78917 |
| **OECD - Total** | 0.15222 | 0.5 | 2.3 | 60 | 0.84032 | 0.86931 | 0.55033 | 0.68713 | 0.86378 |
| **OECD + Major Six NME** | 0.14431 | 0.2 | 1 | 251 | 0.93848 | 0.95574 | 0.75401 | 0.8853 | 0.80085 |
| **Four Big European** | 0.17985 | 0.7 | 1.3 | 177 | 0.89125 | 0.91135 | 0.65548 | 0.69258 | 0.88405 |
| **Japan** | 0.28725 | 0.7 | 1 | 159 | 0.7661 | 0.82503 | 0.11489 | 0.46803 | 0.43737 |

Table 8.2: The results from the cutoff-value-optimization for the combined PLS-VIP-Beta method. The optimal cutoff values $\mu^*$ and $\eta^*$ are the ones where the Root Mean Squared Error is at its minimum, RMSE$^*$. The optimal values for mean correlation $\bar{\rho}_{\text{train}}$ and number of predictors $M$, resulting from the optimal cutoff values, are marked with $^*$. The two rightmost columns are the resulting mean and total correlation of the validation set $\rho_{\text{val}}$. OECD - Total is the combined economy of the OECD member countries, OECD + Major Six NME stands for OECD plus Major Six Non-Member Economies, and consist of the 30 OECD countries plus Brazil, China, India, Indonesia, the Russian Federation and South Africa. Four Big European is the combined economies of France, Germany, Italy and United Kingdom


Australia

Austria

OECD + Major Six NME

Finland

Four Big European

Italy

Japan

OECD - Total

Figure 8.1: Forecasting performance for out-of-sample tests for the PLS-Beta, green, and PLS-VIP, red, against the actual CLI in blue.

