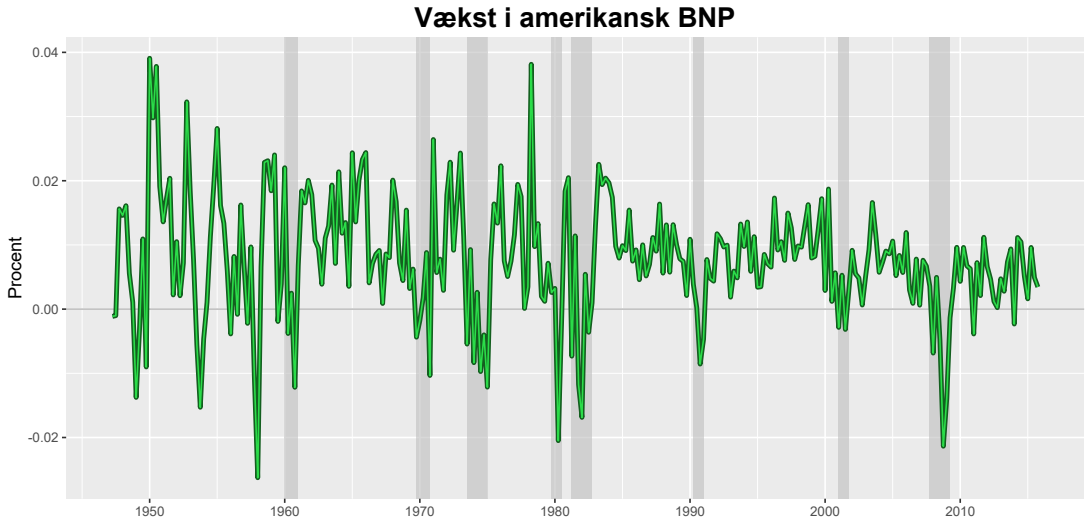


Aalborg Universitet
Bachelorprojekt

Makroøkonomiske prognoser og store datamængder

Forfattere

Nana Sofie Aarøe
Louise K. Nielsen
Rasmus D. Frederiksen
Janus S. Valberg-Madsen

Vejleder

Johannes T. Kristensen

27. maj 2016

Institut for matematiske fag
Fredrik Bajers Vej 7G
DK-9220 Aalborg Ø
http://math.aau.dk

AALBORG UNIVERSITET
STUDENTERRAPPORT

Titel:
Makroøkonomiske prognoser
og store datamængder

Tema:
Statistik modellering og analyse

Projektperiode:
Forårssemestret 2016

Projektgruppe:
MAOK6 G4-115

Deltager(e):
Nana Sofie Aarøe,
Louise Kragh Nielsen,
Rasmus Dovnborg Frederiksen,
Janus S. Valberg-Madsen

Vejleder(e):
Johannes Tang Kristensen

Oplagstal: 5

Sidetal: 60

Afleveringsdato:
2016-05-27

Summary

In this report we seek to forecast the quarterly and annual growth in US real GDP using freely available data from the database of economic variables, FRED. We will start with a simple autoregression as a benchmark model, and our main focus will then be to consider whether or not one can obtain better forecasts by conditioning on additional data series through two different models: the LASSO and the factor model.

The LASSO model is an extension of a classic, linear model, in which a shrinking constraint is imposed on the parameter estimates. This shrinkage makes parameters that would normally be close to zero become exact zero, and as such the LASSO can reduce the model to only include the most important explanatory variables.

The factor model limits the number of model parameters by collecting information about variance in the explanatory variables in some common factors.

The additional data that we want to condition on is a large collection of curated time series from FRED describing monthly variables. We find that relative to the benchmark model the LASSO can't improve on the forecasts of GDP conditioned on these time series, but we do see an improvement by using the factor model.

Since GDP is quarterly data and released with delay, we will — in addition to forecasting — perform so-called *nowcasting*, in which we condition on contemporary data in order to "predict the present". We find that nowcasts in general are more precise than forecasts.

The first half of the report will introduce some basic definitions pertaining time series analysis as well as some properties for the models we consider, and in the second half we will describe and analyse our results in detail. The most important parts of the R-code used to obtain these results are shown in an appendix in the back of the report, and all the code, along with datasets used, is available as an accompanying `zip`-file from the project database.

Forord

Dette projekt er udarbejdet i foråret 2016 af fire studerende på 6. semester af bachelor-uddannelsen i Matematik-Økonomi, Aalborg Universitet, under vejledning af Johannes Tang Kristensen, adjunkt.

Rapporten er skrevet i LaTeX. Bemærk, at vi bruger punktum som decimalseperator i stedet for komma. Til databehandling og modellering bruger vi programmet R[1] samt R-pakkerne:

- `dplyr`[2], `tidyr`[3] og `magrittr`[4], til behandling og transformering af data
- `ggplot2`[5] til figurer
- `glmnet`[6] til tilpasning af LASSO-modeller
- `R.utils`[7] og `forecast`[8] til diverse hjælpefunktioner

R-koden, der danner grundlag for resultaterne i denne rapport, er afviklet på en 64-bit Windows®-maskine, og er tilgængelig som download samme sted som rapporten.

Rapporten er opdelt i to dele, hvoraf første del omhandler den relevante teori, og anden del beskæftiger sig med den praktiske anvendelse. Bagerst i rapporten er et appendiks, der indeholder centrale dele af R-koden, samt en litteraturliste.

Aalborg University, 27. maj 2016

Nana Sofie Aarøe
<nsaj13@student.aau.dk>

Louise Kragh Nielsen
<lkni13@student.aau.dk>

Rasmus Dovnborg Frederiksen
<rfrede13@student.aau.dk>

Janus S. Valberg-Madsen
<jvalbe13@student.aau.dk>

[1] R Core Team [2015]
[2] Wickham & Francois [2015]
[3] Wickham [2016]
[4] Bache & Wickham [2014]
[5] Wickham [2009]
[6] Friedman *et al.* [2010]
[7] Bengtsson [2016]
[8] Hyndman [2015]

Indhold

1

Indledning

Denne rapport omhandler forecasting af makroøkonomiske data. Vi vil undersøge hvordan nogle forskellige modeller kan forecaste en makroøkonomisk tidsrække på baggrund af et større antal andre tidsrækker. Vi har tænkt os at forecaste væksten i USA's BNP på baggrund af 135 månedsbaserede økonomiske tidsrækker.

Vi benytter en autoregressiv model som benchmark, da forecasts fra denne kun betinger på tidligere data af samme tidsrække. For at udvide modellen med data fra alle de andre tidsrækker, ser vi på anvendelsen af LASSO- og faktor-modellen. Begge disse modeller vil udover autoregressive led inddrage de 135 forklarende tidsrækker. Vi formoder, at disse modeller vil give bedre forecasts, idet de anvender mere information.

Problemformulering: *I hvor høj grad kan forecasts af vækst i amerikansk BNP forbedres ved brugen af ekstra data, fra forskellige økonomiske sektorer?*

Yderligere vil vi se på hvordan LASSO- og faktor-modellerne kan bruges til såkaldte *nowcasts*, hvor man forecaster med samtidigt data; da opgørelsen af BNP er en kvartalsudgivet dataserie, vil månedsbaseret data være hurtigere tilgængelig. Dette giver mulighed for at betinge på data fra samme tidsperiode.

1.1 Problemafgrænsning

I dette projekt vil vi fokusere på den praktiske del af modellering, og vi vælger kun at se på den simpleste udgave af henholdsvis LASSO- og faktor-modellerne. Der findes mange udvidelser til begge, men disse ser vi bort fra, herunder modeller, der kan håndtere manglende datapunkter i tidsrækkerne og tidsrækker med forskellig frekvens. Vi vælger at håndtere problemet med forskellig frekvens i tidsrækkerne ved at aggregere månedsdata til kvartalsbasis.

Vi vælger ikke at indsamle og behandle vores egne data. I stedet benyttes datasættet fra [McCracken, 2016], som kommer med en guide til hvordan de enkelte tidsrækker transformeres til stationære tidsrækker. Vi antager, at data i disse tidsrækker ikke har været opgjort med tilbagevirkende kraft. Under databehandling vil alle tidsrækker blive aggregeret ens, idet vi ikke vil betragte og gøre os overvejelser om hver enkel tidsrækkes datatype. Yderligere antagelser omkring data er beskrevet i Kapitel 6.

Del I

Teori

2

Tidsrækker

Dette kapitel er baseret på [Shumway & Stoffer, 2011].
Mange størrelser, især økonomiske, bliver observeret regelmæssigt over tid, og kan derfor beskrives som såkaldte *tidsrækker*. Vi indfører herunder begrebet om tidsrækker og tilhørende egenskaber.

Definition 2.1: Tidsrække

Lad $\mathcal{T}$ betegne en indeksmængde, som beskriver tid.
For $d \geq 1$ definerer vi en d-dimensionel tidsrække til at være en følge

$$X \equiv \{X_t\}_{t \in \mathcal{T}},$$

hvor $X_t \in \mathbb{R}^d$ for alle $t \in \mathcal{T}$.

En tidsrække er således en sekvens af målinger eller observationer af samme fænomen, altså samme variabel, målt til ækvidistante tidspunkter. Tidsrækker bliver også kaldt for stokastiske processer. Vi beskæftiger os kun med en-dimensionelle tidsrækker indekseret i diskret tid, altså hvor $\mathcal{T} \subseteq \mathbb{Z}$ og $d = 1$. Teorien vil herfra af notationsmæssige årsager definere begreberne for en-dimensionelle, diskrete tidsrækker, men det kan alt sammen generaliseres til multidimensionelle, kontinuerte tidsrækker.

Den første egenskab omkring tidsrækker, vi vil betragte, er beskrevet ved den såkaldte *autokovariansfunktion*, som måler kovariansen mellem to værdier i samme tidsrække.

Definition 2.2: Autokovariansfunktion

Lad $\mu_t = \mathbb{E}[X_t]$. Autokovariansfunktionen for X er defineret som

$$\gamma(s,t) = \mathrm{Cov}[X_s, X_t] = \mathbb{E}[(X_s - \mu_s)(X_t - \mu_t)]$$

for alle $s, t \in \mathcal{T}$.

Bemærk, at $\gamma(s,t) = \gamma(t,s)$ for alle $s, t \in \mathcal{T}$. Autokovariansen forklarer den lineære afhængighed mellem to punkter fra den samme tidsrække observeret i forskellige tider. Hvis $\gamma(s,t) = 0$, så er X_s og X_t lineært uafhængige, men der kan stadig være en ikke-lineær afhængighedsstruktur mellem dem. Med visse kendte fordelinger af punkterne i X, f.eks. normalfordelingen, vil $\gamma(s,t) = 0$ betyde, at X_s og X_t er uafhængige. Når $s = t$ vil autokovariansfunktionen angive variansen for X_t:

$$\gamma(t,t) = \mathbb{E}\left[(X_t - \mu_t)^2\right] = \mathrm{Var}[X_t] = \sigma_t^2. \tag{2.1}$$

Eftersom størrelsen af en kovarians afhænger af observationernes standardafvigelser, vil den ikke i sig selv angive en størrelse, som er direkte sammenlignlig med andre kovarianser. Derfor vil vi i stedet bruge en normaliseret størrelse.

Definition 2.3: Autokorrelationsfunktion (ACF)

Autokorrelationsfunktionen for X er defineret som

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}} = \frac{\gamma(s,t)}{\sqrt{\sigma_s^2\sigma_t^2}},$$

hvor σ_s^2 og σ_t^2 er variansen for hhv. X_s og X_t.

Det kan let vises, at $-1 \leq \rho\,(s,t) \leq 1$. ACF bliver brugt til at måle, hvor god den lineære forudsigelse af en værdi til tiden t er; hvis vi kan forudsige X_t perfekt udfra X_s gennem den lineære relation $X_t = \beta_0 + \beta_1 X_s$, så er korrelationen 1 når $\beta_1 > 0$ og -1 når $\beta_1 < 0$. Dette er et groft mål for evnen til at forudsige en værdi til tidspunktet t udfra værdien i tiden s.

Ofte vil vi gerne måle, hvor god forudsigelsen af en anden tidsrække Y er udfra tidsrækken X. Vi antager, at begge tidsrækker har endelige varianser og definerer *krydskovariansfunktionen*, som måler kovariansen mellem to værdier til forskellige tidspunkter i to forskellige tidsrækker.

