Detecting Structural Breaks with a Fusion Penalty

Jana Mareckova
University of Konstanz, Graduation School of Decision Sciences

16th IWH-CIREQ Macroeconometric Workshop,
Halle (Saale),

December 8, 2015

Motivation (1)

- Linear model with time-varying coefficients:

$$y_t = x_t'\beta_t + \varepsilon_t, \qquad t = 1, \ldots, T,$$

  $y_t \ldots$ dependent variable, $x_t \ldots p \times 1$ vector of regressors, $\beta_t \ldots p \times 1$ vector of coefficients, $\varepsilon_t \ldots$ error term.
- Parameters to estimate: Tp
- Number of observations: T

Motivation (2)

- Matrix notation:

$$\underset{T\times 1}{y} = \underset{T\times Tp}{X} \underset{Tp\times 1}{\beta} + \underset{T\times 1}{\varepsilon}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_T \end{pmatrix} = \begin{pmatrix} x_1' & 0 & 0 & \cdots & 0 \\ 0 & x_2' & 0 & \cdots & 0 \\ 0 & 0 & x_3' & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & x_T' \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_T \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_T \end{pmatrix},$$

- Diagonal VC matrix of the error term (dynamics captured correctly in the mean function):

$$V[\varepsilon] = \Sigma.$$

Motivation (3)

- Model:
$$y_t = x_t'\beta_t + \varepsilon_t, \qquad t = 1, \ldots, T,$$
- Assumption: β_t's stable for some period of time, i.e. $m < T$ breaks ($m+1$ different regimes):
$$\beta_{T_{j-1}} = \cdots = \beta_{T_j - 1}, \qquad \text{for } j = 1, \ldots, m,$$
$$\beta_{T_{j-1}} = \cdots = \beta_{T_j}, \qquad \text{for } j = m+1,$$
$T_0 = 1$, $T_{m+1} = T$, $T_j \ldots$ break date.
- Idea: Capture the restrictions in an L_1-norm regularization of a high-dimensional linear model $\Rightarrow$ (weighted) fusion penalty.

(Weighted) Fusion Penalty

- Land and Friedman (1997)
- Penalized least squares:

$$\widehat{\beta}_\lambda = \underset{\beta}{\arg\min} \quad \frac{1}{T}(y - X\beta)'\Sigma^{-1}(y - X\beta) + \lambda P(\beta), \qquad (1)$$

where $P(\beta) = \sum_{t=2}^{T} \sum_{k=1}^{p} w_{t,k} |\beta_{t,k} - \beta_{t-1,k}|$,

$$\lambda \geq 0 \ \ldots \ \text{given tuning parameter,}$$
$$w_{t,k} \ldots \ \text{given non-negative weights.}$$

- Convex in β_t's but non-smooth (for fixed λ and $w_{t,k}$'s).
- An MM algorithm by Yu et al. (2013)

(Weighted) Fusion Penalty - Generalized Error Term

- Procedure for heteroscedastic error terms:
  - First, estimate (1) under homoscedasticity.
  - Model squared residuals on a time-varying constant.
  - Re-estimate the (1) with $\widehat{\Sigma}$ and get new $\widehat{\beta}_\lambda$.
- For now: $\Sigma = \sigma^2 I$.

(Weighted) Fusion Penalty - Weights (1)

- An unweighted fusion penalty ($w_{t,k} = 1, \forall t, k$) is not selecting the true model consistently (Viallon et al., 2013)

$$\limsup_n P(\widehat{B} = B) \leq c < 1,$$

where c is a constant depending on the true model and

$$B = \{(t, k) : \beta_{t,k} \neq 0, \beta_{t,k} \neq \beta_{t-1,k}\},$$

$$\widehat{B} = \{(t, k) : \widehat{\beta}_{t,k} \neq 0, \widehat{\beta}_{t,k} \neq \widehat{\beta}_{t-1,k}\}.$$

(Weighted) Fusion Penalty - Weights (2)

- Therefore, the weighted fusion penalty is implemented (consistent in variable selection)

$$P(\widehat{B} = B) \to 1 \text{ as } n \to \infty.$$

  under a proper choice of weights (Viallon et al., 2013).
- Two step estimation:
  1. Set $w_{t,k} = 1$, get $\widehat{\beta}_{\widehat{\lambda}_1}$ given optimal $\widehat{\lambda}_1$.
  2. Set $w_{t,k} = 1/|\widehat{\beta}_{t,k,\widehat{\lambda}_1} - \widehat{\beta}_{t-1,k,\widehat{\lambda}_1}|$, given optimal $\widehat{\lambda}_2$ get $\widehat{\beta}_{\widehat{\lambda}_2}$.