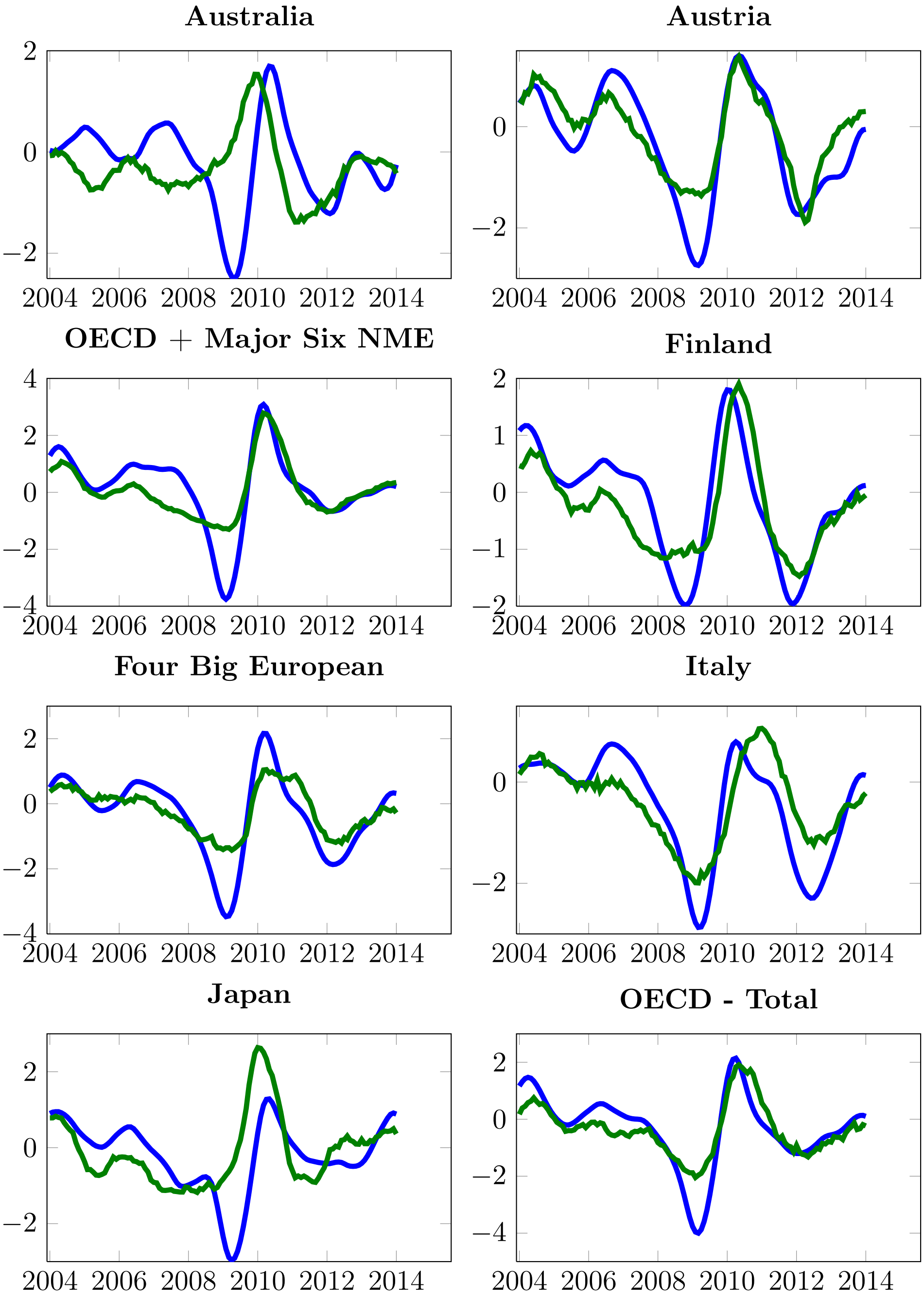


Figure 8.2: Forecasting performance for out-of-sample tests for the combined PLS-VIP-Beta in green against the actual CLI in blue.

One of our objectives with this thesis was to see if the lead of the CLI could be improved, or rather, the CLI itself could be forecasted, by a model including non-domestic data. To inspect the ratio of internal and external predictors, i.e. domestic and non-domestic, chosen by the automated variable selection, the nationality of the predictors are presented in Table 8.3. The results are shown for the most accurate version of the model, i.e. the predictors selected by the PLS-DA with highest $\bar{\rho}_{\text{val}}$ for each region. The number of used internal predictors are shown along with the total number of predictors used, total number of internal predictors available and the percentage of internal predictors among the ones used. The larger zone aggregates clearly have more internal than external variables in the dataset, and the region OECD + Major Six Non-Member Economiest of course includes all of the regions covered in the MEI, and thus all 5012 economic indicators are internal.

| | Internal | Total | Total Internal | Pct Interior |
|---|---|---|---|---|
| **Australia** | 25 | 500 | 244 | 5 % |
| **Austria** | 13 | 100 | 147 | 13 % |
| **Finland** | 89 | 2947 | 160 | 3 % |
| **Italy** | 18 | 329 | 156 | 5.5 % |
| **OECD - Total** | 51 | 60 | 4619 | 85 % |
| **OECD + MSNME** | 251 | 251 | 5012 | 100 % |
| **Four Big European** | 61 | 243 | 553 | 25.1 % |
| **Japan** | 0 | 392 | 242 | 0 % |

Table 8.3: The number of selected Internal predictors are shown along with the Total number of variables selected by the model. Total Internal are the number of available domestic variables in the MEI for the specific region, and Pct Internal is the percentage of the selected predictors that are domestic.

When available the model often choses a combination of external and internal data, and often discards many of the internal predictors in favor of external ones. The exception, again, is Japan, where no internal predictors where chosen.

At this point it might be interesting to see exactly what kind of economical indicators of the MEI the full model uses, and in which way they are included in the prediction. Since different indicators are picked for different regions, and some regions using indicators in the thousands, a complete overview of the resulting modeling variables is too extensive. Instead we give an example of how they may look by presenting one of the most accurately predicted regions with relatively few predictors, namely the PLS-VIP-Beta prediction of Austria. The complete set of variables for modeling, namely $\hat{\boldsymbol{\beta}}$, $s_m$ and $\boldsymbol{I}_{\text{select}}$ (represented here by the actual name of the indicators), are presented in Appendix A.

# Chapter 9

# Discussion

When inspecting the model performance on the validation set in Figure 8.1 and 8.2, the financial crisis of 2008 is evident in the middle of the validation set for every region's CLI. The model recognizes the trough and the following expansion of this period, but fails somewhat to predict the extent of the recession, in most cases. Since the goal of the model is, as mentioned, to predict the turning points and the movements rather than the actual values of the composite leading indicator, this is not a big setback. Further one may discuss the possibility or even the validity of a model being able to predict the extent of the financial crisis based solely on historical data. With hindsight a model could well be constructed to put heavier weights on variables known to play a big part in triggering the crisis, however, only using data available at the time, one may argue that these variables have not had the same extreme impact on the economy in the previous observations of the dataset employed. Therefore a model recognizing the exact extent of the financial crisis, might have weighted certain variables, e.g. interest rate spread and mortgage rate, higher than would be historically correct, and thus might not give accurate predictions in a post crisis forecast.