Definition 2.4: Krydskovariansfunktion

Krydskovariansfunktionen mellem to rækker, X og Y, er givet ved

$$\gamma_{XY}(s,t) = \mathrm{Cov}(X_s, Y_t) = \mathbb{E}\left[\left(X_s - \mu_{X_s}\right)\left(Y_t - \mu_{Y_t}\right)\right],$$

hvor μ_{X_s} og μ_{Y_t} angiver middelværdierne for hhv. X_s og Y_t.

Som ved ACF kan vi endvidere definere den normaliserede *krydskorrelationsfunktionen* (CCF), som måler graden af hvor meget to tidsrækker er korrelerede.

Definition 2.5: Krydskorrelationsfunktion (CCF)

Krydskorrelationsfunktionen mellem to tidsrækker X og Y er givet ved

$$\rho_{XY}\,(s,t) = \frac{\gamma_{XY}\,(s,t)}{\sqrt{\sigma_{X_s}^2\sigma_{Y_t}^2}},$$

hvor $\sigma_{X_s}^2$ og $\sigma_{Y_t}^2$ angiver variansen for hhv. X_s og Y_t.

2.1 Stationære tidsrækker

Alle de tidsrækker, som vi vil beskæftige os med, vil være (eller vil blive transformeret til) såkaldte *stationære* tidsrækker, og de vil have egenskaber, som forsimpler ovenstående notation. En strengt stationær tidsrække er kendetegnet ved, at fordelingen for en delmængde af rækken er invariant under tidsforskydning.

Definition 2.6: Streng stationaritet

En tidsrække X er *strengt stationær*, hvis den simultane fordeling af

$$(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$$

er den samme som den simultane fordeling af

$$(X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_k+h})$$

for alle $t_1 < t_2 < \cdots < t_k$, hvor $k \geq 1$, og for alle $h \in \mathbb{Z}$.
Udtrykt ved sandsynlighedsmålet, skal der gælde, at

$$\mathcal{P}(X_{t_1+h} \leq x_1, X_{t_2+h} \leq x_2, \ldots, X_{t_k+h} \leq x_k)$$

er ens for alle $h \in \mathbb{Z}$.

Begrebet om streng stationaritet er dog ikke særligt anvendeligt, så vi vil i stedet indføre et svagere, men mere praktisk kritirie.

Definition 2.7: Svag stationaritet

En tidsrække X er *svagt stationær*, hvis

- $\mu_t = \mu$, dvs. middelværdien er ens for alle X_t
- $\sigma_t^2 = \sigma^2$, dvs. variansen er ens for alle X_t
- $\gamma(s,t) = \gamma(s+h, t+h)$ for alle $s, t \in \mathcal{T}$, og $h \in \mathbb{Z}$

Vi vil kun beskæftige os med tidsrækker, som er svagt stationære, og ordet "stationær" vil således fremover implicit betyde svagt stationær. En stationær tidsrække kan betragtes som en følge, der hverken konvergerer eller divergerer.

Eftersom $\gamma(s,t)$ er invariant under tidsforskydning, afhænger den kun af differensen $|s-t|$, og vi kan således definere autokovarians og -korrelation for stationære tidsrækker på følgende måde:

Definition 2.8: Autokovarians og -korrelation for en stationær tidsrække

For en stationær tidsrække X defineres autokovariansen som

$$\gamma(h) = \mathrm{Cov}\left(X_t, X_{t+h}\right) = \mathbb{E}\left[\left(X_t - \mu\right)\left(X_{t+h} - \mu\right)\right]$$

og autokorrelationen som

$$\rho(h) = \frac{\gamma(t, t+h)}{\sqrt{\gamma(t,t)\gamma(t+h,t+h)}} = \frac{\gamma(h)}{\gamma(0)} = \frac{\gamma(h)}{\sigma^2}.$$

Begge disse funktioner er symmetriske, altså

$$\gamma(h) = \gamma(-h) \quad \text{og} \quad \rho(h) = \rho(-h) \tag{2.2}$$

for alle h. Dette ses let, idet

$$\begin{aligned}
\gamma\left(h\right) &= \gamma\left(\left(t+h\right) - t\right) \\
&= \mathbb{E}\left[\left(x_{t+h} - \mu\right)\left(x_t - \mu\right)\right] \\
&= \mathbb{E}\left[\left(x_t - \mu\right)\left(x_{t+h} - \mu\right)\right] \\
&= \gamma\left(t - \left(t+h\right)\right) \\
&= \gamma\left(-h\right), \qquad (2.3)
\end{aligned}$$

og når det holder for γ følger det direkte for ρ. Hvis vi er interesserede i at betragte hvor stor en del af korrelationen mellem X_{t+h} og X_t, som ikke allerede er forklaret af de mellemliggende værdier, skal vi i steder for ACF bruge den såkaldte *partielle autokorrelationsfunktion*.

Definition 2.9: Den partielle autokorrelationsfunktion (PACF)

Antag, at tidsrækken X har middelværdi 0. Lad $\widehat{X}_{t+h}$ for $h \geq 2$ betegne OLS-regressionen af X_{t+h} på $\{X_{t+h-1}, X_{t+h-2}, \ldots, X_{t+1}\}$, som kan skrives som

$$\widehat{X}_{t+h} = \beta_1 X_{t+h-1} + \beta_2 X_{t+h-2} + \cdots + \beta_{h-1} X_{t+1},$$

og lad $\widehat{X}_t$ betegne OLS-regressionen af X_t på $\{X_{t+1}, X_{t+2}, \ldots, X_{t+h-1}\}$, som skrives som

$$\widehat{X}_t = \beta_1 X_{t+1} + \beta_2 X_{t+2} + \cdots + \beta_{h-1} X_{t+h-1}.$$

Den partielle autokorrelationsfunktion for en stationær tidsrække X er nu givet ved

$$\rho_{11} = \mathrm{Cor}\left(X_{t+1}, X_t\right) = \rho(1)$$

og

$$\rho_{hh} = \mathrm{Cor}\left(X_{t+h} - \widehat{X}_{t+h}, X_t - \widehat{X}_t\right), \quad h \geq 2.$$

I definitionen ovenfor er der uden tab af generalitet antaget, at tidsrækken X har middelværdi 0. Hvis dette ikke er tilfældet, så betrages tidsrækken $X - \mu$ i stedet for.

3

Autoregressionsmodellen

En *autoregression* er en model, der beskriver datapunkter som en vægtet sum af tidligere datapunkter. For økonomiske variable vil det være anvendeligt, da man ofte må formode, at der vil være en sammenhæng mellem nuværende og foregående observationer. Vi vil derfor bruge en autoregressionsmodel som benchmark for vores andre modeller i den praktiske del af rapporten. Vi starter med at betragte en første ordens autoregressiv model, AR(1),

$$Y_t = \phi_1 Y_{t-1} + \omega_t, \tag{3.1}$$

hvor ω_t er hvid støj. Den rekursive form giver anledning til at folde udtrykket ud og få

$$\begin{aligned} Y_t &= \phi_1(\phi_1 Y_{t-2} + \omega_{t-1}) + \omega_t \\ &= \phi_1^2(\phi_1 Y_{t-3} + \omega_{t-2}) + \phi_1\omega_{t-1} + \omega_t \\ &\vdots \\ &= \phi_1^k Y_{t-k} + \phi_1^{k-1}\omega_{t-(k-1)} + \cdots + \phi_1\omega_{t-1} + \omega_t. \end{aligned} \tag{3.2}$$

Dette omskrives til

$$Y_t = \phi_1^k Y_{t-k} + \sum_{i=0}^{k-1} \phi_1^i \omega_{t-i} \tag{3.3}$$

Fra dette kan vi se, at hvis $|\phi_1| > 1$ vil processen vokse eksponentielt, når k bliver større, hvormed den ikke er stationær. Vi har derimod en stabil proces, hvis $|\phi_1| < 1$, da vægten af Y_{t-k} formindskes som k vokser. For tilfældet hvor $|\phi_1| = 1$ er processen blot en random walk.

Vi betragter nu en autoregression af generel orden, AR(p).

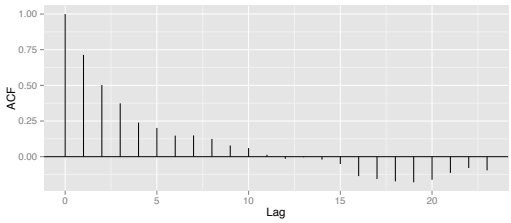
Definition 3.1: Autoregressiv model

En autoregressiv model af orden $p \geq 0$ er givet ved

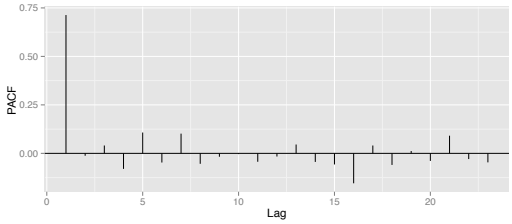
$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \omega_t, \quad t \in \mathbb{Z},$$

hvor parametrene $\phi_0, \phi_1, \ldots, \phi_p \in \mathbb{R}$ og ω_t er hvid støj.
Vi betegner en autoregressiv model af orden p som AR(p).

En måde at bestemme parameteren p, er ved hjælp af graferne for ACF og PACF. For en AR(p) vil ACF aftage eksponentielt, mens PACF vil vise høj grad af korrelation i p lags, men 0 herefter. På Figur 3.1 og 3.2 vises et eksempel på dette for en AR(1) proces.



Figur 3.1: ACF for en AR(1) proces



Figur 3.2: PACF for en AR(1) proces

3.1 Stationaritet af en AR(p)

For en simpel AR(1)-model var en nødvendig og tilstrækkelig betingelse for stationaritet, at $|\phi_1| < 1$, men for en generel AR(p)-model skal vi bruge det såkaldte *lag-polynomium*. For at definere dette, indfører vi *lag-operatoren* L, som afbilder værdier i en tidsrække til tidligere værdier, sådan at

$$LX_t = X_{t-1}. \tag{3.4}$$

Det ses let, at

$$X_{t-h} = LX_{t-h+1} = \underbrace{LL\dots L}_{h} X_t = L^h X_t. \tag{3.5}$$

Med denne notation kan en AR(p)-model for en tidsrække X omskrives til

$$\begin{aligned} \left(1 - \sum_{h=1}^{p} \phi_k L^h\right) X_t &= \phi_0 + \epsilon_t \\ \Phi(L) X_t &= \phi_0 + \epsilon_t, \end{aligned} \tag{3.6}$$

hvor

$$\Phi(L) \coloneqq 1 - \sum_{h=1}^{p} \phi_h L^h, \tag{3.7}$$

definerer lag-polynomiet. En autoregression er stationær hvis og kun hvis rødderne i dens lag-polynomium ligger strengt uden for enhedscirklen. Beviset for dette følger fra [Kasparis, 2008].

3.2 Moving-average model

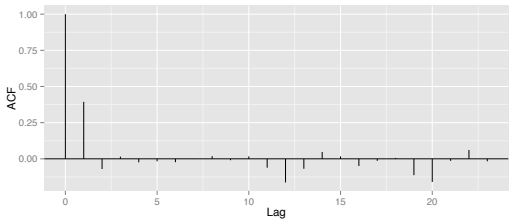
Definition 3.2: Moving-average model

En *moving-average model* af orden $q \geq 0$ er givet ved

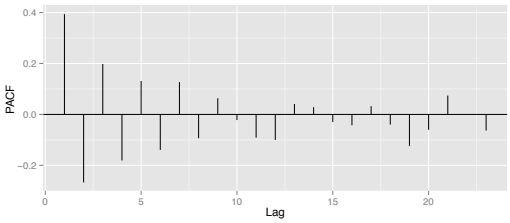
$$Y_t = \theta_0 + \omega_t + \theta_1 \omega_{t-1} + \theta_2 \omega_{t-2} + \cdots + \theta_q \omega_{t-q}, \quad t \in \mathbb{Z},$$

hvor parametrene $\theta_0, \theta_1, \dots, \theta_q \in \mathbb{R}$ og ω_t er hvid støj.
Vi betegner en moving-average model af orden q som MA(q).

