University of Duisburg-Essen
Faculty of Business Administration and Economics
Chair of Econometrics

P-Approximation

Seminar in Econometrics

Term Paper

Submitted to the Faculty of
Business Administration and Economics
at the
University of Duisburg-Essen

from:

Jens Klenke and Janine Langerbein

Reviewer:	Christoph Hanck
Deadline:	Jan. 17th 2020

Name:	Jens Klenke	Janine Langerbein
Matriculation Number:	3071594	307
E-Mail:	jens.klenke@stud.uni-due.de	janine.langerbein@stud.uni-due.de
Study Path:	M.Sc. Economics	M.Sc. Economics
Semester:	5th	5th
Graduation (est.):	Winter Term 2020	Winter Term 2020

Contents

List of Figures

List of Tables

List of Abbreviations

1 Introduction

Meta tests have been shown to be a powerful tool when testing for the null of non-cointegration. The distribution of their test statistic, however, is mostly not available in closed form. This might pose difficulties when implementing the meta tests in econometric software packages, as one has to include tables of critical values and p-values for each combination of the underlying tests. Software package size limitations are therefore quickly exceeded.

In this paper we propose supervised Machine Learning Algorithms to approximate the p-values of the meta test by Bayer and Hanck (2012) which tests for the null of non-cointegration. This approach might reduce the size of associated software packages considerably. The algorithms are trained on simulated data for various specifications of the aforementioned test.

Ergebnis der Models (1-2 Sätze)

Inhalt Paper

2 Bayer Hanck Test

The choice as to which of the available cointegration tests to use is a recurrent issue in econometric time series analysis. Bayer and Hanck (2012) propose powerful meta tests which provide unambiguous test decisions. They combine several residual- and system-based tests in the manner of Fisher's (1932) Chi-squared test.

Bayer and Hanck build their work on results from Pesavento (2004), who defines the underlying model as $z_t' = [x_t', y_t]$. x_t, an $n_1 \times 1$ vector, describes the regressor dynamics, while y_t is a scalar which defines the cointegrating relation. They can be written as

$$\Delta x_t = \tau_1 + v_{1t}, \tag{2.1}$$

$$y_t = (\mu_2 - \gamma'\mu_1) + (\tau_2 - \gamma'\tau_1)t + \gamma' x_t + u_t, \tag{2.2}$$

$$u_t = \rho u_{t-1} + v_{2t}. \tag{2.3}$$

μ_1, μ_2 τ_1 and τ_2 are the deterministic parts of the model. They are subject to the following restrictions: (i) $\mu_2 - \gamma'\mu_1$ and $\tau = 0$ which translates to no deterministics, (ii) $\tau = 0$ which corresponds to a constant in the cointegrating vector, (iii) $\tau_2 - \gamma'\tau_1 = 0$, a constant plus trend.

$v_t = [v_{1t}' v_{2t}]'$ with Ω the long-run covariance matrix of v_t. For derivation of v_t see Pesavento (2004). Pesavento shows that $\{v_t\}$ satisfies an FCLT, i.e. $T^{-1/2} \sum_{t=1}^{[T\cdot]} v_t \Rightarrow \Omega^{1/2} W(\cdot)$. It is further assumed that the x_t are not cointegrated.

It clearly follows from (2.3) that z_t is cointegrated if $\rho < 1$. Hence the null hypothesis of no cointegration is $H_0 : p = 1$.

Furthermore, Pesavento introduces two other parameters. First, R^2 measures the squared correlation of v_{1t} and v_{2t}. It can be interpreted as the influence of the right-hand side variables in (2.2). It ranks between zero and one. When there is no long-run correlation between those variables and the errors from the cointegration regression, R^2 equals zero. Secondly, the number of lags is approximated by a finite number k.

Assumptions (BH S. 84)?

Bayer and Hanck's (2012) meta test combines the test statistics of four stand-alone tests. Namely, these are the tests of Engle and Granger (1987), Johansen (1988), Boswijk (1994) and Banerjee et al. (1998). For the sake of brevity the detailed derivation of the underlying tests has been deliberately omitted here.

Engle and Granger (1987) propose a two-step procedure to test the null hypothesis of no cointegration against the alternative of at least one cointegrating vector. First, the long-run relationship between y_t and $\mathbf{x}_t$ is estimated by least squares regression. The obtained residuals $\hat{u}_t$ are then tested for a unit root. For this, Engle and Granger suggest the use of the t-statistic t_γ^{ADF} in the Augmented Dickey-Fuller (ADF) regression:

$$\Delta \hat{u}_t = \gamma \hat{u}_{t-1} + \sum_{i=1}^{k} \pi_i \Delta \hat{u}_{t-i} + \varepsilon_t. \tag{2.4}$$

The rejection of a unit root points to a cointegration relationship.