While the overall forecast performance is good, a small number of periods show less satisfying results, where the model's forecast deviates from the actual CLI. We will not go into detail for each time period of the concerned regions, in order to find real world explanations for every deviation. However, two regions stand out with very unsatisfying forecast performances that needs to be addressed, namely Australia and Japan.

The most pronounced deviation in the forecasted and actual values of the CLI is the model for Australia, in the period of the 2008 financial crisis. Compared to the composite leading indicator, the forecast does not even signal a trough, but rather an expansion followed by a lowest period a couple of years later. While this is an exceptionally bad forecast of the CLI, the actual economy of Australia did not suffer a severe recession in this period, but is rather the only developed nation to grow in the first half of 2009 [11]. Thus the CLI itself is at fault with its prediction, and the deviation of our forecast might be excused for this period. With this said, the overall performance of Australia and Japan is still not satisfactory for the validation sets, as well as the training set of Japan. Looking at the larger picture, one explanation can be the location of these regions in relation to the available data. The majority of the OECD member regions are located in Europe, leaving most of Asia and the Pacific unrepresented in the MEI dataset. This leaves Australia and Japan with little data about occurrences of their neighboring markets, and with that a large part of their international trade.

We have deliberately focused on the quantitative analysis, and not letting qualitative knowledge or assumptions about individual economic indicators and regional markets affect the design of the model or variable selection. Similarly, choosing the amount of predictors used in the optimal model is decided by the automated variable selection method, when the forecast is optimized in the training set.

Reasons for our cutoff values in some cases being significantly smaller than other studies, and does not fulfill the greater than one rule in some cases, can be the iterations of discrimination over multiple time steps. With many iterations, each discrimination puts relatively weak demands of significance on the predictors, the idea being; it is better for a predictor to be moderately significant in all of the time steps than being extremely significant for just one time step.

# Chapter 10

# Conclusion

This thesis assessed the regression and variable selection methods of partial least squares for financial forecasting. Using these methods to select and use numerous economic variables made it possible to accurately predict the movements of the OECD composite leading indicators ahead of time for most of the regions under study.

Overall, the partial least squares approach shows to be very useful in financial analysis. The regression method does not demand dimension reduction when the variables used are of significance, and the variable selection method does not need any prior knowledge or individual studies of the variables nor their collinearity, to be evaluated for significance. Therefore more data is never a disadvantage, and the number of variables to be studied can preferably be as large as possible.

The out-of-sample tests implies that there is much information to be found in inter-regional data, which is not taken into account today by the OECD's forecast. This is in agreement with the results of Fichtner *et al.* (2011), claiming that complementing the forecasting model with non-domestic variables can increase the lead of the composite leading indicators for many regions, or equivalently, forecast the leading indicator itself.

The method developed in this thesis to find historically significant variables has not been previously employed. It is developed with the specific dataset and objectives in mind, and thus its external validity might be discussed. However, it shows the effectiveness of the Variable Importance on Projection and the Beta coefficients as econometric parameters for financial analysis. The length of the iteration intervals as well as the number of iterations in the model training should be closer examined if this method is to be applied in further studies. Conclusively it is an idea for algorithmically dealing with the non-stationary properties of financial time series, that shows potential.

# Bibliography

[1] Barker, M., and Rayens, W., *Partial least squares for discrimination* Journal of Chemometrics (2003); 17: 166-173

[2] Boulesteix, A.L., and Stimmer, K., *Partial least squares: a versatile tool for the analysis of high-dimensional genomic data* Briefings in Bioinformatics. Vol 8. No 1. 32-44 (2006)

[3] Chong, I.-G., and Jun, C.-H., *Performance of some variable selection methods when multicollinearity is present* Chemometrics and Intelligent Laboratory Systems, 78 (2005) 103-112

[4] Delaigle, A., and Hall, P., *Methodology And Theory For Partial Least Squares Applied To Functional Data* Annals of Statistics 2012, Vol. 40, No. 1, 322-352 (2012)