Enhver stationær AR(p)-proces kan omskrives til en til MA(∞)-proces. Udledningen af resultatet fra (3.3) viser hvordan dette kan lade sig gøre for en AR(1). Da $|\phi_1| < 1$, vil vi have, at $\lim_{k\to\infty} \phi_1^k = 0$, hvilket gør det til en MA(∞)-proces.
En måde at bestemme q, er ved hjælp af ACF og PACF. Hvis ACF viser høj grad af korrelation i q lags, men 0 herefter, vil vi med stor sandsynlighed have med en MA(q) proces at gøre. PACF vil aftage eksponentielt for en MA(q) proces. På figur 3.3 og 3.4 vises et eksempel på dette for en MA(1) proces.



Figur 3.3: ACF for en MA(1) proces



Figur 3.4: PACF for en MA(1) proces

3.3 ARMA

ARMA modellen er en sammensætning af AR og MA processerne.

Definition 3.3

En ARMA(p, q) proces er givet ved

$$\begin{aligned}\text{ARMA}(p,q) = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \omega_t \\ + \theta_0 + \theta_1 \omega_{t-1} + \cdots + \theta_q \omega_{t-q}\end{aligned}$$

Der gælder, at ARMA($p, 0$) = AR(p) og ARMA($0, q$) = MA(q). Dog er det ikke muligt at bestemme p og q for en ARMA(p, q) ud fra ACF og PACF ligesom med AR(p) og MA(q), da både ACF og PACF vil aftage eksponentielt. Forholdene mellem modellerne og ACF/PACF er samlet i Tabel 3.1.

Model	ACF	PACF
AR(p)	Aftager eksponentielt	Nul for $h > p$
MA(q)	Nul for $h > p$	Aftager eksponentielt
ARMA(p, q)	Aftager eksponentielt	Aftager eksponentielt

Tabel 3.1: Forhold mellem modeller og ACF/PACF

4

LASSO-MODELLEN

Dette kapitel er baseret på [Hastie et al.*, 2015].*
Antag, at tidsrækken $\boldsymbol{y}$ følger en lineær struktur med N forklarende variable X_i givet ved

$$y_t = \alpha + \sum_{i=1}^{N} X_{t,i}\beta_i + e_t, \tag{4.1}$$

hvor $\alpha \in \mathbb{R}$ er niveau og $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_N) \in \mathbb{R}^N$. Den normale OLS-estimator til dette problem er

$$\left(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}\right) = \arg\min_{\alpha,\boldsymbol{\beta}} \left\{ \frac{1}{2T} \sum_{t=1}^{T} \left(y_t - \alpha - \sum_{i=1}^{N} X_{t,i}\beta_i \right)^2 \right\} \tag{4.2}$$

Når man arbejder med mange variable kan det være en fordel at udelade nogle af disse og kun fokusere på de relevante. Til at bestemme, hvilke der er relevante, bruger vi LASSO.

LASSO står for *Least Absolute Shrinkage and Selection Operator*. Denne metode tilføjer en bibetingelse til OLS-parameterestimationen, som sørger for at parametrene, der er tæt på 0 under normal OLS, vil blive eksakt 0. Vi får dermed et færre antal betydende variable at arbejde med, og sikrer at de er relevante. Dette reducerer modellen til en simplere model, uden at vi manuelt skal udvælge variable. LASSO-estimatoren er

$$\begin{aligned} \left(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}\right) &= \arg\min_{\alpha,\boldsymbol{\beta}} \left\{ \frac{1}{2T} \sum_{t=1}^{T} \left(y_t - \alpha - \sum_{i=1}^{N} X_{t,i}\beta_i \right)^2 \right\} \\ &\textit{u.b.b.} \sum_{i=1}^{n} |\beta_i| \leq K. \end{aligned} \tag{4.3}$$

K er en konstant, der har betydning for hvor meget problemet bliver begrænset. Et højt K vil give mere frihed i modellen, men med risiko for at overparameterisere. Et lavt K vil restringere modellen til færre prædiktorer, men med risiko for at modellen ikke opfanger tendensen i responsvariablen. På matrixform er dette

$$\begin{aligned} \left(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}\right) &= \arg\min_{\alpha,\boldsymbol{\beta}} \left\{ \frac{1}{2T} \|\boldsymbol{y} - \mathbf{1}\alpha - X\boldsymbol{\beta}\|_2^2 \right\} \\ &\textit{u.b.b.} \ \|\boldsymbol{\beta}\|_1 \leq K. \end{aligned} \tag{4.4}$$

Det er muligt at standardisere X således, at vi kan undgå at have α, ved at sætte middelværdien $\bar{X}_i = 0$ for alle $i = 1, \ldots, N$ og variansen $\frac{1}{T}\sum_{t=1}^{T} X_{t,i}^2 = 1$. Dette er mest anvendeligt, hvis man har forskellige enheder i variablene, da løsningen afhænger af

enhederne. Standardiseres X er tidsrækkerne således sammenlignelige. Findes $\widehat{\boldsymbol{\beta}}$ ud fra standardiserede tidsrækker, kan man bestemme $\widehat{\alpha}$ til de ikke-standardiserede tidsrækker ved

$$\widehat{\alpha} = \bar{y} - \sum_{i=1}^{N} \bar{X}_i \widehat{\beta}_i. \tag{4.5}$$

For at løse (4.4), skriver vi ligningen på Lagrange-form

$$\left(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}\right) = \arg\min_{\alpha,\boldsymbol{\beta}} \left\{ \frac{1}{2T} \|\boldsymbol{y} - \mathbf{1}\alpha - X\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}, \tag{4.6}$$

hvor $\lambda \geq 0$. For hver løsning af (4.4) med en given K, vil der være en entydig løsning af (4.6), da vi har Lagrange-dualitet.

5

Faktor-modellen

Dette kapitel er baseret på [Stock & Watson, 2002a].
I en faktormodel erstattes en gruppe af observerede variable med en uobserveret faktor, som er repræsentativ for variansen i den pågældende gruppe. Dette forsimpler modellen ved at reducere antallet af variable i modellen.

Estimering forløber over to skridt. Først estimeres tidsrækkerne af faktorerne ud fra de forklarende variable, og herefter kan relationen mellem responsvariablen og faktorerne bestemmes vha. OLS. I faktor-modellen er de forklarende variable repræsenteret ved

$$X_{t,i} = F_{t,1}\Lambda_{i,1} + \cdots + F_{t,r}\Lambda_{i,r} + e_{t,i} = F_t\Lambda_i^\top + e_{t,i} \tag{5.1}$$

hvor F_t er en rækkevektor af r faktorer, der erstatter de N variable. Λ_i er en rækkevektor af r faktorvægte og $e_{t,i}$ er et fejlled. Betragtes alle observationer til tiden t på én gang bliver dette

$$X_t = F_t\Lambda^\top + \boldsymbol{e}_t, \tag{5.2}$$

hvor Λ er en $N \times r$ matrix og F_t er en $1 \times r$ vektor af faktorerne. X_t er en $1 \times N$ vektor med observerede variable til tiden t. Her er $\boldsymbol{e}_t$ en $1 \times N$ vektor af fejlled.

5.1 Modelantagelser

(I) $\frac{1}{N}\Lambda^\top\Lambda \longrightarrow I_r$

(II) $\mathbb{E}\left[F_t^\top F_t\right] = \Sigma_F$, hvor Σ_F er diagonal kovariansmatrix med indgange $\sigma_{ii} > \sigma_{jj} > 0$ for $i < j$

(III) $|\Lambda_{i,m}| \leq M < \infty$ for $i = 1, \ldots, N$ og $m = 1, \ldots, r$, hvor M er en konstant

(IV) $\frac{1}{T}\sum_{t=1}^{T} F_t^\top F_t \xrightarrow{\mathcal{P}} \Sigma_F$

Man kan omskrive 5.2 som $F_t\Lambda^\top = F_tRR^{-1}\Lambda^\top$ for enhver invertibel matrix R. Dette betyder, at vi skal anvende en normalisering af modellen for at finde entydige faktorer. Antagelse (I) restringerer R til at være en ortonormal matrix, og antagelse (II) restringerer yderligere R til at være en diagonal matrix med indgange ± 1. Herfra vil vi kunne bestemme faktorerne op til fortegnsfejl. Antagelse (II) tillader også korrelation af faktorerne og lags af faktorerne. Dette kan give anledning til diverse udvidelser af modellen.

I økonomiske sammenhænge har tidsrækker sjældent i.i.d. normalfordelte fejlled, men vil i stedet udvise korrelation. Vores antagelser for fejlledene er derfor som følger:

(V) $\lim_{N\to\infty} \sup_t \sum_{h=-\infty}^{\infty} \left| \mathbb{E}\left[\frac{\boldsymbol{e}_t \boldsymbol{e}_{t+h}^\top}{N} \right] \right| < \infty.$

(VI) $\lim_{N\to\infty} \sup_t \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} |\gamma_{\boldsymbol{e}_t}(i,j)| < \infty$

(VII) $\lim_{N\to\infty} \sup_{t,s} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} |\mathrm{Cov}(e_{is}e_{it}, e_{js}e_{jt})| < \infty$

Når antagelse (V) er opfyldt, må fejlledstidsrækkerne godt udvise autokorrelation, antagelse (VI) tillader svag korrelation mellem fejlledene på tværs af variablene og antagelse (VII) begrænser størrelsen af de fjerde momenter.

5.2 Forecasting med faktor-modellen

Under faktor-modellen antager vi følgende struktur for responsvariablen:

$$y_{t+h} = F_t \boldsymbol{\beta}_F + \boldsymbol{w}_t \boldsymbol{\beta}_{\boldsymbol{w}} + \epsilon_{t+h}, \tag{5.3}$$

hvor $\boldsymbol{w}_t = (y_t, \ldots, y_{t-p+1})$, $\boldsymbol{\beta}_F$ og $\boldsymbol{\beta}_{\boldsymbol{w}}$ er OLS-koefficienterne, der hører til hhv. F og $\boldsymbol{w}$. Der er herunder antaget, at hvis F var observeret, ville OLS estimere modelkoefficienterne konsistent. Lad $\boldsymbol{z}_t = (F_t, \boldsymbol{w}_t)$ og $\boldsymbol{\beta} = (\boldsymbol{\beta}_F^\top, \boldsymbol{\beta}_{\boldsymbol{w}}^\top)^\top$, da gør følgende antagelser (5.3) mulig:

(VIII) $\mathbb{E}\left[\boldsymbol{z}_t^\top \boldsymbol{z}_t\right] = \Sigma_{\boldsymbol{z}} = \begin{bmatrix} \Sigma_F & \Sigma_{F\boldsymbol{w}} \\ \Sigma_{\boldsymbol{w}F} & \Sigma_{\boldsymbol{w}} \end{bmatrix}$ er en positiv definit matrix

(IX) $\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{z}_t^\top \boldsymbol{z}_t \xrightarrow{\mathcal{P}} \Sigma_{\boldsymbol{z}}$

(X) $\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{z}_t \epsilon_{t+h} \xrightarrow{\mathcal{P}} \boldsymbol{0}^\top$

(XI) $\frac{1}{T} \sum_{t=1}^{T} \epsilon_{t+h}^2 \xrightarrow{\mathcal{P}} \sigma^2$

(XII) $|\boldsymbol{\beta}| < \infty$

Antagelse (VIII) til (X) sørger for konsistens i OLS-estimeringen ved regression af y_{t+h} ud fra $\boldsymbol{z}_t$. De yderligere antagelser bliver senere brugt til at vise konsistens i OLS-estimeringen, når F erstattes med estimerede faktorer, $\widehat{F}$.

