

CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Economic Studies



**Noise reduction and feature extraction
with principal component analysis for
cryptocurrency price modeling**

Bachelor's thesis

Author: Tomáš Barhoň

Study program: Economics and Finance

Supervisor: prof. PhDr. Ladislav Křišťoufek Ph.D.

Year of defense: 2024

Declaration of Authorship

The author hereby declares that he or she compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis in whole or in part and agrees with the thesis being used for study and scientific purposes.

Prague, February 28, 2024

Tomas Barhon

Abstract

The abstract should concisely summarize the contents of a thesis. Since potential readers should be able to make their decision on the personal relevance based on the abstract, the abstract should clearly tell the reader what information he can expect to find in the thesis. The most essential issue is the problem statement and the actual contribution of described work. The authors should always keep in mind that the abstract is the most frequently read part of a thesis. It should contain at least 70 and at most 120 words (200 when you are writing a thesis). Do not cite anyone in the abstract.

JEL Classification	F12, F21, F23, H25, H71, H87
Keywords	keywordone, keywordtwo, keywordthree, keywordfour
Title	Noise reduction and feature extraction with principal component analysis for cryptocurrency price modeling
Author's e-mail	tomas.barhon@hotmail.cz
Supervisor's e-mail	ladislav.kristoufek@fsv.cuni.cz

Abstrakt

Nutnou součástí práce je anotace, která shrnuje význam práce a výsledky v ní dosažené. Anotace práce by neměla být delší než 200 slov a píše se v jazyce práce (tj. česky, slovensky či anglicky) a v překladu (tj. u anglicky psané práce česky či slovensky, u česky či slovensky psané práce anglicky). Anotace práce by neměla být delší než 200 slov a píše se v jazyce práce (tj. česky, slovensky či anglicky) a v překladu (tj. u anglicky psané práce česky či slovensky, u česky či slovensky psané práce anglicky). V abstraktu by se nemělo citovat.

Klasifikace JEL	F12, F21, F23, H25, H71, H87
Klíčová slova	klicjedna, klicdva, klictri, klicctyri
Název práce	Redukce šumu a extrakce rysů pomocí analýzy hlavních komponent pro modelování cen kryptoměn
E-mail autora	tomas.barhon@hotmail.cz
E-mail vedoucího práce	ladislav.kristoufek@fsv.cuni.cz

Acknowledgments

The author is grateful especially to prof. PhDr. Ladislav Křišťoufek Ph.D., prof. Lars Hansen, prof. David Zilberman, and participants at several conferences and seminars for their comments. The usual caveat applies.

Slightly modified ?? has been awarded the Czech Economic Society prize for young authors.

Typeset in FSV L^AT_EX template with great thanks to prof. Zuzana Havrankova and prof. Tomas Havranek of Institute of Economic Studies, Faculty of Social Sciences, Charles University.

Bibliographic Record

Barhon, Tomas: *Noise reduction and feature extraction with principal component analysis for cryptocurrency price modeling*. Bachelor's thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2024, pages ??. Advisor: prof. PhDr. Ladislav Křišťoufek Ph.D.

Contents

List of Tables

List of Figures

Acronyms

FDI Foreign Direct Investment

MNC Multinational Company

OFDI Outward Foreign Direct Investment

OLI Ownership, Location, Internalization

Bachelor's Thesis Proposal

Author	Tomáš Barhoň
Supervisor	prof. PhDr. Ladislav Krištoufek Ph.D.
Proposed topic	Noise reduction and feature extraction with principal component analysis for cryptocurrency price modeling

Motivation Crypto assets have always been exceptionally volatile compared to traditional assets such as stocks or gold. The historical window is relatively short, thus modeling their price or volatility proposes quite a difficult challenge. It is generally believed that noise in any data decreases the precision of predictions. This effect might be reduced, which will improve the performance of traditional models that are used for cryptocurrency price modeling.

The main motivation for researching this topic is that there is still an ongoing discussion about the role of different features in crypto pricing dynamics. (Kukacka; Kristoufek 2023) have shown that a lot of the pricing dynamic emerges from complex interactions between fundamental and speculative components. They also show the different correlations between all of the explanatory variables which have a direct connection to principal component analysis. It is crucial to study the real impact of those variables in different models as many of them might turn out to be obsolete.