Selection of λ - Information criteria (1)

EBIC

- Chen and Chen (2008) - Extended BIC (EBIC) for highdimensional models with polynomially increasing number of parameters.

$$EBIC(\lambda) = T\log\left(\frac{1}{T}SSE(\widehat{\beta}_\lambda)\right) + (\log(T) + 2\log(Tp))\,|\widehat{\beta}_\lambda|.$$

- $|\widehat{\beta}_\lambda|$ represents the number of nonzero unique parameters

$$|\widehat{\beta}_\lambda| = \sum_{k=1}^{p} 1(\widehat{\beta}_{1,k,\lambda} \neq 0) + \sum_{t=2}^{T}\sum_{k=1}^{p} 1(\widehat{\beta}_{t-1,k,\lambda} \neq \widehat{\beta}_{t,k,\lambda} | \widehat{\beta}_{t,k,\lambda} \neq 0),$$

- Consistent under certain regularity conditions.

Selection of λ - Information criteria (2)

IC by Qian and Su (2014)

- Qian and Su (2014) - IC for detecting structural breaks with a Frobenius norm penalty in a model with time-varying parameters

$$IC_{QS}(\lambda) = \log\left(\frac{1}{T} SSE(\widehat{\beta}_\lambda)\right) + \rho_T |\widehat{\beta}_\lambda|, \quad \rho_T = 1/\sqrt{T}.$$

- $|\widehat{\beta}_\lambda|$ represents the number of nonzero unique parameters
- Consistent under certain regularity conditions.

Selection of λ - Information criteria (3)

- Observation: *EBIC* and IC_{QS} prone to underfitting
- **1. step - Augmented** IC_{QS}

$$IC_1(\lambda) = \log\left(\frac{1}{T}\sum_{t=1}^{T} SSE(\widehat{\beta}_\lambda)\right) + \frac{1}{\sqrt{T}}(|\widehat{\beta}_\lambda| - c),$$

where

$$c = \sum_{t=2}^{T}\sum_{k=1}^{p} 1\Big\{|\widehat{\beta}_{t,k,\lambda} - \widehat{\beta}_{ad,k,\lambda}| < \delta(\max_i \widehat{\beta}_{i,k,\lambda} - \min_i \widehat{\beta}_{i,k,\lambda}) \mid \widehat{\beta}_{t,k,\lambda} \neq 0, \widehat{\beta}_{t,k,\lambda} \neq \widehat{\beta}_{t-1,k,\lambda}\Big\},$$

$ad \ldots$ time when parameter k added 1 to c the last time ($t = 2$, $ad = 1$) and $\delta \in (0, 1)$.

Selection of the Tuning Parameters (4)

- **2. step - Augmented** IC_{QS}

$$IC_2(\lambda) = \log\left(\frac{1}{T} SSE(\widehat{\beta}_\lambda)\right) + \frac{1}{\sqrt[3]{T^2}} |\widehat{\beta}_\lambda|.$$

Underlying Assumptions (incompl.)

- m grows reasonably slow with T,
- the smallest break has to be reasonably large,
- $\exists \delta > 0$ such that $E[x_t^{4+\delta}]$, $E[\varepsilon_t^{4+\delta}]$ exist

Defining a Break

- bootstrap computationally too demanding, problems with inconsistency for the L_1-norm estimates (Camponovo, 2014)
- simple comparison of the estimated coefficients (adaptive part should be able to suppress the insignificant breaks rather well)
- i.e. when $\widehat{\beta}_t \neq \widehat{\beta}_{t+1}$, then $\widehat{T}_j = t + 1$ for $t \geq \widehat{T}_{j-1}$, where $j = 1, \dots, m$.