5.3 Estimering af faktorer

De sande faktorer F er ikke kendte. I stedet estimeres derfor faktorer $\widehat{F}$ ud fra følgende minimeringsproblem:

$$\widehat{F}, \widehat{\Lambda} = \arg\min_{F,\Lambda} \left\{ V(F,\Lambda) \right\}, \tag{5.4}$$

hvor

$$V(F,\Lambda) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{\top} \left(X_{t,i} - F_t \Lambda_i^\top \right)^2, \tag{5.5}$$

som på matrixform er

$$= \frac{1}{NT} \text{Trace}\Big[\Big(X - F\Lambda^\top\Big)^\top \Big(X - F\Lambda^\top\Big)\Big]. \tag{5.6}$$

For at løse (5.4) vil vi koncentrere F ud og minimere mht. Λ. Dette gør vi ved først at differentiere V med hensyn til F og sætte lig 0:

$$\frac{\partial V(F,\Lambda)}{\partial F} = \frac{2}{NT}\Big[F\Lambda^\top\Lambda - X\Lambda\Big] = 0. \tag{5.7}$$

Vi kan her se bort fra $\frac{2}{NT}$, da dette ingen indflydelse har på minimering af (5.6), og isolerer nu F:

$$\begin{aligned} F\Lambda^\top\Lambda &= X\Lambda \\ F &= X\Lambda(\Lambda^\top\Lambda)^{-1} \end{aligned}$$

Fra modelantagelserne har vi, at $\frac{1}{N}\Lambda^\top\Lambda \longrightarrow I_r$, og vi får:

$$F = \frac{1}{N}X\Lambda. \tag{5.8}$$

Vi har nu en lukket form for F, og denne indsætter vi tilbage i (5.6) og får

$$\begin{aligned} V(\Lambda) &= \frac{1}{NT}\text{Trace}\Big[\Big(X - X\Lambda\Lambda^\top/N\Big)^\top \Big(X - X\Lambda\Lambda^\top/N\Big)\Big] \\ &= \frac{1}{NT}\text{Trace}\Big[X^\top X + \Lambda\Lambda^\top X^\top X\Lambda\Lambda^\top/N^2 - X^\top X\Lambda\Lambda^\top/N - \Lambda\Lambda^\top X^\top X/N\Big] \\ &= \frac{1}{NT}\Big(\text{Trace}\Big[X^\top X\Big] + \text{Trace}\Big[\Lambda\Lambda^\top X^\top X\Lambda\Lambda^\top/N^2\Big] - \text{Trace}\Big[X^\top X\Lambda\Lambda^\top/N\Big] \\ &\qquad - \text{Trace}\Big[\Lambda\Lambda^\top X^\top X/N\Big]\Big). \end{aligned}$$

Da $\text{Trace}\Big[\Lambda\Lambda^\top X^\top X\Lambda\Lambda^\top/N^2\Big] = \text{Trace}\Big[X^\top X\Lambda\Lambda^\top/N\Big]$ får vi

$$V(\Lambda) = \frac{1}{NT}\Big(\text{Trace}\Big[X^\top X\Big] - \text{Trace}\Big[\Lambda\Lambda^\top X^\top X/N\Big]\Big). \tag{5.9}$$

At minimere denne funktion er ækvivalent med at maksimere

$$\text{Trace}\Big[\Lambda^\top X^\top X\Lambda\Big], \tag{5.10}$$

da $\text{Trace}\Big[X^\top X\Big]$ og konstantleddene er uafhængige af Λ og dermed kan ignoreres. Dette maksimeringsproblem løses ved at sætte $\widehat{\Lambda}$ lig med egenvektorene fra $X^\top X$, der svarer til de r største egenværdier. Med denne $\widehat{\Lambda}$ kan vi udregne $\widehat{F}$ ved hjælp af (5.8) og får

$$\widehat{F} = \frac{1}{N}X\widehat{\Lambda}. \tag{5.11}$$

Hvis $N > T$ kan det udregnes lettere ved at koncentrere Λ ud i stedet for F. Vi får da at minimere (5.5) er lig med at maksimere $\text{Trace}\Big[F^\top\Big(XX^\top\Big)F\Big]$ under betingelsen at $F^\top F/T = I_r$. Dette giver estimatoren $\tilde{F}$ som er en matrix af egenvektorer tilsvarende de r største egenværdier fra $XX^\top$ [Stock & Watson, 2002a, s. 1169]. Da $\widehat{F}$ og $\tilde{F}$ har ækvivalente søjlerum, kan disse bruges i flæng når der forecastes.

5.3.1 Konsistens af estimatorer

Vi gennemgår her et par sætninger, som sikrer, at vi har konsistens i vores estimater i faktor-modellen. Med konsistens mener vi, at estimerede værdier asymptotisk vil forløbe imod de sande værdier.

Sætning 5.1

Lad S_i være en variabel af værdi ± 1 og lad $N, T \longrightarrow \infty$. Vi antager, at k faktorer bliver estimeret. Da kan S_i vælges således, at følgende er opfyldt

a) $S_i \widehat{F}_{t,i} \xrightarrow{\mathcal{P}} F_{t,i}, \quad \text{for } i = 1, \ldots, r$

b) $\frac{1}{T} \sum_{t=1}^{\top} (S_i \widehat{F}_{t,i} - F_{t,i})^2 \xrightarrow{\mathcal{P}} 0, \quad \text{for } i = 1, \ldots, r$

c) $\frac{1}{T} \sum_{t=1}^{\top} \widehat{F}_{t,i}^2 \xrightarrow{\mathcal{P}} 0, \quad \text{for } i = r+1, \ldots, k, \quad \text{hvis } k > r$

Bevis findes i [Stock & Watson, 2002a, s. 1174-1177]. Da vi kun kan estimere $F_{t,i}$ op til en fortegnsfejl, bruger vi S_i til at korrigere for dette, hvilket bruges i punkt a). Punkt b) siger, at de korrigerede $\widehat{F}_{t,i}$ går imod den sande værdi $F_{t,i}$ når N, T vokser, som ses ved middelkvadratafvigelsen. Punkt c) siger, at de sidste $k - r$ estimerede faktorer har varians 0 og dermed ingen betydning.

Sætning 5.2

Lad $\widehat{\boldsymbol{\beta}}_F$ og $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}}$ betegne OLS-estimaterne af $\boldsymbol{\beta}_F$ og $\boldsymbol{\beta}_{\boldsymbol{w}}$ fra regression af $\{y_{t+h}\}_{t=1}^{T-h}$ på $\{\widehat{F}_t, \boldsymbol{w}_t\}_{t=1}^{T-h}$. Vi har da, at følgende holder

a) $\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}} - \boldsymbol{\beta}_{\boldsymbol{w}} \xrightarrow{\mathcal{P}} 0$

b) S_i kan vælges således, at $S_i \widehat{\boldsymbol{\beta}}_{F,i} - \boldsymbol{\beta}_{F,i} \xrightarrow{\mathcal{P}} 0, \quad \text{for } i = 1, \ldots, r$

c) $\left(\widehat{F}_T \widehat{\boldsymbol{\beta}}_F + \boldsymbol{w}_T \widehat{\boldsymbol{\beta}}_{\boldsymbol{w}}\right) - \left(F_T \boldsymbol{\beta}_F + \boldsymbol{w}_T \boldsymbol{\beta}_{\boldsymbol{w}}\right) \xrightarrow{\mathcal{P}} 0$

Bevis. Fra sætning 5.1 har vi, at $\widehat{F}$ er konsistent, og dermed gælder pga. tidligere antagelser:

- $\frac{1}{T} \sum_{t=1}^{T} \widehat{F}_t^{\top} \widehat{F}_t \xrightarrow{\mathcal{P}} \Sigma_{FF}$
- $\frac{1}{T} \sum_{t=1}^{T} S \widehat{F}_t^{\top} \boldsymbol{w}_t \xrightarrow{\mathcal{P}} \Sigma_{F\boldsymbol{w}}$
- $\frac{1}{T} \sum_{t=1}^{T} S \widehat{F}_t^{\top} \epsilon_{t+h} \xrightarrow{\mathcal{P}} \mathbf{0}$,

hvor $S = \text{diag}\,(S_1, \ldots, S_r)$.

Bevis for a) og b). Vi har, at

$$\begin{bmatrix} S\widehat{\boldsymbol{\beta}}_F \\ \widehat{\boldsymbol{\beta}}_{\boldsymbol{w}} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\beta}_F \\ \boldsymbol{\beta}_{\boldsymbol{w}} \end{bmatrix} = \begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T} \widehat{F}_t^\top \widehat{F}_t & S\frac{1}{T}\sum_{t=1}^{T} \widehat{F}_t^\top \boldsymbol{w}_t \\ \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{w}_t^\top \widehat{F}_t S & \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{w}_t^\top \boldsymbol{w}_t \end{bmatrix}^{-1} \begin{bmatrix} S\frac{1}{T}\sum_{t=1}^{T} \widehat{F}_t \epsilon_{t+h} \\ \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{w}_t \epsilon_{t+h} \end{bmatrix} \tag{5.12}$$

$$\xrightarrow{\mathcal{P}} \begin{bmatrix} \Sigma_F & \Sigma_{F\boldsymbol{w}} \\ \Sigma_{F\boldsymbol{w}} & \Sigma_{\boldsymbol{w}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} = \mathbf{0}. \tag{5.13}$$

Bevis for c). Lad $\widehat{\boldsymbol{z}}_t = \left(\widehat{F}_t, \boldsymbol{w}_t\right)$ og $\widehat{\boldsymbol{\beta}} = \left[\sum_{t=1}^{T-h} \widehat{\boldsymbol{z}}_t^\top \widehat{\boldsymbol{z}}_t\right]^{-1} \left[\sum_{t=1}^{T-h} \widehat{\boldsymbol{z}}_t^\top y_{t+h}\right]$, da vil

$$\widehat{\boldsymbol{z}}_T\widehat{\boldsymbol{\beta}} - \boldsymbol{z}_T\boldsymbol{\beta} \xrightarrow{\mathcal{P}} 0. \tag{5.14}$$

Dette ses ved at lade $R = \begin{bmatrix} S & 0 \\ 0 & I_p \end{bmatrix}$, hvor p er antal elementer i $\boldsymbol{w}$, og betragte

$$\begin{aligned} \widehat{\boldsymbol{z}}_T\widehat{\boldsymbol{\beta}} - \boldsymbol{z}_T\boldsymbol{\beta} &= \widehat{\boldsymbol{z}}_T RR\widehat{\boldsymbol{\beta}} - \boldsymbol{z}_T\boldsymbol{\beta} \\ &= \widehat{\boldsymbol{z}}_T RR\widehat{\boldsymbol{\beta}} + \boldsymbol{z}_T\left(R\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \boldsymbol{z}_T R\widehat{\boldsymbol{\beta}} \\ &= \underbrace{\left(\widehat{\boldsymbol{z}}_T R - \boldsymbol{z}_T\right)}_{\xrightarrow{\mathcal{P}} 0} R\widehat{\boldsymbol{\beta}} + \boldsymbol{z}_T \underbrace{\left(R\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)}_{\xrightarrow{\mathcal{P}} 0} \\ &\xrightarrow{\mathcal{P}} 0. \end{aligned}$$

□

5.3.2 Estimering af r

I praksis er det muligt at estimere r ved hjælp af AIC. Man udvælger den r, der giver modellen med mindst AIC. Dette er dog bevist af [Bai & Ng, 2002] til ikke at være et konsistent estimat. I stedet bruges funktionen fra (5.5), men denne gang som funktion af F og k, hvor $0 < k \leq k_{max}$ er antallet af faktorer, der estimeres:

$$V(k, \widehat{F}) = \min_{\Lambda} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{t,i} - \widehat{F}_t \Lambda_i^\top)^2, \tag{5.15}$$

hvor Λ_i og $\widehat{F}_t$ nu er af længde k i stedet for r. Dette skal sammen med en straffunktion $g(N,T)$ danne et kriterie på formen

$$IC(k) = \ln\left(V(k, \widehat{F})\right) + kg(N,T). \tag{5.16}$$

[Bai & Ng, 2002] identificerer tre kriterier, der asymptotisk vil estimere det sande r:

$$IC_1(k) = \ln\left(V(k, \widehat{F})\right) + k\left(\frac{N+T}{NT}\right)\ln\left(\frac{NT}{N+T}\right) \tag{5.17}$$

$$IC_2(k) = \ln\left(V(k, \widehat{F})\right) + k\left(\frac{N+T}{NT}\right)\ln\left(\min\{N,T\}\right) \tag{5.18}$$

$$IC_3(k) = \ln\left(V(k, \widehat{F})\right) + k\left(\frac{\ln\left(\min\{N,T\}\right)}{\min\{N,T\}}\right). \tag{5.19}$$

Da vil det k, der minimerer $IC(k)$, være det bedste estimat af r:

$$\widehat{k} = \underset{0<k\leq k_{max}}{\arg\min}\ IC(k). \tag{5.20}$$

Hvis man i praksis oplever, at $\widehat{k} = k_{max}$, kan det være en indikation for, at $k_{max} < r$. Dette medfører, at man ikke har estimater for alle de sande faktorer, men de $\widehat{k}$ faktorer, man har, vil stadig opfylde punkt a) og b) i Sætning 5.1.