Johansen's (1988) maximum eigenvalue test is a system-based test that allows for several cointegration relationships. Take the vector error correction model (VECM)

$$\Delta \mathbf{z}_t = \mathbf{\Pi} \mathbf{z}_{t-1} + \sum_{i=1}^{k} \mathbf{\Gamma}_p \Delta \mathbf{z}_{t-p} + \mathbf{d}_t + \varepsilon_t. \tag{2.5}$$

blabla Johansen test statistic

Banerjee and Boswijk

To combine the results from the underlying tests Bayer and Hanck draw upon Fisher's combined probability test (Fisher, 1932). It merges the tests using the formula

$$\tilde{\chi}^2_{\mathcal{I}} := -2 \sum_{i \in \mathcal{I}} \ln{(p_i)}. \tag{2.6}$$

Let t_i be the i^{th} test statistic. If test i rejects for large values, take $\xi_i := t_i$. If test i rejects for small values, take $-\xi_i := t_i$. With $\Xi_i(x) := \Pr_{\mathcal{H}_\prime}(\xi_i \geq x)$ the p-value of the i^{th} test is $p_i := \Xi_i(\xi_i)$.

Fisher (1932) shows that under the assumption of independence the null distribution of $\tilde{\chi}^2_{\mathcal{I}}$ follows a chi-squared distribution with $2\mathcal{I}$ degrees of freedom. If this assumption is violated the null distribution is less evident. Here, the latter case occurs, as the ξ_i are not independent. The $\tilde{\chi}^2_{\mathcal{I}}$, however, have well-defined asymptotic null distributions $F_{\mathcal{F}_{\mathcal{I}}}$, as $\tilde{\chi}^2_{\mathcal{I}} \to_d \mathcal{F}_{\mathcal{I}}$ under $\mathcal{H}_0$ if $T \to \infty$, with $\mathcal{F}_{\mathcal{I}}$ some random variable. It is therefore feasible to simulate the joint null distribution of the ξ_i to obtain the distribution $F_{\mathcal{F}_{\mathcal{I}}}$ of 2.6. The $F_{\mathcal{F}_{\mathcal{I}}}$ depend on which and how many tests are combined. The distributions of the ξ_i depend on $K - 1$ and the deterministic case.

3 Simulation

In this section, we describe the simulation of the null distribution of the Bayer Hanck meta test. The objective is to obtain data for training machine learning algorithms on approximating the p-values of the aforementioned test. In consideration of the different forms of the meta test we generated six data sets. These vary according to the specific combination of the underlying tests and also account for the above-mentioned restrictions on the deterministic parts of the model.

This simulation relies largely on previous work by Pesavento (2004). We consider $\mathrm{R}^2 \in \{0, 0.05, 0.1, ..., 0.95\}$, $k = 11$ and $c = 0$[1] and set the number of repetitions to 1,000,000. N? c vielleicht mal definieren

To calculate the Bayer Hanck test statistic we first simulate the null distributions of the underlying test statistics. It can be shown that asymptotically these are non-standard but a function of standard Brownian motions. The latter is approximated by step functions using Gaussian random walk with

[1] Since we solely aim at simulating the distribution of the null of no cointegration we will not consider any further values of c here.

$N = 1000$ observations. Referenz Theorem? OU Prozess? Nochmal auf Transformation je nach case eingehen? p values durch cdf, Fisherstat + pvalues berechnen.

4 Models

5 Package

References

Banerjee, A., Dolado, J., & Mestre, R. (1998). Error-correction mechanism tests for cointegration in a single-equation framework. *Journal of Time Series Analysis*, *19*(3), 267–283. https://EconPapers.repec.org/RePEc:bla:jtsera:v:19:y:1998:i:3:p:267-283

Bayer, C., & Hanck, C. (2012). Combining non-cointegration tests. *Journal of Time Series Analysis*.

Boswijk, H. P. (1994). Testing for an unstable root in conditional and structural error correction models. *Journal of Econometrics*, *63*(1), 37–60. https://EconPapers.repec.org/RePEc:eee:econom:v:63:y:1994:i:1:p:37-60

Engle, R., & Granger, C. W. (1987). Co-integration and error correction: Representation, estimation and testing. *Econometrica*, *55*, 251–276.