Simulation Study

- $T = \{50, 100, 200\}$,
- 1000 samples for each T,
- $\delta = \{\mathbf{0.025}, 0.05, 0.075\}$,
- Grid of λ:
  - covers 50 values between 0.01 and 0.5 (1. and 2. Step)

Simulation Study - Comparison

- Compare the performance of fused Lasso and tests introduced by Bai and Perron (1998):
- Double maximum (DM) test:

$$
\begin{aligned}
H_0 &: \text{no breaks} \\
H_A &: \text{unknown number of breaks} \\
&\quad \text{given some upper bound } M = 5
\end{aligned}
$$

- Testing sequentially:

$$H_0 : \ell \text{ breaks} \qquad \text{against} \qquad H_A : \ell + 1 \text{ breaks}$$

until H_0 not rejected.

Simulation Study - Motivation

- Potential advantages of fused Lasso in comparison to the structural break tests:
  - Not necessary to know the date of break in advance.
  - Not necessary to set number of breaks in advance.
  - Estimates simultaneously breaks and the coefficients.
  - No need to trim the sample, potential to detect break even at the beginning or the end of the time period (Bai and Perron, 1998).

Evaluation Criteria

- *FN* . . . False Negatives (Too Few Breaks),
- *TP* . . . True Positives (Correct Number of Breaks),
- *FP* . . . False Positives (Too Many Breaks),
- *Q* . . . Average Relative Distance from the True Break,
- *P* . . . Criterion Measuring if the Break is in the Correct Parameter.

Simulation 1 - 2 Parameters - 1 Break in 1 Parameter - Middle - SNR = 2

- DGP with no heteroscedasticity

$$
\begin{aligned}
y_t &= 2x_{t1} + x_{t2} + \varepsilon_t, &\qquad t &= 1, \dots, T/2, \\
y_t &= 4x_{t1} + x_{t2} + \varepsilon_t, &\qquad t &= T/2 + 1, \dots, T,
\end{aligned}
$$

$x_t \dots$ iid $N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.3 \\ 0.3 & 0 \end{bmatrix}\right)$, $\varepsilon_t \dots$ iid Gaussian, zero mean, $\sigma_\varepsilon^2 = 6.4$.

Simulation 1 - Bias

Table 1: Average Squared Bias, Break - Middle, 1000 draws, SNR=2, $\delta = 0.025$

	1 step			2 step				
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.590	0.644	0.452	0.595	0.630	0.592	0.548	0.565
100	0.450	0.448	0.292	0.419	0.408	0.267	0.346	0.311
200	0.377	0.376	0.206	0.340	0.332	0.138	0.206	0.191

Simulation 1 - EC

Table 2: Evaluation Criteria, Break - Middle, 1000 draws, SNR=2, $\delta = 0.025$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
	FN	51.5	46.6	7.3	12.9	10.8	39.1
	TP	26.8	27.3	29.6	45.2	41.1	43.3
50	FP	21.7	26.1	63.1	41.9	48.1	17.6
	Q	0.56	0.52	0.19	0.23	0.21	0.47
	P	0.41	0.45	0.67	0.70	0.69	0.30

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
100	FN	51.5	49.5	4.0	14.4	10.4	16.9
	TP	30.1	29.4	40.2	50.2	49.3	69.1
	FP	18.4	21.1	55.8	35.4	40.3	14.0
	Q	0.54	0.52	0.12	0.20	0.17	0.24
	P	0.45	0.46	0.79	0.79	0.81	0.42
200	FN	48.8	47.9	2.0	6.0	5.2	0.7
	TP	30.0	29.5	39.0	50.8	49.4	87.9
	FP	21.2	22.6	59.0	43.2	45.4	11.4
	Q	0.51	0.50	0.07	0.10	0.10	0.05
	P	0.48	0.49	0.88	0.90	0.90	0.50