Del II

Resultater

6

Databehandling

Vi beskæftiger os med to datasæt, begge hentet fra Federal Reserve Bank of St. Louis' database over økonomiske data (FRED):

- `GDPC96`: BNP for USA på kvartalsbasis, justeret for inflation og sæson, mellem 1947-Q1 og 2015-Q4 [US. Bureau of Economic Analysis, 2016]
- `FRED-MD`, 2016-03: en samling af 135 makroøkonomiske og finansielle tidsrækker på månedsbasis, mellem 1959-01 og 2016-02 [McCracken, 2016]. Databasen er beskrevet i detaljer i [McCracken & Ng, 2015]

Formålet vil være at forecaste ændringen i `GDPC96` betinget på tidsrækkerne i `FRED-MD`. Opgørelse af BNP er en længere proces, og derfor bliver data for BNP først udgivet senere end den periode, de beskriver. Det er således heller ikke uinteressant at foretage nowcasting, dvs. prædiktion af nutiden.

Det er ikke selve `GDPC96`, som den ser ud i FRED, vi vil betragte, men i stedet den procentuelle ændring over h kvartaler, hvilket er givet ved

$$y_{t+h}^h = 100 \log \left(\frac{\texttt{GDPC96}_{t+h}}{\texttt{GDPC96}_t} \right). \tag{6.1}$$

Her bruger vi notationen fra [Stock & Watson, 2002b, s. 149], som bruger et toptegn til at angive antallet af perioder, som ændringen i variablen betragtes over. Vi vil desuden også anvende dette i parameterestimaterne i vores modeller for at skelne imellem forskellige fits. Når vi betragter hele tidsrækken samlet, betegner vi den $\boldsymbol{y}$.

For at kunne anvende de månedsbaserede data til at prædiktere y_{t+h}^h, er det nødvendigt at aggregere dem til kvartalsbasis. Dette gør vi på to forskellige måder, afhængigt af, om vi forecaster eller nowcaster:

Til nowcasting aggregerer vi ved at betragte observationen for den første måned i kvartalet som repræsentation for hele kvartalet. Dette gør vi for at mindske risikoen for at betinge på data, som i virkeligheden ikke var tilgængelige i den periode, de nowcaster. Langt de fleste tidsrækker i `FRED-MD` er up-to-date, men enkelte halter nogle gange nogle perioder bagefter. Vi antager herunder, at tidsrækkernes opgørelse har været konsistente i løbet af årene. Hvis en tidsrække derfor mangler for meget nylig data, antager vi, at den altid har gjort det, og dermed frasorteres den.

Til forecasting har vi ikke med samtidige data at gøre, og derfor aggregerer vi ved at tage et uvægtet gennemsnit over tre måneder ad gangen og anvende dette gennemsnit som observationen for det pågældende kvartal. Dette har vi valgt at gøre ud fra en betragtning om, at et gennemsnit over hele kvartalet giver et bedre billede af udviklingen end blot den første måned. Man kunne også vælge andre metoder, f.eks. at tage sidste

måned i kvartalet, eller summen over hele kvartalet. Vi er opmærksomme på, at aggregeringsmetoden kan have indflydelse på resultatet, men i henhold til Afsnit 1.1 har vi ikke betragtet tidsrækkerne individuelt for at bestemme "korrekte" aggregeringsmetoder til hver tidsrække. I Appendiks A.1 er vores aggregeringsfunktion vist.

Efter disse aggregeringer, transformerer vi tidsrækkerne hver især, således at de bliver stationære. Disse transformationer er i datasættet angivet ved et "transform flag" til hver tidsrække, som er et tal, der indikerer, hvilken transformation, der er anbefalet i [McCracken & Ng, 2015]. De forskellige transformationer er:

(1) Ingen transformation

(2) Første ordens differens, ΔX_t

(3) Anden ordens differens, $\Delta^2 X_t$

(4) Den naturlige logaritme, $\log(X_t)$

(5) Første ordens log-differens, $\Delta \log(X_t)$

(6) Anden ordens log-differens, $\Delta^2 \log(X_t)$

(7) Procentuel ændring, $\frac{X_t}{X_{t-1}} - 1.0$

I det datasæt, vi står tilbage med, vil mange af tidsrækkerne have manglende værdier i starten som resultat af deres transformation. Desuden indeholder flere af tidsrækkerne i forvejen manglende værdier som følge af, at den variabel, de beskriver, først er indført senere end de andre. Vi vælger derfor at beskære alle tidsrækkerne (inkl. responsvariablen), således at de starter i 1960-Q1. De få tidsrækker, der stadig har manglende værdier efter dette, bliver frasorteret.

For at vores prædikterende tidsrækker har samme længde som responsvariablen, beskærer vi også datasættet, så det slutter i 2015-Q4. Dermed har alle tidsrækkerne længde 224, hvilket svarer til 56 års data. De prædikterende tidsrækker samles som søjlerne i en matrix, som vi herfra betegner X, og forecasting indebærer således at estimere

$$\widehat{y}^h_{t+h|t} \coloneqq \mathbb{E}\left[y^h_{t+h} \middle| \mathcal{F}_t \right], \tag{6.2}$$

hvor $\mathcal{F}_t = \{y_t, X_t, y_{t-1}, X_{t-1}, \ldots, y_1, X_1\}$. X_t angiver her rækkevektorer i X, som repræsenterer observationerne fra alle prædiktorer til tid t.

7

Benchmark-model

For at kunne udtale os om, hvor godt en model forecaster, vil vi betragte *mean square forecast error* (MSFE) for hver af de modeller, vi bruger. Fremgangsmåden for at beregne MSFE er, at vi iterativt forecaster $\widehat{y}^h_{t+h|t}$ og betragter forskellen mellem disse og de tilsvarende observerede værdier. Denne iterative proces starter til en tid t_0 og løber op til tid $T-h$, som er den sidste tid, hvortil vi kan observere y^h_{t+h}, og

$$MSFE^h = \frac{1}{T-(t_0+h-1)} \sum_{t=t_0}^{T-h} \left(y^h_{t+h} - \widehat{y}^h_{t+h|t}\right)^2. \tag{7.1}$$

For at have tilstrækkeligt mange observationer at tilpasse modellerne over, starter vi vores iterative forecasts midt i 80'erne, hvilket svarer til $t_0 = 100$. Vi mener desuden, at dette er et passende sted at starte, idet denne tid angiver begyndelsen af "the great moderation", hvor man så et fald i volatiliteten i økonomien. I Appendiks A.2 er koden, der foretager forecasting-skridtet, vist.

Vi starter med at betragte en simpel autoregressiv model, AR(p), idet vi har med tidsrækkedata at gøre, og den vil komme til at udgøre et sammenligningsgrundlag for de udvidede modeller. Modellen vil således blive tilpasset udelukkende på tidsrækken med væksten i `GDPC96`, og først i de andre modeller vil vi betinge på andre tidsrækker.

AR(p)-modellen estimeres med mindste kvadraters metode (OLS), og ordenen $p \leq 4$ bestemmes ud fra AIC. Den maksimale orden på 4 betyder, at modellen kun betinger med højst ét års tidligere data. Denne begrænsning har vi valgt at pålægge modellen, fordi vi ikke mener, at en model af højere orden er realistisk for data af denne type. Forecastligningen for benchmark-modellen er

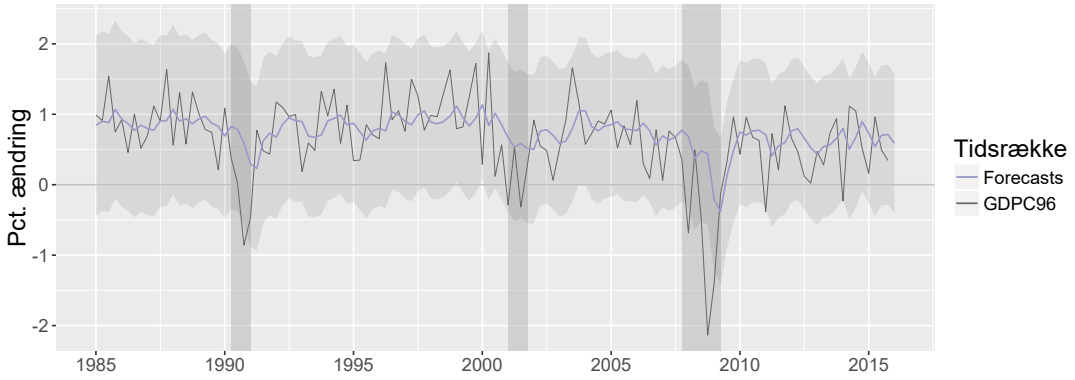
$$\widehat{y}^h_{t+h|t} = \widehat{\alpha}^h + \sum_{j=1}^{p} \widehat{\beta}^h_j y_{t-j+1}. \tag{7.2}$$

Vi vil med alle modellerne forecaste $\widehat{y}^1_{T+1|T}$ og $\widehat{y}^4_{T+4|T}$, som hhv. beskriver væksten i 2016-Q1 og væksten i hele 2016 samlet. Resultaterne for benchmark-modellen er vist i Tabel 7.1, og disse vil blive sammenlignet og diskuteret i Kapitel 10.

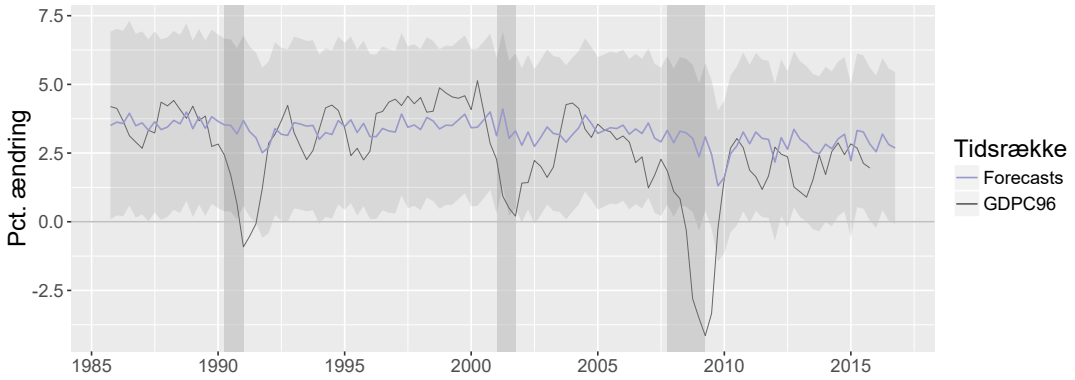
	$\widehat{y}^h_{t+h\|t}$	$MSFE^h_{\text{AR}}$
$h=1$	0.587 %	2.933 E-05
$h=4$	2.685 %	2.946 E-04

Tabel 7.1: Resultater for forecasting med benchmark-modellen

På Figur 7.1 og 7.2 er vist de iterative forecasts for hhv. kvartalsmæssig og årlig vækst i `GDPC96`. På figurerne er yderligere anført et approksimeret forecast-interval, som er dannet på baggrund af en antagelse om normalitet i residualerne.



Figur 7.1: Benchmark-modellens forecasts for den kvartalsmæssige vækst i `GDPC96` vist sammen med observerede værdier og et approksimeret 80 % forecast-interval



Figur 7.2: Benchmark-modellens forecasts for den årlige vækst i `GDPC96` vist sammen med observerede værdier og et approksimeret 80 % forecast-interval

8

Modellering med LASSO-modellen

Vi anvender funktionen `glmnet` fra R-pakken af samme navn [Friedman *et al.*, 2010] til at estimere LASSO-koefficienterne i vores model. Funktionen genererer ud fra datasættet en sekvens af 100 λ-værdier og tilpasser en model til hver af disse ved maksimum likelihood estimation med minimeringsalgoritmen *coordinate descent*. Ud fra BIC vælger vi dernæst den værdi $\widehat{\lambda}$, som giver den bedste model. De vigtigste dele af den hertil anvendte R-kode er vist i Appendiks A.3.