Fisher, R. A. (1932). *Statistical methods for research workers*. Oliver; Boyd, Edinburgh; London.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, *12*(2), 231–254. https://doi.org/https://doi.org/10.1016/0165-1889(88)90041-3

Pesavento, E. (2004). Analytical evaluation of the power of tests for the absence of cointegration. *Journal of Econometrics*, *122*(2), 349–384.

Software-References

Breiman, L., Cutler, A., Liaw, A., & Wiener., M. (2018). *Randomforest: Breiman and cutler's random forests for classification and regression* [R package version 4.6-14]. https://CRAN.R-project.org/package=randomForest

Croissant, Y., Millo, G., & Tappe, K. (2019). *Plm: Linear models for panel data* [R package version 2.1-0]. https://CRAN.R-project.org/package=plm

Friedman, J., Hastie, T., Tibshirani, R., Simon, N., Narasimhan, B., & Qian, J. (2019). *Glmnet: Lasso and elastic-net regularized generalized linear models* [R package version 2.0-18]. https://CRAN.R-project.org/package=glmnet

Greenwell, B., Boehmke, B., Cunningham, J., & Developers, G. (2019). *Gbm: Generalized boosted regression models* [R package version 2.1.5]. https://CRAN.R-project.org/package=gbm

Henry, L., & Wickham, H. (2019). *Purrr: Functional programming tools* [R package version 0.3.2]. https://CRAN.R-project.org/package=purrr

Hlavac, M. (2018). *Stargazer: Well-formatted regression and summary statistics tables* [R package version 5.2.2]. https://CRAN.R-project.org/package=stargazer

Izrailev, S. (2014). *Tictoc: Functions for timing r scripts, as well as implementations of stack and list structures.* [R package version 1.0]. https://CRAN.R-project.org/package=tictoc

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., & Hunt., T. (2019). *Caret: Classification and regression training* [R package version 6.0-84]. https://CRAN.R-project.org/package=caret

Lumley, T., & Miller, A. (2017). *Leaps: Regression subset selection* [R package version 3.0]. https://CRAN.R-project.org/package=leaps

Mevik, B.-H., Wehrens, R., & Liland, K. H. (2019). *Pls: Partial least squares and principal component regression* [R package version 2.7-1]. https://CRAN.R-project.org/package=pls

Milborrow, S. (2019a). *Plotmo: Plot a model's residuals, response, and partial dependence plots* [R package version 3.5.5]. https://CRAN.R-project.org/package=plotmo

Milborrow, S. (2019b). *Rpart.plot: Plot 'rpart' models: An enhanced version of 'plot.rpart'* [R package version 3.0.7]. https://CRAN.R-project.org/package=rpart.plot

R Core Team. (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Ripley, B. (2019a). *Class: Functions for classification* [R package version 7.3-15]. https://CRAN.R-project.org/package=class

Ripley, B. (2019b). *Mass: Support functions and datasets for venables and ripley's mass* [R package version 7.3-51.4]. https://CRAN.R-project.org/package=MASS

Ripley, B. (2019c). *Tree: Classification and regression trees* [R package version 1.0-40]. https://CRAN.R-project.org/package=tree

RStudio Team. (2019). *Rstudio: Integrated development environment for r* [Version 1.2.1541]. RStudio, Inc. Boston, MA. http://www.rstudio.com/

Rushworth, A. (2019). *Inspectdf: Inspection, comparison and visualisation of data frames* [R package version 0.0.4]. https://CRAN.R-project.org/package=inspectdf

Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., & Despouy, P. (2019). *Plotly: Create interactive web graphics via 'plotly.js'* [R package version 4.9.0]. https://CRAN.R-project.org/package=plotly

Therneau, T., & Atkinson, B. (2019). *Rpart: Recursive partitioning and regression trees* [R package version 4.1-15]. https://CRAN.R-project.org/package=rpart

Ushey, K., Allaire, J., Wickham, H., & Ritchie, G. (2019). *Rstudioapi: Safely access the rstudio api* [R package version 0.10]. https://CRAN.R-project.org/package=rstudioapi

Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string operations* [R package version 1.4.0]. https://CRAN.R-project.org/package=stringr

Wickham, H., François, R., Henry, L., & Müller, K. (2019). *Dplyr: A grammar of data manipulation* [R package version 0.8.0.1]. https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2019). *Tidyr: Easily tidy data with 'spread()' and 'gather()' functions* [R package version 0.8.3]. https://CRAN.R-project.org/package=tidyr

Xie, Y. (2019). *Knitr: A general-purpose package for dynamic report generation in r* [R package version 1.23]. https://CRAN.R-project.org/package=knitr