Simulation 2 - 1 Break - End - SNR = 2

- DGP with no heteroscedasticity, no autocorrelation

$$y_t = 2x_t + \varepsilon_t, \qquad t = 1, \dots, T-11,$$
$$y_t = 4x_t + \varepsilon_t, \qquad t = T-10, \dots, T,$$

$x_t \dots$ iid $N(0,1)$, $\varepsilon_t \dots$ iid Gaussian, zero mean,

$$\sigma_\varepsilon^2 = 3.2 \text{ (T=50)},$$
$$\sigma_\varepsilon^2 = 2.6 \text{ (T=100)},$$
$$\sigma_\varepsilon^2 = 2.3 \text{ (T=200)}.$$

Simulation 2 - Bias

Table 3: Average Squared Bias, 1000 draws, Break - End, SNR=2, $\delta = 0.025$

	1 step			2 step				
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.590	0.615	0.378	0.602	0.596	0.414	0.482	0.460
100	0.370	0.369	0.201	0.371	0.367	0.216	0.273	0.257
200	0.203	0.203	0.126	0.204	0.204	0.133	0.169	0.165

Simulation 2 - EC

Table 4: Evaluation Criteria, 1000 draws, Break - End, SNR=2, $\delta = 0.025$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	FN	57.9	55.2	13.8	23.4	21.0	29.6
	TP	20.7	18.7	50.8	56.9	52.9	62.5
	FP	21.4	26.1	35.4	19.7	26.1	7.9
	Q	0.63	0.62	0.23	0.31	0.29	0.37

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
100	FN	89.6	88.8	21.6	40.9	36.8	58.3
	TP	4.1	3.7	41.2	42.7	42.7	38.7
	FP	6.3	7.5	37.2	16.4	20.5	3.0
	Q	0.91	0.90	0.30	0.45	0.42	0.64
200	FN	99.7	99.7	44.2	70.1	67.3	83.7
	TP	0.0	0.0	24.7	21.7	22.7	15.6
	FP	0.3	0.3	31.1	8.2	10.0	0.7
	Q	1.00	1.00	0.49	0.72	0.69	0.88

Simulation 3 - No Breaks - SNR = 2

- DGP with no heteroscedasticity, no autocorrelation

$$y_t = 2x_t + \varepsilon_t, \qquad t = 1, \ldots, T.$$

$x_t \ldots$ iid $N(0,1)$, $\varepsilon_t \ldots$ iid Gaussian, zero mean, $\sigma_\varepsilon^2 = 2$.

Simulation 3 - FN

Table 5: Rates of falsely detected breaks, 1000 draws, SNR=2, $\delta = 0.025$

	2 step					BP	
T	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	DM	Seq
50	0.7	0.9	22.4	11.0	12.8	49.3	24.4
100	0.0	0.0	8.4	2.2	2.7	31.3	13.6
200	0.0	0.0	2.8	0.4	0.4	16.9	9.0

Simulation 3 - Bias

Table 6: Average Squared Bias, 1000 draws, No Break, SNR=2, $\delta = 0.025$

	1 step			2 step				
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.044	0.044	0.067	0.043	0.046	0.088	0.058	0.065
100	0.022	0.022	0.032	0.022	0.022	0.031	0.023	0.024
200	0.010	0.010	0.014	0.010	0.010	0.012	0.011	0.011

Simulation - Information Criteria

1. step	2. Step	Outcome
$EBIC$	$EBIC$	High Ratio of FN
IC_{QS}	IC_{QS}	High Ratio of FN
IC_1	IC_2	Low Ratio of FN at a cost of more FP
IC_1	$EBIC$	Low Ratio of FN at a cost of more FP
IC_1	IC_{QS}	Low Ratio of FN at a cost of more FP

Application (1)

- Detection of changes in the US labor productivity in the period from 1955-2004,

Application (2)

- CES Production Function

$$Y_t = a(\varepsilon_t)\left[a_K(t)K_t^{\left(\frac{\sigma-1}{\sigma}\right)} + a_L(t)L_t^{\left(\frac{\sigma-1}{\sigma}\right)}\right]^{\left(\frac{\sigma}{\sigma-1}\right)}, \qquad (2)$$

where
Y_t = output in t,
K_t = capital input in t,
L_t = labor input in t,
$a_K(t)$ = capital augmenting technical progress,
$a_L(t)$ = labor augmenting technical progress,
$a(\varepsilon_t)$ = productivity shock,
σ = elasticity of substitution, $\sigma \geq 0$.