There is currently little use of this dimensionality reduction technique in the academic literature about cryptocurrencies. However, for more traditional financial series this technique is already quite established as a preprocessing technique to reduce noise and dimensionality from which financial data inherently suffer (Chowdhury, U. ; Chakravarty, S. and Hossain, M. 2018). Moreover (Bouri, E.; Kristoufek, L.; Ahmad, T. et al. 2022) studied the effect of microstructural noise on idiosyncratic volatility in cryptocurrencies which further supports the need for a technique that will mitigate this effect on the predictions.

The research will address the problem of variable selection for different types of predictive models with respect to the analysis of the principle components aiming to reduce the dimensionality and simultaneously increase precision. The second question is whether it is more appropriate to transform the high dimensionality with

PCA into lower dimensionality or simply omit the variables with high multicollinearity from the models. These approaches are fundamentally different and the answer is not clear.

Hypotheses

Hypothesis #1: The literature estimating gasoline demand elasticities is affected by publication bias.

Hypothesis #2: The publication bias exaggerates the mean reported elasticity.

Hypothesis #3: The extent of publication bias decreases in time.

Methodology The data will come from various sources because the aim is to look at all the possible variables even if they might not seem useful at first glance. As already mentioned the dynamic is driven by a lot of completely different effects. Most of it will be collected from: coinmetrics.io, studio.glassnode.com, and for macroeconomic indicators <https://fred.stlouisfed.org/>. Some of the data might need to be interpolated to daily observations. Lastly, the observations will need to be sliced to different window sizes and shifted by one so that the predictions can be made for the next day with the data available on that day.

Afterward, the multicollinearity in the data will be examined and different approaches to solve it will be used. The two main ones are using only a smaller set of uncorrelated variables (simple dimensionality reduction) and the second being employing PCA transformation to preserve a predefined threshold of variance or directly targeting the number of principal components.

All the different setups will be compared across three models: linear regression, SVM, and LSTM neural network. The hypothesis is that PCA transformation will substantially lower the measured errors for linear regression and SVM although for LSTM it will result in lower performance as it will only decrease the capacity of the model as the model is powerful enough to create such uncorrelated features without the PCA transformation.

Expected Contribution Existing research agrees that financial data and especially cryptocurrency data are significantly affected by noise. The main goal is to extend the research on the topic of variable selection for algorithmic trading models as there are still a lot of unanswered questions. It will most likely become clearer which approach to dimensionality reduction is the most efficient concerning cryptocurrencies.

Also, not only the underlying pricing dynamics will be detected but the results can be used for investors that are trying to lower their risk of loss which is rela-

tively high in crypto markets. The effect of having a more stable and precise model might significantly cut the transaction costs that are associated with more frequent exchanges as the predictions will become less volatile. That is a desirable property needed to maximize profit and increase credibility towards its customers.

Outline

1. Motivation: there are meta-analyses on the price elasticity of gasoline demand, but they do not correct their estimates for publication bias. Publication bias has been shown to distort most areas of empirical economics, so there is a good chance it will be important here as well.
2. Studies on gasoline demand: I will briefly describe how people estimate the price elasticity of gasoline demand.
3. Data: I will explain how I will collect estimates from studies estimating the elasticity.
4. Methods: I will explain modern meta-analysis methods, including the funnel asymmetry test, precision effect test, and multilevel variants of these regressions.
5. Results: I will discuss my baseline regressions and robustness checks.
6. Concluding remarks: I will summarize my findings and their implications for policy and future research.

Core bibliography

Ashenfelter, O., Harmon, C., Oosterbeek, H., 1999. A review of estimates of the schooling/earnings relationship, with tests for publication bias. *Labour Econ.* 6 (4), 453-470.

Brons, M., Nijkamp, P., Pels, E., Rietveld, P., 2008. A meta-analysis of the price elasticity of gasoline demand. A SUR approach. *Energy Econ.* 30 (5), 2105-2122.

Card, D., Kluve, J., Weber, A., 2010. Active labour market policy evaluations: a meta-analysis. *Econ. J.* 120 (548), F452-F477.

Card, D., Krueger, A.B., 1995. Time-series minimum-wage studies: a meta-analysis. *Am. Econ. Rev.* 85 (2), 238-243.