A Appendices

Number	Functional form	Range of γ
1	$p = \text{poly}(t, \gamma) + (1/k)$	$\gamma \in \{3, 4, 5, 6\}$
2	$p = \text{poly}(t, \gamma) + (1/k) + \text{poly}(t, \gamma) * 1/k$	$\gamma \in \{3, 4, 5, 6\}$
3	$p = \text{poly}(t, \gamma) + \log(k) + \text{poly}(k, \gamma) * \log(k)$	$\gamma \in \{3, 4, 5, 6\}$
4	$p = \text{poly}(t, \gamma) + k + (1/k)$	$\gamma \in \{3, 4, 5, 6\}$
5	$p = \text{poly}(\log(t), \gamma) + \log(k)$	$\gamma \in \{3, 4, 5, 6, 8, 9, 10\}$
6	$p = \text{poly}(\log(t), \gamma) * \log(k)$	$\gamma \in \{3, 4, 5, 6, 8, 9, 10\}$

Table A1: Description of all tested models....

Functional form	RMSE	RMSE_cor	RMSE_0.2	RMSE_cor_0.2
$\text{bc}(p) = \text{poly}(\text{bc}(t), 10) * \log(k) + \text{poly}(\text{bc}(t), 10) * \sqrt{k}$	$4.97 \cdot 10^{-4}$	$4.69 \cdot 10^{-4}$	$8.05 \cdot 10^{-4}$	$7.16 \cdot 10^{-4}$
$\text{bc}(p) = \text{poly}(\text{bc}(t), 10) * \log(k) + \text{poly}(\text{bc}(t), 10) * 1/k$	$5.39 \cdot 10^{-4}$	$5.11 \cdot 10^{-4}$	$8.54 \cdot 10^{-4}$	$7.61 \cdot 10^{-4}$
$p = \text{poly}(\text{bc}(t), 10) * \log(k) + \text{poly}(\text{bc}(t), 10) * \sqrt{k}$	$7.68 \cdot 10^{-4}$	$6.91 \cdot 10^{-4}$	$1.01 \cdot 10^{-3}$	$8.97 \cdot 10^{-4}$
$p = \text{poly}(\text{bc}(t), 10) * \log(k) + \text{poly}(\text{bc}(t), 10) * 1/k + \sqrt{k}$	$7.79 \cdot 10^{-4}$	$7.04 \cdot 10^{-4}$	$1.05 \cdot 10^{-3}$	$9.31 \cdot 10^{-4}$
$p = \text{poly}(\text{bc}(t), 10) * \log(k) + \text{poly}(\text{bc}(t), 10) * 1/k$	$7.82 \cdot 10^{-4}$	$7.07 \cdot 10^{-4}$	$1.06 \cdot 10^{-3}$	$9.41 \cdot 10^{-4}$

Table A2: The 5 best models for the combined Non-Cointegration test of Bayer and Hanck (2012), where *all* underlying test are included and case 1.

Functional form	RMSE	RMSE_cor	RMSE_0.2	RMSE_cor_0.2
$\log(p) = \text{poly}(\text{bc}(t), 10) * \log(k)$	$1.27 \cdot 10^{-3}$	$1.25 \cdot 10^{-3}$	$1.05 \cdot 10^{-3}$	$9.52 \cdot 10^{-4}$
$\log(p) = \text{poly}(\text{bc}(t), 10) * \log(k) + \text{poly}(\text{bc}(t), 10) * \sqrt{k}$	$6.82 \cdot 10^{-4}$	$6.22 \cdot 10^{-4}$	$1.28 \cdot 10^{-3}$	$1.12 \cdot 10^{-3}$
$\log(p) = \text{poly}(\text{bc}(t), 10) * \log(k) + \text{poly}(\text{bc}(t), 10) * 1/k$	$7.32 \cdot 10^{-4}$	$6.63 \cdot 10^{-4}$	$1.39 \cdot 10^{-3}$	$1.20 \cdot 10^{-3}$
$\log(p) = \text{poly}(\text{bc}(t), 10) * \log(k) + \text{poly}(\text{bc}(t), 10) * 1/k + \sqrt{k}$	$8.38 \cdot 10^{-4}$	$7.78 \cdot 10^{-4}$	$1.48 \cdot 10^{-3}$	$1.31 \cdot 10^{-3}$
$\text{bc}(p) = \text{poly}(\text{bc}(t), 10) * \log(k) + \text{poly}(\text{bc}(t), 10) * 1/k$	$9.08 \cdot 10^{-4}$	$8.42 \cdot 10^{-4}$	$1.69 \cdot 10^{-3}$	$1.50 \cdot 10^{-3}$

Table A3: The 5 best models for the combined Non-Cointegration test of Bayer and Hanck (2012), where *all* underlying test are included and case 2.