[5] Fichtner, F., Rasmus, R., and Schnatz, B., *The Forecasting Performance of Composite Leading Indicators: Does Globalisation Matter?* OECD Journal: Journal of Business Cycle Measurement and Analysis, Vol. 2011/1 (2011)

[6] Fujiwara, K., Sawada, H., and Kano, M., *Input variable selection for PLS modeling using nearest correlation spectral clustering* Chemometrics and Intelligent Laboratory Systems, 118 (2012) 109-119

[7] Manne, R., *Analysis of Two Partial-Least-Squares Algorithms for Multivariate Calibration* Chemometrics and Intelligent Laboratory Systems, 2 (1987) 187-197

[8] Mevik, B.H., and Wehrens, R., *The pls Package: Principal Component and Partial Least Squares Regression in R* Journal of Statistical Software, January (2007), Volume 18, Issue 2.

[9] Moore, G., and Zarnowitz, V., *The development and role of the National bureau's Busines cycle chronologies* National Bureau of Economic Research (1984)

[10] Naes, T., and Martens, H., *Comparison of prediction methods for multicollinear data* Communications in Statistics - Simulation and Computation, 14:3, 545-576, (1985)

[11] Nanto, D., *The Global Financial Crisis: Analysis and Policy Implications*, in: J. Gallagher, and E. Wilkins (Eds.), The Global Financial Crisis: Policies and Implications, (2011)

[12] Pérez-Enciso, M., and Tenenhaus, M., *Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach* Human Genetics (2003) 112 : 581Ð592

[13] Saxena, A.K., and Prathipati, P., *Comparison of MLR, PLS and GA-MLR in QSAR analysis**, SAR and QSAR in Environmental Research, 14:5-6, 433-445 (2007)

[14] Stock, J.H. and Watson, M.W., *Forecasting with many predictors*, in: G. Elliott, and A. Timmermann (Eds.), Handbook of Forecasting, (2006)

[15] Wentzell, P., and Vega Montoto, L., *Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures* Chemometrics and Intelligent Laboratory Systems 65 257-279 (2003)

[16] Wold, S., Geladi, P., Esbensen, K., and Öhman, J., *Multi-Way Principal Components- and PLS-analysis* Journal of Chemometrics, Vol. 1, 41-56 (1987)

[17] Wold, S., Sjöström, M., and Eriksson, L., *PLS-regression: a basic tool of chemometrics* Chemometrics and Intelligent Laboratory Systems 58 (2001)

# Appendix A

The below table shows the resulting full model of the Composite Leading Indicator for Austria, represented by the complete set of predictors used, by OECD referred to as subjects, and their respective regression coefficients $\beta$ and lead $s$. The time series of the predictors can be obtained from the OECD website iLibrary from the package named Main Economic Indicators - Complete Database. Acronyms for the specific Measures, M, of the predictors include

GPY = 'Growth rate same period previous year'

LRN = 'Level, rate or national currency'

NCM = 'National currency, monthly level, s.a.'

NOR = 'Normalised, seasonally adjusted (normal = 100)'

s.a. = seasonally adjusted.

Once the specific time series have been obtained from the database, the model specifications below can be used to forecast the future CLI up to six months ahead. Firstly the predictors need to be centered and scaled as explained in Section 3.2 before applying Equation 7.3.