Doucoulagos, H., Stanley, T.D., 2009. Publication selection bias in minimum-wage research? A meta-regression analysis. *Br. J. Ind. Relat.* 47 (2), 406-428.

- Espey, M., 1998. Gasoline demand revisited: an international meta-analysis of elasticities. *Energy Econ.* 20 (3), 273-295.
- Goldfarb, R.S., 1995. The economist-as-audience needs a methodology of plausible inference. *J. Econ. Methodol.* 2 (2), 201-222.
- Havranek, T., 2010. Rose effect and the Euro: is the magic gone? *Rev. World Econ.* 146 (2), 241-261.
- Havranek, T., Irsova, Z., 2011. Estimating Vertical Spillovers from FDI: Why Results Vary and What the True Effect Is. *J. Int. Econ.* 85 (2), 234-244.
- Horvathova, E., 2010. Does environmental performance affect financial performance? A meta-analysis. *Ecol. Econ.* 70 (1), 52-59.
- Rosenthal, R., 1979. The "file drawer" problem and tolerance for null results. *Psychol. Bull.* 86, 638-641.
- Stanley, T.D., 2001. Wheat from Chaff: meta-analysis as quantitative literature review. *J. Econ. Perspect.* 15 (3), 131-150.
- Stanley, T.D., 2005. Beyond publication bias. *J. Econ. Surv.* 19 (3), 309-345.
- Stanley, T.D., Doucouliagos, H., Jarrell, S.B., 2008. Meta-regression analysis as the socioeconomics of economics research. *J. Socio-Econ.* 37 (1), 276-292.
- Stanley, T.D., Jarrell, S.B., 1989. Meta-regression analysis: a quantitative method of literature surveys. *J. Econ. Surv.* 3 (2), 161-170.

Chapter 1

Introduction

This document serves two purposes. First, it is a template and example for a master’s thesis. Second, the text in all sections contains some useful information on structuring and writing your thesis.

The introduction should consist of three parts (as paragraphs, not to be structured into multiple headings): The first part deals with the background of the work and describes the field of research. It should also elaborate on the general problem statement and the relevance. The second part should describe the focus of the thesis, typically the paragraph starts with a phrase like “The objective of this thesis is” The last part should describe the structure of the thesis, for instance in the following manner. The thesis is structured as follows: ?? cites some formal requirements of the faculty, ?? gives some hints on basic formatting features and covers also acronyms, figures, boxes and tables. ?? gives a recommendation on the usage of hyphens in English language in L^AT_EX and explains how to use the itemize and quote environments and shows a few enumerate-based environments. ?? presents a checklist of common mistakes to avoid. ?? contains numerous hints. ?? summarizes our findings.

Chapter 2

Title of Chapter Two

2.1 Formal requirements of master's thesis at the Faculty of Social Sciences

According to Dean's Provision no. 18/2017:

- The minimum extent of master's thesis is 60 standard pages (108 thousand characters including spaces) of the text itself, i.e. without an abstract and appendices and a list of literature. In case the master's thesis is written in English, its minimum extent is 50 standard pages (90 thousand characters including spaces) without an abstract and appendices and a list of literature. For bachelor's thesis, these requirements go down to 30 standard pages in any language. When writing a standard text document, the minimum requirement is 60 characters per line and 30 lines per page, i.e. 1,800 characters per page (the so-called standard page). Font size, page layout, margins, and line spacing need to be customized.
- Generally, a standard form of the page of the final thesis applies the fonts of 12 points, the gaps between the paragraphs are recommended to be of the size of 6 points. Notes and footnotes can be written in a 10-point font. The text is aligned on both sides (aligned to a block). Electronic version of the thesis will be entered by a student/applicant for a state doctoral examination through the SIS website interface in the archive format of PDF/A version 1.3 or higher. Further details are stipulated by the rector's provision.

Chapter 3

Title of Chapter Three

3.1 Citations

Text text text text text text text text text text text text. Text text text text text text text text text. Text text text text text text text text text text text text text text text. Text text ?.

Text text text text text text text text text text text text text. Text text text text text text text (see, *inter alia*, ?, pg. 10).