Fremgangsmåden for at forecaste og beregne MSFE for LASSO-modellen med en forecast-horisont h bliver således, at for $t = 100, 101, \ldots, T - h$:

1. Tilpas LASSO-modeller med `glmnet` over $\left(y_{1+h}^h, \ldots, y_t^h\right)$ betinget på $(Z_1, \ldots, Z_{t-h})$, hvor Z er en $T \times (N + p)$ matrix, der består af X og $\boldsymbol{y}, L\boldsymbol{y}, L^2\boldsymbol{y} \ldots, L^{p-1}\boldsymbol{y}$
2. Vælg bedste model ud fra $BIC = t\log(\widehat{\sigma}^2) + \log(t)k$, hvor k er antallet af valgte prædiktorer
3. Forecast $\widehat{y}_{t+h|t}^h$ ud fra forecast-ligningen
$$\widehat{y}_{t+h|t}^h = \widehat{\alpha}^h + Z_t\widehat{\boldsymbol{\beta}}^h \tag{8.1}$$
4. Gem forecast-fejl $y_{t+h}^h - \widehat{y}_{t+h|t}^h$

MSFE beregnes til slut som angivet i (7.1).

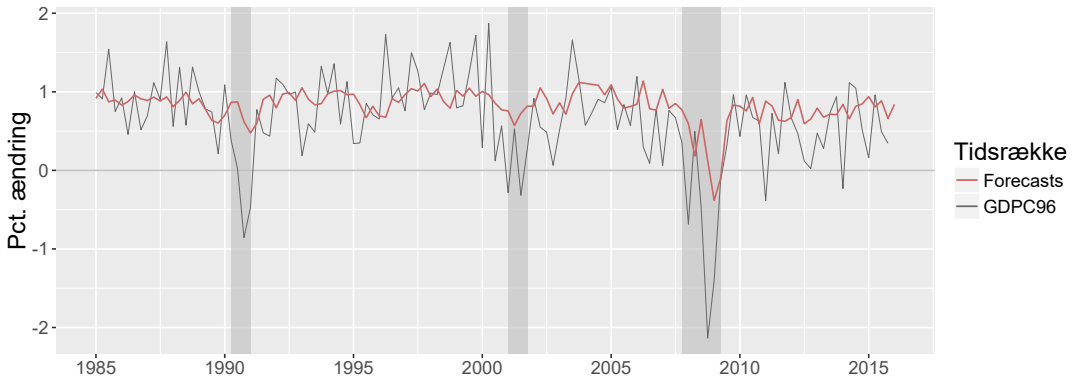
8.1 Forecasting med LASSO-modellen

Modellens forecasts af hhv. den kvartalsmæssige vækst i 2016-Q1 og den samlede vækst i hele 2016 er angivet i Tabel 8.1 sammen med de tilhørende MSFE. T betegner her 2015-Q4.

	$\widehat{y}_{T+h\|T}^h$	$MSFE_{\text{LASSO}}^h$
$h = 1$	0.840 %	2.945 E-05
$h = 4$	3.401 %	3.560 E-04

Tabel 8.1: Resultater for forecasting med LASSO-modellen

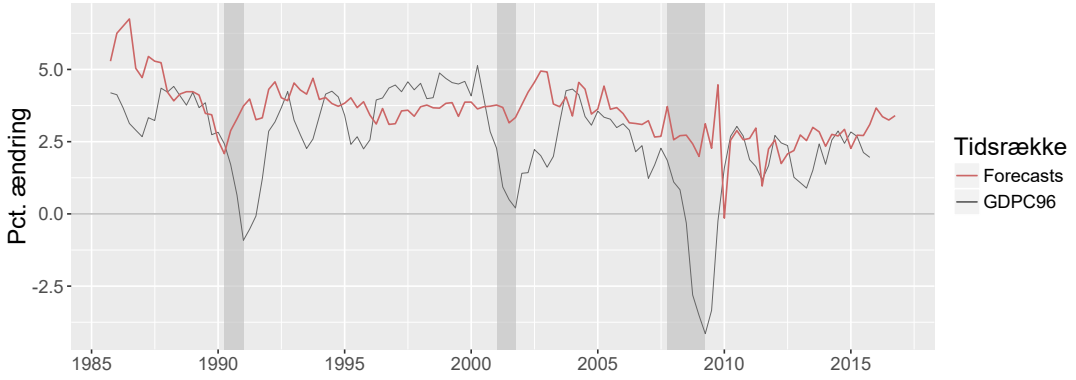
På Figur 8.1 er de iterative forecasts vist sammen med de tilsvarende sande værdier for den kvartalsmæssige ændring i `GDPC96`. Vores forecasts er mere konservative mht.



Figur 8.1: LASSO-modellens forecasts for den kvartalsmæssige vækst i `GDPC96` vist sammen med sande værdier

udsving end de realiserede værdier, hvilket gør, at pludselige konjunkturændringer ikke fanges i sit fulde omfang, og dette er især tydeligt ved finanskrisen i 2008.

En interessant bemærkning omkring vores iterative forecasts er, at der ud af de 125 forecasts ikke var ét eneste, hvor modellen medtog laggede værdier af $\boldsymbol{y}$ iblandt de signifikante prædiktorer. Dette kan måske skyldes, at den dynamik, som de laggede værdier kan forklare, bedre indfanges af nogle af de andre foreklarende variable. Hvis disse har højere korrelation med responsvariablen, bliver de laggede værdier således overflødige for modellen.



Figur 8.2: LASSO-modellens forecasts for den årlige ændring i `GDPC96` vist sammen med sande værdier

På Figur 8.2 er de iterative forecasts for årlige ændringer vist, hvor vi kan se, at de i knap så høj grad som ovenfor fanger tendenser og shocks. Dette er også at forvente, idet forecast-horisonten netop er et helt år. På figuren er også vist fire out-of-sample forecasts, som viser en positiv udvikling henover 2016.

8.2 Nowcasting med LASSO-modellen

Ved nowcasting har vi tilgængelige observationer af vores forklarende variable for den tidsperiode, vi prædiktorer responsvariablen til, så derfor er notationen her en smule anderledes end ved forecasting. Nu angiver T perioden 2016-Q1, og X indeholder en række med observationer for denne, sådan at prædiktorerne har længde 225. Vi prædikterer således

$$\widehat{y}_t^{\text{nc}} \coloneqq \mathbb{E}\left[y_t \middle| \mathcal{F}_t^{\text{nc}} \right], \tag{8.2}$$

hvor $\mathcal{F}_t^{\text{nc}} = \{X_t, y_{t-1}, X_{t-1}, \ldots, y_1, X_1\}$, ud fra nowcast-ligningen

$$\widehat{y}_t^{\text{nc}} = \widehat{\alpha}^{\text{nc}} + Z_t^{\text{nc}} \widehat{\boldsymbol{\beta}}^{\text{nc}}, \tag{8.3}$$

hvor Z^{nc} er en $T \times (N + p)$ matrix, der består af X og $L\boldsymbol{y}, L^2\boldsymbol{y} \ldots, L^p\boldsymbol{y}$. Bemærk, at de medtagede værdier af y_t er laggede en ekstra gang i forhold til ved forecasting.

Hvis vi betragter korrelationen mellem ændringen i `GDPC96` og tidsrækkerne i `FRED-MD` hver især, kan vi se, at der overordnet er større korrelation til samtidige observationer end til tidligere observationer; de gennemsnitlige absolutte korrelationer er hhv.

$$\frac{1}{N} \sum_{i=1}^{N} \left| \widehat{\rho}_{y,X_i}(t,t) \right| = 0.317 \tag{8.4}$$

$$\frac{1}{N} \sum_{i=1}^{N} \left| \widehat{\rho}_{y,X_i}(t,t-1) \right| = 0.203. \tag{8.5}$$

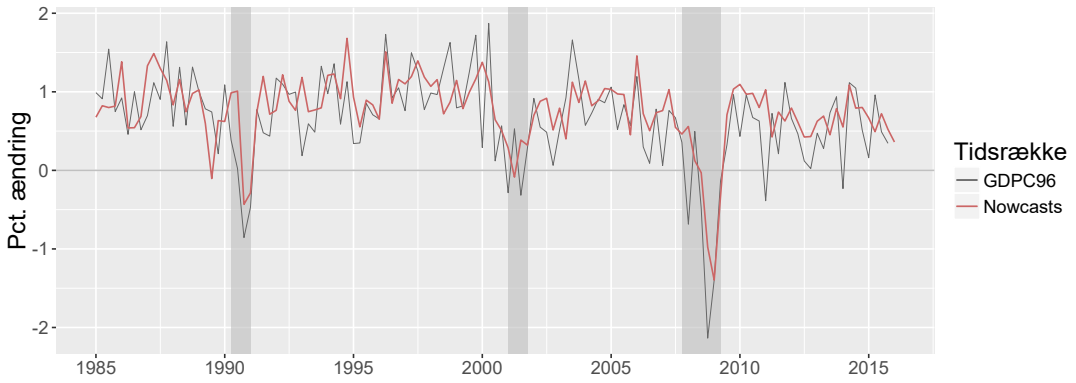
Her angiver $\widehat{\rho}$ den empiriske krydskorrelationsfunktion, som anvender stikprøveudgaver af middelværdier, varianser og kovarianser. Fordi vi ser større korrelationer til samtidige observationer, forventer vi, at vores nowcasts overordnet vil være bedre end vores forecasts.

Svarende til MSFE under forecasting, betragter vi her *mean square nowcast error* (MSNE), som for de iterative nowcasts under LASSO-modellen evalueres til

$$MSNE_{\text{LASSO}} = 1.986\,\text{E-}05. \tag{8.6}$$

De iterative nowcasts er plottet sammen med observerede værdier på Figur 8.3, inkl. out-of-sample nowcastet for 2016-Q1. I forhold til de tilsvarende forecasts følger de nowcastede værdier udsvingene i højere grad. I Tabel 8.2 er de 10 prædiktorer, som blev udvalgt af LASSO-modellen ved flest nowcasts, angivet. Herunder er også angivet, hvor mange gange det konkret drejer sig om, samt hvor stor en procentdel, dét er ud af det samlede antal iterationer. Overordnet har de fleste af dem en stor korrelation med responsvariablen. En detaljeret korrelationsmatrix for $\boldsymbol{y}$ og top 10-prædiktorerne er angivet i Appendiks B.

I modsætning til ved forecasting ser vi under de iterative nowcasts, at der i starten (1985-2000) nogle steder bliver medtaget laggede værdier iblandt de signifikante forklarende variable. Alle de pågældende steder drejer det sig udelukkende om 3. lag, $L^3\boldsymbol{y}$, og dette lag rangerer som #19 iblandt de oftest inkluderede prædiktorer. Modellen ser således ingen signifikante bidrag fra 1. og 2. lag på noget tidspunkt.



Figur 8.3: Nowcastede værdier af væksten i `GDPC96` plottet sammen med observerede værdier

Prædiktor	Kor. med $\boldsymbol{y}$	#	Pct.	Beskrivelse
ISRATIOx	-0.588	125	100 %	Lager/salg-kvotient for alle virksomheder
DPCERA...	0.563	125	100 %	Private forbrugsudgifter
INDPRO	0.712	125	100 %	Industriel produktionsindeks
CLAIMSx	-0.563	125	100 %	Antal nye modtagere af arbejdsløshedsstøtte
RPI	0.525	123	98 %	Personlig indkomst
TB3SMFFM	0.301	121	97 %	3-måneders treasury bonds minus diskonto
UEMPLT5	-0.360	121	97 %	Antal arbejdsløse, under 5 uger
S.P.500	0.293	112	90 %	S&P 500 indeks
IPFPNSS	0.708	102	82 %	Industriel prod. af ikke-industrielle varer
AWOTMAN	0.454	87	70 %	Gns. overarbejdstimer, produktionssektoren

Tabel 8.2: Top 10 oftest inkluderede prædiktorer i de iterative nowcasts med LASSO-modellen

Vi vil herunder kort beskrive det sidste nowcasts i rækken, som er out-of-sample nowcastet for 2016-Q1, i lidt flere detaljer. Værdien for dette nowcast er:

$$0.361\%. \tag{8.7}$$

Modellen udvælger ni signifikante variable til at bestemme dette, og de er beskrevet i Tabel 8.3. Disse prædiktorer har overordnet med produktion, beskæftigelse, privatøkonomi og aktiemarkedet at gøre.