| Country | Subject | M | s | $\beta$ |
|---|---|---|---|---|
| Australia | Leading Indicators OECD > Leading indicators > CLI > Trend restored | GPY, s.a. | 6 | 8.479989e-02 |
| Austria | Leading Indicators OECD > Leading indicators > CLI > Amplitude adjusted | LRN, s.a. | 15 | 2.016549e-02 |
| Austria | Leading Indicators OECD > Leading indicators > CLI > Normalised | LRN, s.a. | 15 | 2.095110e-02 |
| Austria | Leading Indicators OECD > Component series > BTS - Business situation > Normalised | LRN, s.a. | 14 | 2.421221e-02 |
| Austria | Leading Indicators OECD > Component series > BTS - Business situation > Original series | LRN, s.a. | 14 | 7.805710e-03 |
| Austria | Leading Indicators OECD > Component series > BTS - Order books > Normalised | LRN, s.a. | 13 | 3.683123e-02 |
| Austria | Leading Indicators OECD > Component series > BTS - Order books > Original series | LRN, s.a. | 13 | -8.031397e-04 |
| Austria | Business tendency surveys (manufacturing) > Production > Tendency > National indicator | LRN, s.a. | 14 | -2.215161e-02 |
| Austria | Business tendency surveys (manufacturing) > Finished goods stocks > Level > National indicator | LRN, s.a. | 14 | 3.066804e-02 |
| Austria | Business tendency surveys (manufacturing) > Order books > Level > National indicator | LRN, s.a. | 13 | -8.031397e-04 |
| Austria | Business tendency surveys (manufacturing) > Export order books or demand > Level > National indicator | LRN, s.a. | 12 | -2.880231e-02 |
| Austria | Business tendency surveys (manufacturing) > Selling prices > Future tendency > National indicator | LRN, s.a. | 13 | -5.747147e-02 |
| Austria | Business tendency surveys (manufacturing) > Confidence indicators > Composite indicators > National indicator | LRN, s.a. | 14 | -2.673673e-02 |
| Austria | Business tendency surveys (manufacturing) > Confidence indicators > Composite indicators > OECD Indicator | NOR | 14 | -4.285786e-02 |
| Belgium | Leading Indicators OECD > Reference series > Gross Domestic Product (GDP) > Ratio to trend | LRN, s.a. | 14 | -1.150100e-01 |
| Belgium | Leading Indicators OECD > Reference series > Gross Domestic Product (GDP) > Normalised | LRN, s.a. | 14 | -1.151701e-01 |
| Belgium | Leading Indicators OECD > Leading indicators > CLI > Amplitude adjusted | LRN, s.a. | 15 | -6.808495e-02 |
| Belgium | Leading Indicators OECD > Leading indicators > CLI > Normalised | LRN, s.a. | 15 | -6.811692e-02 |
| Belgium | Leading Indicators OECD > Component series > CS - Confidence indicator > Normalised | LRN, s.a. | 12 | -4.752689e-02 |
| Belgium | Business tendency surveys (manufacturing) > Export order books or demand > Level > National indicator | LRN, s.a. | 14 | 1.708603e-02 |
| Chile | Share Prices > All shares/broad > Total > Total | GPY | 16 | -3.557249e-02 |
| Denmark | Leading Indicators OECD > Component series > BTS - Employment > Normalised | LRN, s.a. | 14 | -6.089601e-02 |
| Denmark | Leading Indicators OECD > Component series > CS - Confidence indicator > Normalised | LRN, s.a. | 13 | -1.895632e-02 |
| Denmark | Business tendency surveys (manufacturing) > Order books > Level > National indicator | LRN, s.a. | 13 | 2.108563e-02 |
| Denmark | Business tendency surveys (manufacturing) > Export order books or demand > Level > National indicator | LRN, s.a. | 13 | 4.897350e-03 |
| Denmark | Business tendency surveys (manufacturing) > Employment > Future Tendency > National indicator | LRN, s.a. | 13 | -3.773022e-02 |
| Denmark | Consumer opinion surveys > Economic Situation > Future tendency > National indicator | LRN, s.a. | 14 | -1.441759e-02 |
| Denmark | Currency Conversions > US$ exchange rate > Average of daily rates > National currency:USD | GPY | 6 | 4.170035e-03 |
| Denmark | International Trade > Imports > Value (goods) > Total | GPY, s.a. | 15 | -2.473760e-03 |
| Euro area | Leading Indicators OECD > Leading indicators > CLI > Amplitude adjusted | LRN, s.a. | 14 | -2.000094e-02 |
| Euro area | Leading Indicators OECD > Leading indicators > CLI > Normalised | LRN, s.a. | 14 | -2.349089e-02 |