3.2 Acronyms

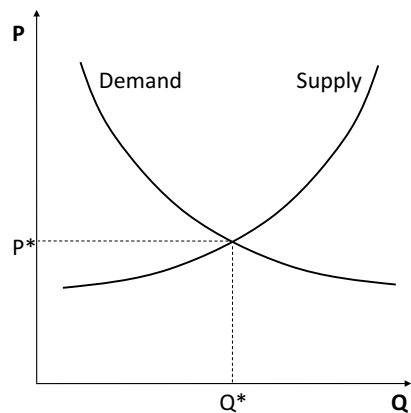
Text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text. Politicians usually like inward **FDI!** (**FDI!**) and an **MNC!** (**MNC!**) appreciates **FDI!** subsidies. Are **MNC!**s greedy?

3.3 Figures

To achieve compatibility with PDF/A 2u, your file must not include links to external fonts, audio, video, or scripts. On the other hand, your file must declare each color environment you use, it must include all the pictures/figures either in jpeg or PDF/A 2u format, used fonts compliant under Unicode (your file cannot use any external fonts), and it must include meta-data in XMP format.

Most troubleshooting comes from the conversion of figures to compliant formats. You can convert from simple PDF using Adobe Acrobat:

Figure 3.1: Market equilibrium



Source: ?.

Look at the ??. Text text text text text text text text text. Text text text text text. Text text text text text text text text text text. Text text text text text text text text text. Text text text text text text text text text.

3.4 Tables

If you use Stata, you might want to check the `sutex`, `outtable`, `outtex`, and `estout` tools, which help you with exporting Stata tables to L^AT_EX.

Table 3.1: Model’s predictions

<i>Case</i>	Y_1	Y_2	τ_1	τ_2	a	n
CR—Slovakia	10.9	10	0.24	0.19	1,000	2.16
CR—Poland	13.3	12	0.24	0.19	1,000	0.38
CR—Hungary	10.4	8	0.24	0.16	1,000	1.10

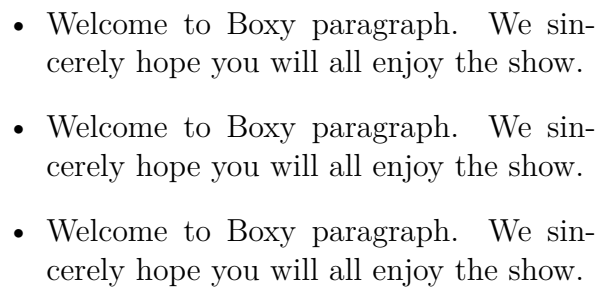
Source: If the source is author himself (like a calculation output), this line is redundant.

Text text text text text text text text text text text text. Text text text text text text text text text. Text text text text text text. Text text text text text text text text text. Text text text text text text text text text text text.

3.5 Boxes

Text text text text text text text text text text text text. Text text text text text text text text text. Text text text text text text. Text text text text text text text text text. Text text text text text text text text text text. Let us make a box:

Figure 3.2: Boxy’s example

- 
- Welcome to Boxy paragraph. We sincerely hope you will all enjoy the show.
 - Welcome to Boxy paragraph. We sincerely hope you will all enjoy the show.
 - Welcome to Boxy paragraph. We sincerely hope you will all enjoy the show.

Source: ?

Text text text text text text text text text text text text. Text text text text text text text text text. Text text text text text text. Text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text text text text text.

3.6 Theorems, Definitions, . . .

Definition 3.1 (My original definition). This is a definition.

Assumption 3.1 (My realistic assumption). This is an assumption.

Proposition 3.1 (My clever proposition). *This is a proposition.*

Lemma 3.1 (My useful lemma). *This is a lemma.*

Example 3.1. This is an example.

Proof. This is a proof.