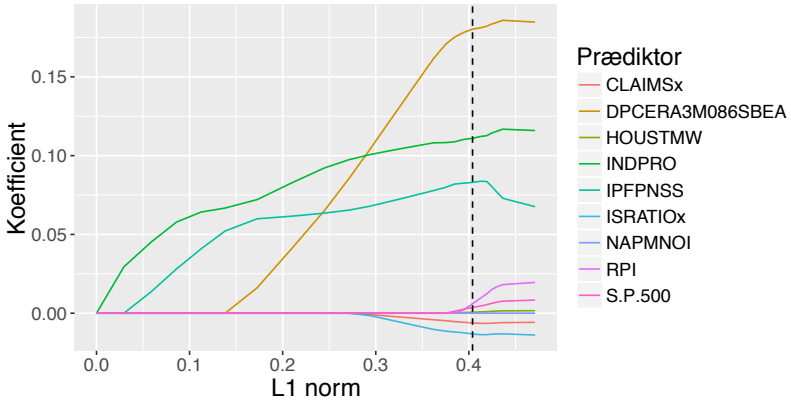
Ikke overraskende ses det, at seks af dem går igen fra top 10-listen. Ifølge LASSO-modellen ser vi således ikke signifikante bidrag til udviklingen i `GDPC96` i 2016-Q1 fra variable, der har med renter eller priser at gøre.

Udviklingen af estimaterne kan ses på Figur 8.4, hvor størrelsen af koefficienterne er plottet over for størrelsen af begrænsningen i minimeringsproblemet. Ved den hårdest straffede model medtages kun tidsrækken `INDPRO`, hvilket indikerer, at den i LASSO-modellen giver det vigtigste bidrag ved nowcasting. Som straffen lempes, medtages

Prædiktor	$\widehat{\beta}$	Beskrivelse
NAPMNOI	6.184696 E-05	ISM-index: nye ordrer
ISRATIOx	-1.315079 E-02	Lager/salg-kvotient for alle virksomheder
HOUSTMW	5.990749 E-04	Nye boliger påbegyndt konstruktion
RPI	5.994419 E-03	Personlig indkomst
DPCERA...	1.802826 E-01	Private forbrugsudgifter
INDPRO	1.110263 E-01	Industriel produktionsindeks
IPFPNSS	8.301466 E-02	Industriel prod. af ikke-industrielle varer
CLAIMSx	-6.196374 E-03	Antal nye modtagere af arbejdsløshedsstøtte
S.P.500	3.649628 E-03	S&P 500 indeks

Tabel 8.3: β-estimaterne i den optimale model

dernæst `IPFPNSS`. Begge disse prædiktorer har høj korrelation med $\boldsymbol{y}$ og har begge med produktion at gøre.



Figur 8.4: LASSO-stien for nowcastet til 2016-Q1; kun de 9 udvalgte prædiktorer er vist. Den vertikale streg angiver ℓ_1-normen tilhørende $\widehat{\boldsymbol{\beta}}$

9

Modellering med faktor-modellen

For faktor-modellen er fremgangsmåden, at for $t = 100, 101, \ldots, T - h$:

1. Estimér r faktorer for $r = 1, \ldots, r_{max}$ ud fra $X_1, X_2, \ldots, X_t$
2. Vælg bedste r ud fra ét af informationskriterierne beskrevet i Afsnit 5.3.2
3. Vælg antal laggede værdier p ud fra $\text{AIC} = t \log(\widehat{\sigma}^2) + 2p$
4. Estimér parametre $\widehat{\alpha}, \widehat{\boldsymbol{\beta}}_F, \widehat{\boldsymbol{\beta}}_w$ med OLS
5. Forecast $\widehat{y}^h_{t+h|t}$ ud fra forecast-ligningen

$$\widehat{y}^h_{t+h|t} = \widehat{\alpha}^h + F_t \widehat{\boldsymbol{\beta}}^h_F + \sum_{j=1}^{p} \widehat{\beta}^h_{w,j} y_{t-j+1} \tag{9.1}$$

6. Gem forecast-fejl $y^h_{t+h} - \widehat{y}^h_{t+h|t}$

og MSFE beregnes til slut som givet i (7.1). Centrale dele af den anvendte kode kan ses i Appendiks A.4.

9.1 Forecasting med faktor-modellen

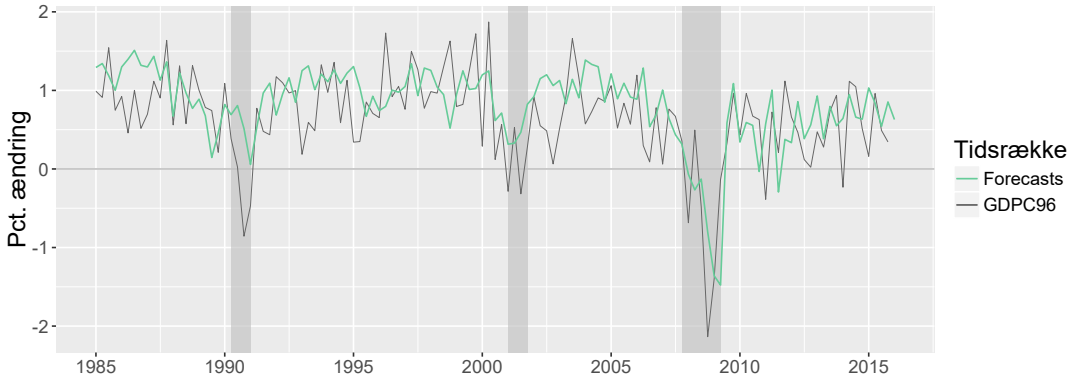
For at bestemme, hvilket informationskriterie, vi vil bruge til at finde r i modellen, ser vi på, hvilken af dem der giver det laveste MSFE. I Tabel 9.1 er for hvert af kriterierne vist MSFE for forecasts af hhv. kvartalsmæssig og årlig vækst i `GDPC96`.

	IC_1	IC_2	IC_3
$\widehat{y}^1_{T+1\|T}$	0.586%	0.629%	0.907%
Antal faktorer / lags til tid T	5 / 1	8 / 1	21 / 1
$MSFE^1_{\text{FAKTOR}}$	2.872 E-05	2.775 E-05	3.614 E-05
$\widehat{y}^4_{T+4\|T}$	2.674%	2.492%	3.273%
Antal faktorer / lags til tid T	5 / 0	8 / 0	21 / 0
$MSFE^4_{\text{FAKTOR}}$	3.406 E-04	2.905 E-04	3.909 E-04

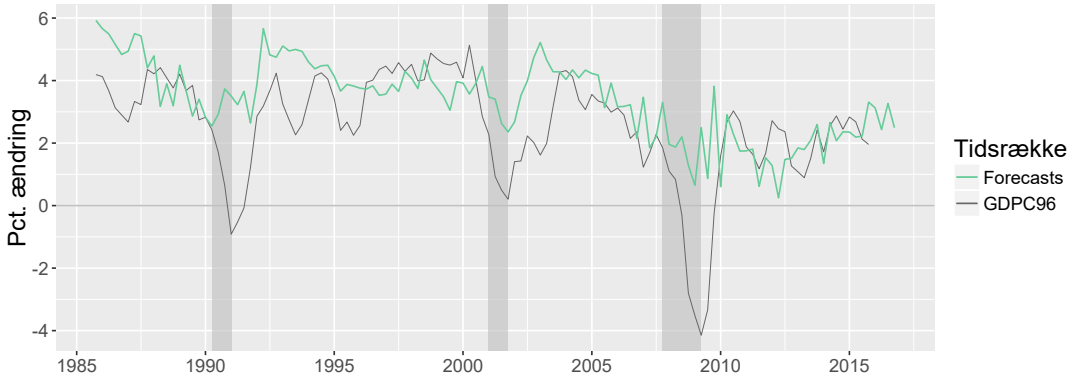
Tabel 9.1: Resultater for forecasting med faktor-modellen

Antallet af faktorer og lags til hver model vælges meget ensartet henover hele forecast-perioden, og vi har i tabellen derfor kun anført de nyeste. Generelt vælger IC_1 få faktorer, IC_2 lidt flere, og IC_3 vælger det maksimale antal. Eftersom vi kun forecaster 125 datapunkter, når vi altså ikke at se den forventede konvergens til samme antal faktorer.

For både kvartalsmæssig og årlig vækst giver IC_2 den laveste MSFE, og derfor vælger vi at fokusere på disse. Kvartalsmæssige og årlige forecasts for faktor-modellen med r bestemt ved IC_2 er hhv. vist på Figur 9.1 og 9.2.



Figur 9.1: Faktor-modellens forecasts for den kvartalsmæssige vækst i `GDPC96` vist sammen med de observerede værdier



Figur 9.2: Faktor-modellens forecasts for den årlige vækst i `GDPC96` vist sammen med de observerede værdier

9.2 Nowcasting med faktor-modellen

Ligesom ved nowcasting med LASSO-modellen angiver T nu perioden 2016-Q1, og vi vil prædiktere iterativt med nowcast-ligningen

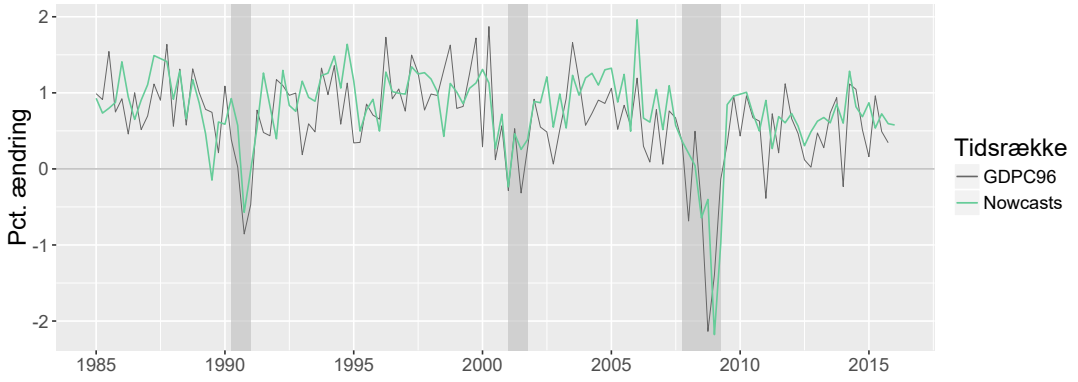
$$\widehat{y}_t^{\text{nc}} = \widehat{\alpha}^{\text{nc}} + \widehat{F}_t^{\text{nc}} \widehat{\boldsymbol{\beta}}_F^{\text{nc}} + \sum_{j=1}^{p} \widehat{\beta}_{w,j}^{\text{nc}} y_{t-j}. \tag{9.2}$$

Resultater for nowcasting med faktor-modellen er vist i Tabel 9.2. Tendensen ved forecasting går igen ved nowcasting: IC_1 og IC_2 vælger relativt få faktorer, mens IC_3 vælger maksimalt antal, og IC_2 giver den laveste MSNE.

	IC_1	IC_2	IC_3
$\widehat{y}_T^{\text{nc}}$	0.577%	0.577%	0.702%
Antal faktorer / lags til tid T	6 / 3	6 / 3	21 / 3
$MSNE_{\text{FAKTOR}}$	2.250 E-05	2.177 E-05	2.737 E-05

Tabel 9.2: Resultater for nowcasting med faktor-modellen

Vi kan se, at for det sidste nowcast udvælger IC_1 og IC_2 det samme antal faktorer, og dermed producerer det samme nowcast. Dette har de dog ikke gjort henover hele nowcast-perioden, og dette afspejles af, at deres MSNE er forskellige. På Figur 9.3 er vist iterative nowcasts med faktor-modellen, som udvælger r ud fra IC_2.