| Euro area | Business tendency surveys (manufacturing) > Finished goods stocks > Level > National indicator | LRN, s.a. | 14 | 5.154920e-02 |
| Euro area | Business tendency surveys (manufacturing) > Order books > Level > National indicator | LRN, s.a. | 12 | 1.054844e-02 |
| Euro area | Business tendency surveys (manufacturing) > Export order books or demand > Level > National indicator | LRN, s.a. | 12 | 8.257322e-03 |
| Euro area | Business tendency surveys (manufacturing) > Selling prices > Future tendency > National indicator | LRN, s.a. | 12 | -5.973115e-02 |
| Euro area | Business tendency surveys (manufacturing) > Confidence indicators > Composite indicators > National indicator | LRN, s.a. | 13 | 1.037582e-02 |
| Euro area | Business tendency surveys (manufacturing) > Confidence indicators > Composite indicators > OECD Indicator | NOR | 13 | 1.285426e-02 |
| Euro area | Currency Conversions > US$ exchange rate > Average of daily rates > National currency:USD | GPY | 7 | 5.099938e-03 |
| European Union | Production > Industry > Total industry > Total industry excluding construction | GPY, s.a. | 15 | 1.149903e-03 |
| France | Leading Indicators OECD > Component series > BTS - Export orders > Normalised | LRN, s.a. | 12 | -7.362950e-02 |
| France | Leading Indicators OECD > Component series > BTS - Export orders > Original series | LRN, s.a. | 12 | 3.745184e-04 |
| France | Leading Indicators OECD > Component series > BTS - Production > Normalised | LRN, s.a. | 14 | -6.521366e-02 |
| France | Leading Indicators OECD > Component series > CS - Confidence indicator > Normalised | LRN, s.a. | 11 | 6.951113e-02 |
| France | Business tendency surveys (manufacturing) > Production > Tendency > National indicator | LRN, s.a. | 12 | 1.969416e-02 |
| France | Business tendency surveys (manufacturing) > Order books > Level > National indicator | LRN, s.a. | 12 | 5.776520e-04 |
| France | Business tendency surveys (manufacturing) > Export order books or demand > Level > National indicator | LRN, s.a. | 12 | 3.745184e-04 |
| Four Big European | Leading Indicators OECD > Leading indicators > CLI > Amplitude adjusted | LRN, s.a. | 15 | -7.559021e-03 |
| Four Big European | Leading Indicators OECD > Leading indicators > CLI > Normalised | LRN, s.a. | 14 | 2.272419e-03 |
| Germany | Leading Indicators OECD > Leading indicators > CLI > Amplitude adjusted | LRN, s.a. | 15 | 2.152083e-02 |
| Germany | Leading Indicators OECD > Leading indicators > CLI > Normalised | LRN, s.a. | 15 | 2.126322e-02 |
| Germany | Leading Indicators OECD > Component series > BTS - Business situation > Normalised | LRN, s.a. | 14 | 2.421221e-02 |
| Germany | Leading Indicators OECD > Component series > BTS - Business situation > Original series | LRN, s.a. | 14 | 7.805710e-03 |
| Germany | Leading Indicators OECD > Component series > BTS - Finished goods stocks > Normalised | LRN, s.a. | 13 | -4.574814e-02 |
| Germany | Business tendency surveys (manufacturing) > Selling prices > Future tendency > National indicator | LRN, s.a. | 12 | -6.475695e-02 |
| Germany | Business tendency surveys (manufacturing) > Business situation > Current > National indicator | LRN, s.a. | 14 | 7.805710e-03 |
| Iceland | Consumer Price Index > OECD Groups > Energy (Fuel, electricity & gasoline) > Total | GPY | 24 | -1.185126e-01 |
| Ireland | Leading Indicators OECD > Component series > Exports of goods > Normalised | LRN, s.a. | 9 | 3.504725e-02 |
| Italy | Leading Indicators OECD > Leading indicators > CLI > Amplitude adjusted | LRN, s.a. | 12 | -4.172970e-02 |
| Italy | Leading Indicators OECD > Leading indicators > CLI > Normalised | LRN, s.a. | 12 | -4.183325e-02 |
| Italy | Leading Indicators OECD > Component series > BTS - Order books > Normalised | LRN, s.a. | 12 | 2.626831e-02 |
| Italy | Leading Indicators OECD > Component series > BTS - Production > Normalised | LRN, s.a. | 14 | -1.589780e-03 |
| Italy | Leading Indicators OECD > Component series > BTS - Production > Original series | LRN, s.a. | 14 | 2.232003e-02 |
| Italy | Leading Indicators OECD > Component series > CS - Confidence indicator > Normalised | LRN, s.a. | 12 | -8.231776e-02 |