Application (3)

- Based on (2) and under following assumptions:

$$a(\varepsilon_t) = [\exp(\alpha_0 + \varepsilon_t)]^{\frac{1}{1-\sigma}},$$
$$a_L(t) = [\exp(\alpha_L \cdot t + \alpha_{L_2} \cdot t^2)]^{\frac{\sigma-1}{\sigma}},$$
$$\frac{Y_t/L_t}{Y_{t-1}/L_{t-1}} = \left(\frac{Y_t^*/L_t^*}{Y_{t-1}/L_{t-1}}\right)^{\gamma},$$

where $*$ denotes the optimal quantity and $\gamma \in (0, 1]$ denotes an adjustment parameter of the Labor Productivity, the FOC for L_t can be log-linearized to get:

Application (4)

- Logarithm of Labor Productivity:

$$\ln\left(\frac{Y_t}{L_t}\right) = \beta_0 + \beta_1 \cdot t + \beta_2 t^2 + \beta_3 \cdot \ln\left(\frac{w_t}{p_t}\right) + \beta_4 \ln\left(\frac{Y_{t-1}}{L_{t-1}}\right) + u_t,$$

where

$$w_t/p_t \dots \text{ real wage in } t,$$
$$\beta_0 = \gamma\alpha_0, \quad \beta_1 = (1-\sigma)\gamma\alpha_L, \quad \beta_2 = (1-\sigma)\gamma\alpha_{L_2},$$
$$\beta_3 = \sigma\gamma, \quad \beta_4 = 1-\gamma, \quad u_t = \gamma\varepsilon_t.$$

Application (5)

- Optimize (under homoscedasticity):

$$\frac{1}{T}\sum_{t=1}^{T}\left(\ln\left(\frac{Y_t}{L_t}\right)-\beta_{t,0}-\beta_{t,1}\cdot t-\beta_{t,2}t^2-\beta_{t,3}\cdot\ln\left(\frac{w_t}{p_t}\right)-\right.$$
$$\left.-\beta_{t,4}\ln\left(\frac{Y_{t-1}}{L_{t-1}}\right)\right)^2+\lambda\sum_{t=2}^{T}\sum_{k=0}^{4}w_{t,k}|\beta_{t,k}-\beta_{t-1,k}|, \tag{3}$$

Application (6)

- Quarterly Data for 1955Q1-2004Q1, $T = 199$,
- OECD Database (seasonally adjusted data):

Y_t	(Real) GDP
L_t	Civilian employment
w_t	(Nominal) wage index
p_t	Consumer Price Index

- For IC_1: $\delta = 0.025$.