□

3.7 Equations

3.7.1 Nonumbered Equations

Text text text text text text text text text text text text. Text text text text text text text text text. Text text text text text text. Text text text text text text text text text text. Text text text text text text text text text text text.

$$U = \underbrace{\int_0^\infty \frac{1}{1-\sigma} (C^{1-\sigma} - 1) e^{-\rho t} dt}_{\text{meaning of life}}$$

3.7.2 Numbered Equations

Text text text text text text text text text text text text. Text text text text text text text text text. Text text text text text text. Text text text text text text text text text text. Text text text text text text text text text text text.

$$U = \int_0^\infty \overbrace{\frac{1}{1-\sigma} (C^{1-\sigma} - 1)}^{\text{instantaneous utility}} e^{-\rho t} dt \quad (3.1)$$

3.7.3 Matrix Equations

Text text text text text text text text text text text text. Text text text text text text text text text. Text text text text text text. Text text text text text text text text text text. Text text text text text text text text text text text.

$$\mathbf{A} = \mathbf{B} + \mathbf{C} \quad (3.2)$$

3.8 Cross-references

- to literature (?, pg. 10) or ?, pg. 10,
- to ??,
- see ??,
- to ??,

- to Definition ??, to Proposition ??, Example ??,
- to equations like this: see (??).

3.9 Source codes

You can input a source code like this:

```
omega = 1;
syms zeta;
jmn = [1 2*zeta*omega omega^2];
figure(1);
    for zeta = 1E-5 : 0.2 : 1+1E-12
        G = tf(omega^2,subs([1 2*zeta*omega omega^2]));
        bode(G); hold on;
    end
legend('\zeta = 0', '\zeta = 0,2', '\zeta = 0,4', '\zeta = 0,6','');

```

Should you prefer a different font size, redefine file `Styles/Mystyle.sty`.

3.10 Paragraphs

Usually you should not use the first person singular (I) in your text, write we instead. As a general recommendation, use the first person sparsely, sometimes it can be replaced by a phrase like “This work presents . . .”

Text text text text text text text text text text text text text. Text
text text text text text text text text text. Text text text text text text. Text
text text text text text text text text text. Text text text text text text text
text text text. Text text text text text text (?). Let us make two paragraphs:

Proin Text text text text text text text text text text text text.
Text text text text text text text text text. Text text text text text text.
Text text text text text text text text text. Text text text text text text
text text text text. Text text text text text text text text text text text
text text text. Text text text text text text text text text. Text text text
text text text. And a subparagraph:

Velit Text text text text text text text text text text text text. Text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text text. Text text text text text text text text text text.

begin tabular llllllllll
top rule BTC-LR - 1 day BTC-LR - 5 days BTC-LR - 10 days BTC-SVR - 1
day BTC-SVR - 5 days BTC-SVR - 10 days BTC-LSTM - 1 day BTC-LSTM
- 5 days BTC-LSTM - 10 days

mid rule dimensionality 960.371278 2392.049220 3765.643891 973.251758 2375.766866
3187.232878 3332.288921 9794.555404 7699.674445

Chapter 4

Title of Chapter Four

4.1 Title of Section One

Many people use simple n-dash in many occasions – like this –, where however typographic convention—it looks a bit strange at first sight—requires m-dash. Text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text ?.

Text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text ?.

Let us describe the following animals:

Item 1 Text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text text.

Item 2 Text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text text.

Text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text text text. Text

Chapter 5

Title of Chapter Five

5.1 Title of Section One

The following checklist should help in avoiding some frequently made mistakes, if any of the following propositions apply for your thesis, there is a problem:

- You have citations in your abstract.
- The introduction does not cover the three parts as described in ??.
- The introduction contains subheadings.
- You described different aspects than promised in the title.
- You copied some parts of the text from other work without proper referencing and citing.
- You used automatic translation tools to produce text by translating it from another language.
- Your thesis contains many typos and grammatical errors. (Use an electronic spell checker. Please!)
- You used color in your figures and refer to the “blue” line (assume that your readers use a monochrome printer).
- You mainly used websites and other unrefereed material as your sources or you used Wikipedia as your source.
- You refer to something in your conclusion which you have not mentioned before.

-
- Some forenames in the references are abbreviated, some not.
 - Some references miss a publishing date.