Functional form	RMSE	RMSE_cor	RMSE_0.2	RMSE_cor_0.2
$\log(p) = \text{poly}(\text{bc}(t), 10) * \log(k) + \text{poly}(\text{bc}(t), 10) * \sqrt{k}$	$4.58 \cdot 10^{-4}$	$4.55 \cdot 10^{-4}$	$3.37 \cdot 10^{-4}$	$3.16 \cdot 10^{-4}$
$\log(p) = \text{poly}(\text{bc}(t), 10) * \log(k) + \text{poly}(\text{bc}(t), 10) * 1/k$	$5.17 \cdot 10^{-4}$	$5.14 \cdot 10^{-4}$	$3.904 \cdot 10^{-4}$	$3.73 \cdot 10^{-4}$
$\log(p) = \text{poly}(\text{bc}(t), 10) * \log(k)$	$1.04 \cdot 10^{-3}$	$1.04 \cdot 10^{-3}$	$6.760 \cdot 10^{-4}$	$6.50 \cdot 10^{-4}$
$\log(p) = \text{poly}(\text{bc}(t), 10) * \log(k) + \text{poly}(\text{bc}(t), 10) * 1/k + \sqrt{k}$	$1.18 \cdot 10^{-3}$	$1.17 \cdot 10^{-3}$	$2.06 \cdot 10^{-3}$	$2.05 \cdot 10^{-3}$
$\text{bc}(p) = \text{poly}(\text{bc}(t), 10) * \log(k) + \text{poly}(\text{bc}(t), 10) * 1/k$	$1.16 \cdot 10^{-3}$	$1.06 \cdot 10^{-3}$	$2.08 \cdot 10^{-3}$	$1.80 \cdot 10^{-3}$

Table A4: The 5 best models for the combined Non-Cointegration test of Bayer and Hanck (2012), where *all* underlying test are included and case 3.

Functional form	RMSE	RMSE_cor	RMSE_0.2	RMSE_cor_0.2
$\log(p) = \text{poly}(\text{bc}(t), 10) * \log(k) + \text{poly}(\text{bc}(t), 10) * 1/k$	$4.75 \cdot 10^{-4}$	$4.44 \cdot 10^{-4}$	$7.81 \cdot 10^{-4}$	$6.84 \cdot 10^{-4}$
$\log(p) = \text{poly}(\text{bc}(t), 10) * \log(k)$	$6.54 \cdot 10^{-4}$	$5.87 \cdot 10^{-4}$	$1.01 \cdot 10^{-3}$	$7.81 \cdot 10^{-4}$
$\log(p) = \text{poly}(\text{bc}(t), 10) * \log(k) + \text{poly}(\text{bc}(t), 10) * \sqrt{k}$	$7.60 \cdot 10^{-4}$	$6.13 \cdot 10^{-4}$	$1.46 \cdot 10^{-3}$	$1.06 \cdot 10^{-3}$
$\log(p) = \text{poly}(\text{bc}(t), 10) * \log(k) + \text{poly}(\text{bc}(t), 10) * 1/k + \sqrt{k}$	$7.64 \cdot 10^{-4}$	$7.45 \cdot 10^{-4}$	$1.29 \cdot 10^{-3}$	$1.23 \cdot 10^{-3}$
$\text{bc}(p) = \text{poly}(\text{bc}(t), 10) * \log(k) + \text{poly}(\text{bc}(t), 10) * 1/k$	$1.01 \cdot 10^{-3}$	$9.17 \cdot 10^{-4}$	$1.89 \cdot 10^{-3}$	$1.65 \cdot 10^{-3}$

Table A5: The 5 best models for the combined Non-Cointegration test of Bayer and Hanck (2012), where *EG-J* underlying test are included and case 3.

Eidesstattliche Versicherung

Ich versichere an Eides statt durch meine Unterschrift, dass ich die vorstehende Arbeit selbständig und ohne fremde Hilfe angefertigt und alle Stellen, die ich wörtlich oder annähernd wörtlich aus Veröffentlichungen entnommen habe, als solche kenntlich gemacht habe, mich auch keiner anderen als der angegebenen Literatur oder sonstiger Hilfsmittel bedient habe. Die Arbeit hat in dieser oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen.

Essen, den ____________________   ________________________________

Jens Klenke and Janine Langerbein