Application (7) - $\widehat{\beta}_0$

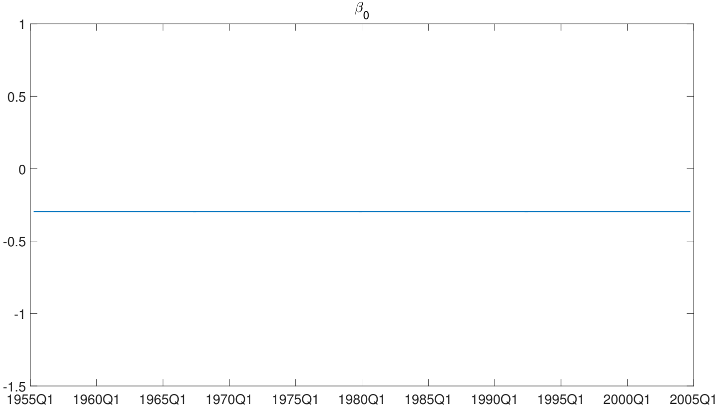


Figure 1: $\widehat{\beta}_0 = -0.0297$

Application (8) - $\widehat{\beta}_1$ - t

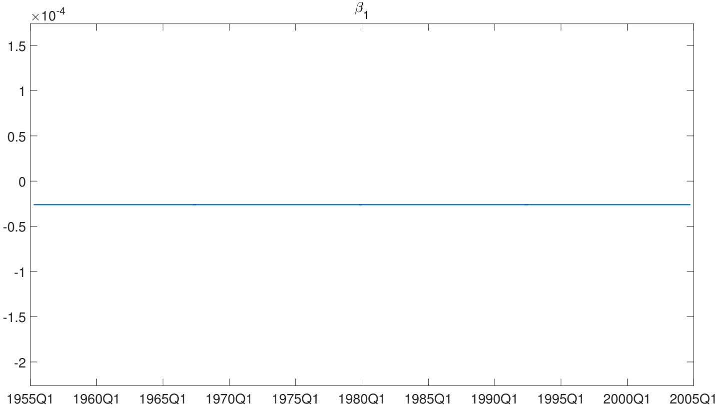


Figure 2: $\widehat{\beta}_1 = -0.000026$

Application (9) - $\widehat{\beta}_2$ - $t^2/100$

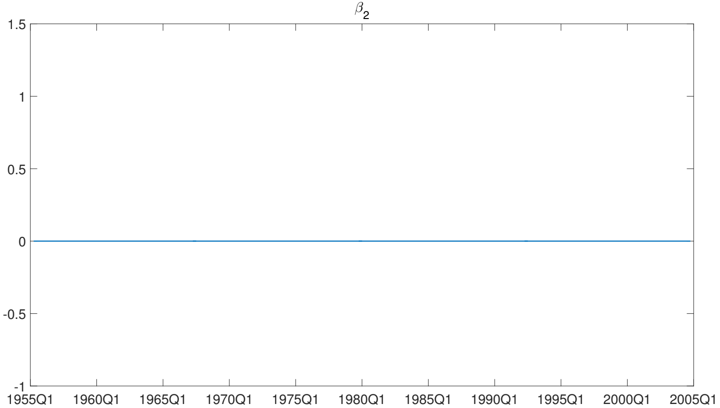


Figure 3: $\widehat{\beta}_2 = 0.000012$

Application (10) - $\widehat{\beta}_4$ - $\ln(Y_{t-1}/L_{t-1})$

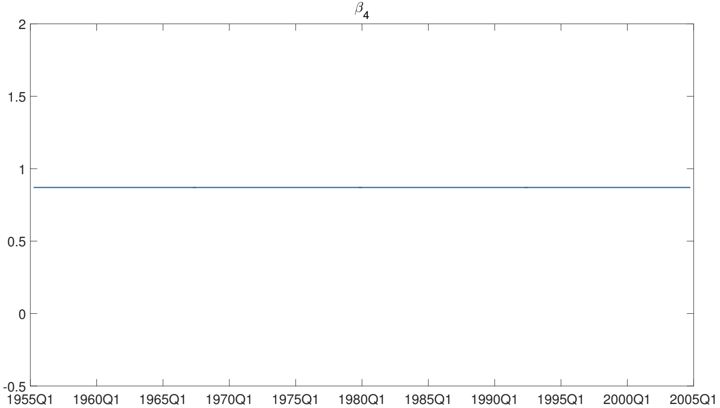


Figure 4: $\widehat{\beta}_4 = 0.8704 \Rightarrow \widehat{\gamma} = 0.1296$

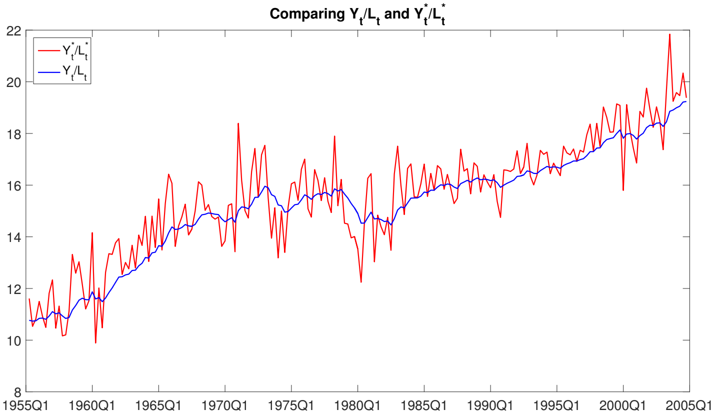
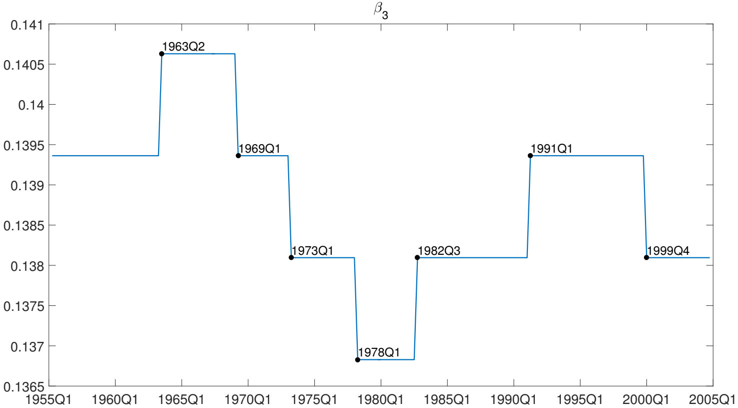


Figure 5: Comparison of optimal and realized values of Y_t/L_t, $\widehat{\gamma} = 0.1296 \Rightarrow$ 4.91% above or below optimal value on average

Application (11) - $\hat{\beta}_3$ - $\ln(w_t/p_t)$



- $\sigma = [1.0757 \quad 1.0855 \quad 1.0757 \quad 1.0660 \quad 1.0562 \quad 1.0660 \quad 1.0757 \quad 1.0660] \Rightarrow K$ and L (gross) substitutes.

Application (12) - Breaks Interpretation

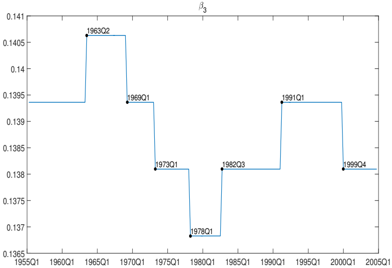


Figure 6: $\widehat{\beta}_3$

- mid 1960s - economic expansion
- 1973 Oil Crisis
- 1978 Oil Shock $\Rightarrow$ 1979 Oil Crisis
- 1983 - Rebound from the Early 1980s Recession
- 1991 - 2001 = 1990s economic boom
- Early 2000s recession

Conclusion

- Application of the fusion penalty to detect structural breaks in the coefficients of a standard linear regression model and estimate the coefficients.