Chapter 6

Useful Hints

If you write in English, you might find the following hint useful: The indefinite article *a* is used as an before a vowel sound—for example an apple, an hour, an unusual thing, an **MNC!** (because the acronym is pronounced Em-En-See). Before a consonant sound represented by a vowel letter *a* is usual—for example a one, a unique thing, a historic chance. Few more tips to follow:

- Don't give orders—don't write in the imperative mood—unless you are training to be a teacher.
- Avoid the use of questions. You may know the answer: does your reader? It's much safer to tell her, or him.
- Do not become entangled in the problems of 'sexist' language. It is much easier to write in the plural. "Students should check their work" is good English. "A student should check—" is also good English, but now the problems begin: "—her work?" "—his work?" Which? You can write "his or her," but that seems clumsy. Stick to the plural.
- If you must refer to yourself, use the third person such as "The present writer would recommend that . . ." may be useful.
- Use the full forms of words and phrases, not contractions like "he's," "don't," etc. Keep the apostrophe to indicate possession—and use it correctly. Academics really sneer at students who use the "Greengrocer's apostrophe."
- Do not despise short, workmanlike, and effective plain English words. If they mean what you want to say. Accurately.

- Avoid the use of humor in academic writing—unless you are very sure of yourself.
- Even when you are not being funny, avoid the use of irony or sarcasm.
- Paragraphs in academic English should contain more than one sentence. (Short paragraphs look as if you are writing for a tabloid newspaper—or a simple Template!) I guess that the average academic book runs to two or three paragraphs per page. Look at the books in your subject, and get a feel for how long your own paragraphs should be when you are imitating the academic style.
- Use the word that more in formal writing than most of us do in speech—particularly after such verbs of utterance as to say, to report, to think etc. It can help to make your writing much clearer.
- Develop an academic vocabulary. The ‘long words’ you learn in the course of your studies are long usually because they have more precise meanings than their less formal equivalents. They are therefore better when you want to be accurate. (Also they allow you to sound like someone who deserves a degree.)
- Use as few words as you can; but use enough words to express your meaning as fully as you can. Your judgment of what is appropriate here is part of what you should learn throughout your course.
- Avoid lazy words such as “nice”. It is usually better to say “acquire” or “obtain” than “get;” and it may be better, if you mean “through the use of money,” to say “purchase” or—better still—“buy.”
- A short word like “buy” is better than a long one like “purchase”—unless the long one is more accurate. A “statutory instrument” is better than a “rule”—to a lawyer, at any rate.
- Proof-read with care. Ask someone else to help—you may be too close to your work to be able to see your mistakes.
- If in doubt, choose the more formal, or possibly just the more old-fashioned, of two words. For example, say quotation rather than quote whenever you mean the use of somebody else’s words.

-
- You will often sound more academic if you include doubts in your work—and qualifications. Within the scope of this thesis, the current writer cannot hope to cover all the possible implications of the question.Ô
 - In this context, the use of litotes sounds very academic. This is the construction where a writer uses a negative with a negative adjective, e.g. it is not unlikely that ... This does not mean the same as it is probable that ... It has a shade of meaning and qualification that can be useful to academic writers.

Chapter 7

Conclusion

The conclusion should briefly summarize the problem statement and the general content of the work and the emphasize on the main contribution of the work.

When writing the conclusion keep in mind that some readers may not have gone through the whole thesis, but have jumped directly to the conclusion after having read the abstract in order the decide on the personal relevance of the thesis. Therefore, the conclusion should be self contained, which means that a reader should be able to understand the essence of the conclusion without having to read the whole thesis.

The conclusion typically ends with an outlook that describes possible extensions of the presented approaches and of planned future work.

Bibliography

BLOMSTROM, M. & A. KOKKO (2003): “The Economics of Foreign Direct Investment Incentives.” *NBER Working Papers 9489*, National Bureau of Economic Research, Inc.

HAAPARANTA, P. (1996): “Competition for Foreign Direct Investment.” *Journal of Public Economics* **63**(1): pp. 141–53.

HAUFLER, A. & I. WOOTON (2006): “The Effects of Regional Tax and Subsidy Coordination on Foreign Direct Investment.” *European Economic Review* **50**(2): pp. 285–305.

WELLS, L. T., N. ALLEN, J. MORISSET, & N. PIRNIA (2001): *Using Tax Incentives to Compete for Foreign Investment: Are They Worth the Cost?* Washington, DC: FIAS.

Appendix B

Content of Enclosed DVD

This is optional: you may enclose a DVD to this thesis which contains empirical data and MatLab/R/Stata source codes. Even better so, you can create a special website for your project. Stating in your thesis that the data and source codes are available upon request is enough but please, have them prepared for such requests.

- Folder 1: Source codes
- Folder 2: Empirical data