| Italy | Leading Indicators OECD > Component series > Orders > Normalised | LRN, s.a. | 12 | -3.173817e-02 |
| Italy | Production > Industry > Total industry > Total industry excluding construction | GPY, s.a. | 13 | 3.310629e-02 |
| Italy | Business tendency surveys (manufacturing) > Production > Future Tendency > National indicator | LRN, s.a. | 14 | 2.232003e-02 |
| Italy | Business tendency surveys (manufacturing) > Selling prices > Future tendency > National indicator | LRN, s.a. | 12 | 3.931008e-02 |
| Italy | Business tendency surveys (manufacturing) > Confidence indicators > Composite indicators > National indicator | LRN, s.a. | 13 | 2.563783e-02 |
| Italy | Business tendency surveys (manufacturing) > Confidence indicators > Composite indicators > OECD Indicator | NOR | 13 | 3.279551e-02 |
| Italy | International Trade > Imports > Value (goods) > Total | GPY, s.a. | 14 | -2.956249e-03 |
| Mexico | Leading Indicators OECD > Component series > BTS - Finished goods stocks > Normalised | LRN, s.a. | 10 | -5.370177e-02 |
| Netherlands | Leading Indicators OECD > Component series > BTS - Business situation > Normalised | LRN, s.a. | 14 | 2.421221e-02 |
| Netherlands | Leading Indicators OECD > Component series > BTS - Business situation > Original series | LRN, s.a. | 14 | 7.805710e-03 |
| Netherlands | Leading Indicators OECD > Component series > BTS - Finished goods stocks > Normalised | LRN, s.a. | 12 | -1.098570e-02 |
| Netherlands | Business tendency surveys (manufacturing) > Selling prices > Future tendency > National indicator | LRN, s.a. | 13 | -3.839206e-03 |
| Netherlands | Labour Force Survey - quarterly levels > Harmonised unemployment - monthly levels > Aged 15-24 > Females | LRN | 19 | -6.635479e-03 |
| Netherlands | Labour Force Survey - quarterly levels > Harmonised unemployment - monthly levels > Aged 15-24 > Females | LRN, s.a. | 19 | 2.256546e-02 |
| Netherlands | Labour Force Survey - quarterly rates > Harmonised unemployment - monthly rates > Aged 15-24 > Females | LRN | 19 | 1.615350e-03 |
| Netherlands | Labour Force Survey - quarterly rates > Harmonised unemployment - monthly rates > Aged 15-24 > Females | LRN, s.a. | 19 | 5.464048e-02 |
| Netherlands | Consumer Price Index > OECD Groups > Energy (Fuel, electricity & gasoline) > Total | GPY | 16 | -2.372285e-02 |
| New Zealand | Leading Indicators OECD > Component series > Short-term interest rate > Normalised | LRN, s.a. | 9 | -7.705664e-02 |
| Norway | Monetary aggregates and their components > Broad money and components > Broad money, index | GPY, s.a. | 19 | 2.915270e-02 |
| OECD - Europe | Leading Indicators OECD > Leading indicators > CLI > Amplitude adjusted | LRN, s.a. | 15 | 7.593098e-04 |
| OECD - Europe | Leading Indicators OECD > Leading indicators > CLI > Normalised | LRN, s.a. | 15 | -9.655715e-03 |
| Portugal | Leading Indicators OECD > Component series > BTS - Export orders > Normalised | LRN, s.a. | 14 | -8.682508e-02 |
| Slovak Republic | Leading Indicators OECD > Component series > Imports > Normalised | LRN, s.a. | 17 | 3.747295e-02 |
| Slovak Republic | International Trade > Imports > Value (goods) > Total | GPY, s.a. | 15 | 4.724419e-03 |
| Slovak Republic | International Trade > Net trade > Value (goods) > Total | NCM, s.a. | 14 | -1.714230e-02 |
| South Africa | Leading Indicators OECD > Component series > BTS - Business situation > Normalised | LRN, s.a. | 14 | -2.300338e-03 |
| South Africa | Leading Indicators OECD > Component series > Interest rate spread > Normalised | LRN, s.a. | 13 | -1.711648e-02 |
| Spain | Business tendency surveys (manufacturing) > Export order books or demand > Level > National indicator | LRN, s.a. | 12 | 5.942555e-02 |
| Spain | Business tendency surveys (manufacturing) > Selling prices > Future tendency > National indicator | LRN, s.a. | 12 | 4.283469e-02 |
| Sweden | Leading Indicators OECD > Component series > Long-term interest rate > Normalised | LRN, s.a. | 10 | -1.595587e-01 |
| Switzerland | Business tendency surveys (retail trade) > Order intentions or Demand > Future tendency > National indicator | LRN, s.a. | 16 | -5.414331e-03 |