- ICs controlling better the FN rate for a two step estimation procedure.
- Advantages of fusion penalty: allow for unknown number of breaks, no trimming of the data sample, regarding the estimation of the true position of the break outperform the BP tests.
- Future work: Statistical inference, Generalizing the error term structure

Any questions? Suggestions? Problems?

References I

Bai, J. and P. Perron (1998): "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica*, 66, pp. 47–78.

Camponovo, L. (2014): "On the Validity of the Pairs Bootstrap for Lasso Estimators," *Available at SSRN:*.

Chen, J. and Z. Chen (2008): "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, 95, 759–771.

Land, S. R. and J. H. Friedman (1997): "Variable fusion: A new adaptive signal regression method," Tech. rep., Technical Report 656, Department of Statistics, Carnegie Mellon University Pittsburgh.

References II

QIAN, J. AND L. SU (2014): "Shrinkage Estimation of Regression Models with Multiple Structural Changes," Working Papers 06-2014, Singapore Management University, School of Economics.

VIALLON, V., S. LAMBERT-LACROIX, H. HÖFLING, AND F. PICARD (2013): "Adaptive Generalized Fused-Lasso: Asymptotic Properties and Applications," .

YU, D., J.-H. WON, T. LEE, J. LIM, AND S. YOON (2013): "High-dimensional Fused Lasso Regression using Majorization-Minimization and Parallel Processing," *ArXiv e-prints*.