Figur 9.3: Faktor-modellens nowcasts for den kvartalsmæssige vækst i `GDPC96` vist sammen med de observerede værdier

10

Sammenligning af resultater

For at teste om to modellers forecasts er signifikant forskellige, vil vi anvende *Diebold-Mariano-testen* [Diebold, 2013], som sammenligner forecast-fejlene. Vi antager, at vi har forecasts for $t = t_0, \ldots, T$, og betegner de to modellers forecast-fejl for hhv. $\widehat{\epsilon}^1_{t+h|t}$ og $\widehat{\epsilon}^2_{t+h|t}$. Vi tester nul-hypotesen

$$\mathcal{H}_0 : \mathbb{E}\left[\xi\left(\widehat{\epsilon}^1_{t+h|t}\right)\right] = \mathbb{E}\left[\xi\left(\widehat{\epsilon}^2_{t+h|t}\right)\right] \tag{10.1}$$

imod den alternative hypotese

$$\mathcal{H}_1 : \mathbb{E}\left[\xi\left(\widehat{\epsilon}^1_{t+h|t}\right)\right] \neq \mathbb{E}\left[\xi\left(\widehat{\epsilon}^2_{t+h|t}\right)\right], \tag{10.2}$$

hvor $\xi(\cdot)$ betegner en tabsfunktion. Typisk anvendes $\xi\left(\widehat{\epsilon}^i_{t+h|t}\right) = \left(\widehat{\epsilon}^i_{t+h|h}\right)^2$ eller $\xi\left(\widehat{\epsilon}^i_{t+h|h}\right) = \left|\widehat{\epsilon}^i_{t+h|h}\right|$. Definér

$$d_t = \xi\left(\widehat{\epsilon}^1_{t+h|t}\right) - \xi\left(\widehat{\epsilon}^2_{t+h|t}\right), \tag{10.3}$$

hvormed $\mathcal{H}_0 : \mathbb{E}[d_t] = 0$. Diebold-Mariano teststatistikken DM er givet ved

$$DM = \frac{\bar{d}}{\sqrt{\frac{2\pi \widehat{f}_d(0)}{T - t_0}}}, \tag{10.4}$$

hvor

$$\bar{d} = \frac{1}{T - t_0} \sum_{t=t_0}^{T} d_t, \tag{10.5}$$

$$\widehat{f}_d(0) = \begin{cases} \frac{1}{2\pi}\left(\widehat{\gamma}_d(0) + 2\sum_{k=1}^{h-1} \widehat{\gamma}_d(k)\right), & \text{hvis } h > 1, \\ \frac{1}{2\pi}\widehat{\gamma}_d(0), & \text{hvis } h = 1. \end{cases} \tag{10.6}$$

Størrelsen $\widehat{f}_d(0)$ er ifølge [Diebold, 2013] et konsistent estimat af spektraltætheden for d som defineret i [Shumway & Stoffer, 2011, s. 181]. Under antagelse af, at d er stationær, gælder det, at

$$DM \xrightarrow{\mathcal{D}} N(0, 1), \tag{10.7}$$

og nulhypotesen afvises således ved niveau 5 %, hvis $|DM| > 1.96$.

10.1 Forecasting-resultater

I dette kapitel vil vi fortolke og sammenligne vores resultater fra de forskellige modeller. Vi starter med at betragte de relative MSFE, dvs. forholdet mellem en given models MSFE og den tilsvarende MSFE fra benchmark-modellen. Disse er vist i Tabel 10.1 (tal under 1.000 betyder, at modellen klarer sig bedre end benchmark-modellen).

	$h = 1$	$h = 4$
Benchmark	1.000	1.000
LASSO	1.004	1.208*
Faktor	0.946	0.986

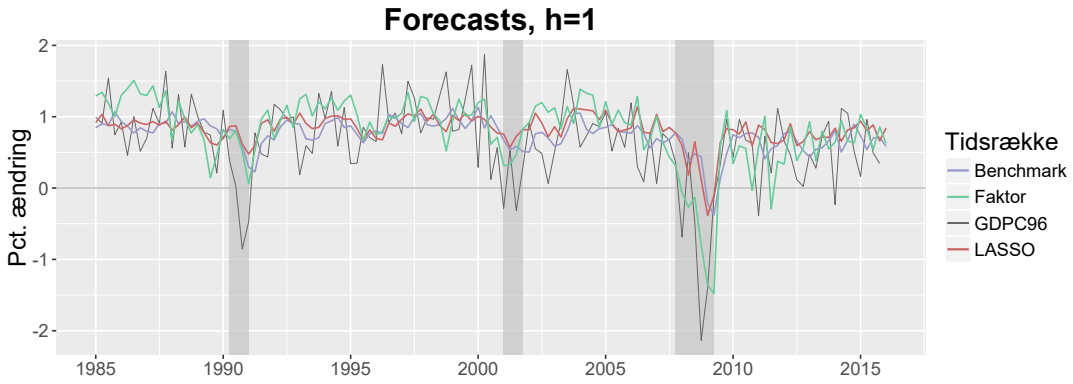
Tabel 10.1: Relative MSFE i forhold til benchmark-modellen

Som det fremgår af tabellen, er LASSO-modellen omtrent lige så god til at forecaste ét kvartal frem som benchmark-modellen, men får markant større forecast-fejl ved forecasting et år frem. Denne er i tabellen markeret med *, fordi en Diebold-Mariano-test afviser, at forecast-fejlene ikke er signifikant forskellige fra benchmark-modellens forecast-fejl ved niveau 5 % (p-værdi: 1.64 %). Alle de andre forecast-fejl var ifølge testen ikke signifikant forskellige. Vi havde en formodning om, at LASSO-modellen ville klare sig bedre, idet den udover laggede værdier betinger på andre tidsrækker, så resultaterne overrasker. Årsagen er ikke direkte klar, men det kunne f.eks. skyldes at estimeringsmetoden i LASSO-modellen egner sig bedre til tværsnitsdata end tidsrækkedata. Det kan også have indflydelse, at vi har relativt få observationer, som vi forecaster.

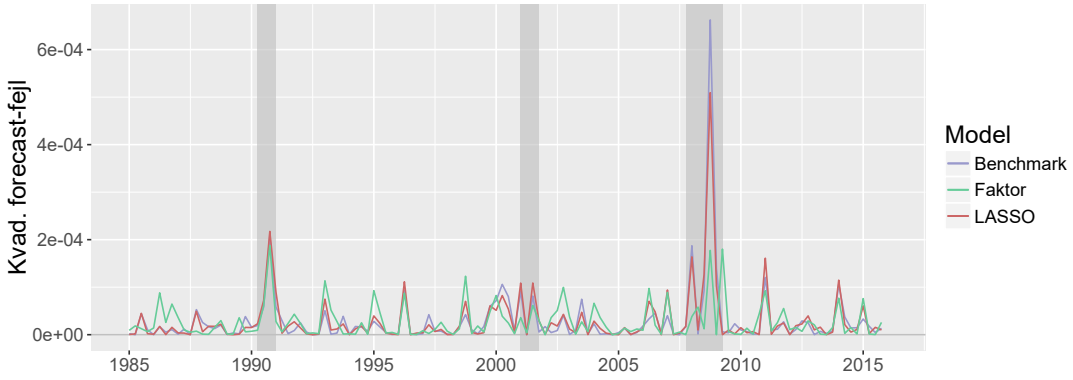
For faktor-modellen ser vi derimod en lille forbedring ved forecasting ét kvartal frem, og en marginal forbedring ved forecasting et år frem. Grunden til, at den klarer sig bedre end LASSO-modellen kan ligge i, at faktorerne er estimerede til at forklare den samlede varians, hvorimod LASSO blot udvælger nogle repræsenterende tidsrækker. Desuden er estimeringsmetoden mere direkte sammenlignlig med benchmark-modellen, idet de begge estimerer $\boldsymbol{\beta}^h$ med urestringeret OLS.

På Figur 10.1 er kvartalsmæssige forecasts for de tre modeller plottet sammen. Vi kan se, at benchmark- og LASSO-forecasts følges nogenlunde ad, men begge forecaster konservativt mht. størrelsen af udsvingene. Faktor-modellens forecasts har større udsving og rammer oftere tæt på peaks i grafen.

På Figur 10.2 er de tre modellers kvadrerede fejl ved kvartalsmæssig forecasting plottet sammen. Generelt bliver fejlene større ved recessionerne, hvilket også er at forvente, idet kriser er svære at forudsige. De tre modellers forecast-fejl ligger overordnet tæt på 0 de samme steder og har peaks de samme steder. Faktor-modellen har i recessionsperioder mindre forecast-fejl end de andre modeller, især ved finanskrisen i 2008. Til gengæld har den større fejl uden for kriserne.

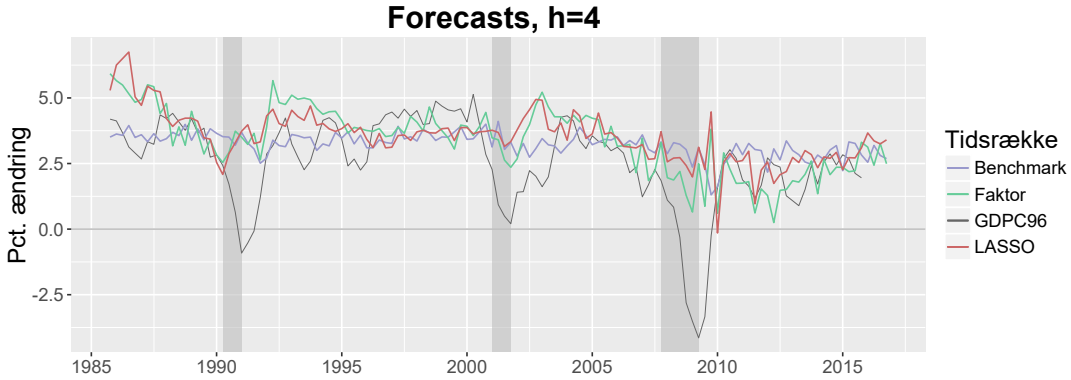


Figur 10.1: Kvartalsmæssige forecasts for de forskellige modeller vist sammen med observerede værdier

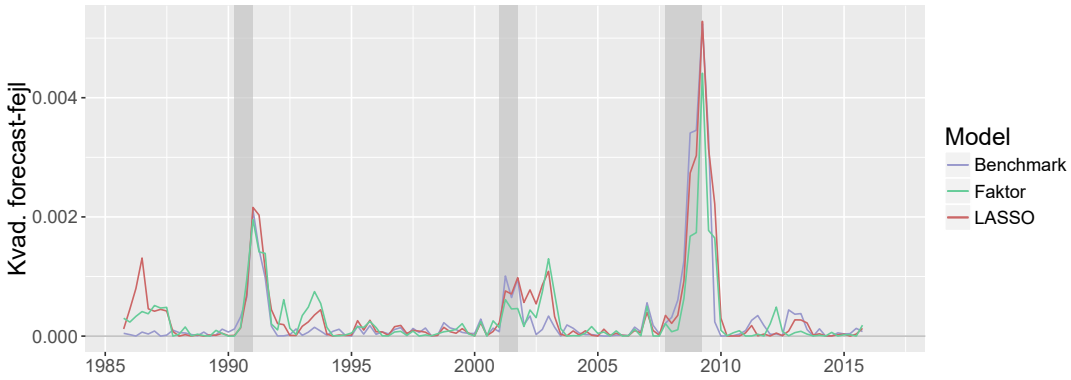


Figur 10.2: Kvadrerede fejl ved kvartalsmæssig forecasting for de forskellige modeller

På Figur 10.3 er de tre modellers årlige forecasts vist sammen. Her ser vi en større forskel mellem de tre modellers forecasts i kraft af den øgede usikkerhed ved forecasting et helt år frem, men de overordnede tendenser er stadig meget lig hinanden. På Figur 10.4 ses det, at de tre modeller får udsving i forecast-fejl omkring de samme tidsperioder, ligesom før. Ingen af modellerne forudsiger recessionerne i særlig høj grad, og dette fører til de store forecast-fejl i slutningen af de pågældende perioder.



Figur 10.3: Årlige forecasts for de forskellige modeller vist sammen med observerede værdier



Figur 10.4: Kvadrerede fejl ved årlig forecasting for de forskellige modeller

10.2 Nowcasting-resultater

Når vi betragter nowcasts, kan vi ikke sammenligne med benchmark-modellen, idet benchmark-modellen ikke kan nowcaste. Vi ser derfor i stedet på forbedringen fra forecasting til nowcasting for hhv. LASSO- og faktor-modellen. De relative forbedringer er vist i Tabel 10.2.

Begge modeller nowcaster som forventet bedre, end de forecaster. LASSO-modellen ser en større forbedring end faktor-modellen, og en Diebold-Mariano-test indikerer, at

	$MSNE/MSFE$[1]	p-værdi for DM test
LASSO	0.674	2.118 %
Faktor	0.785	5.257 %

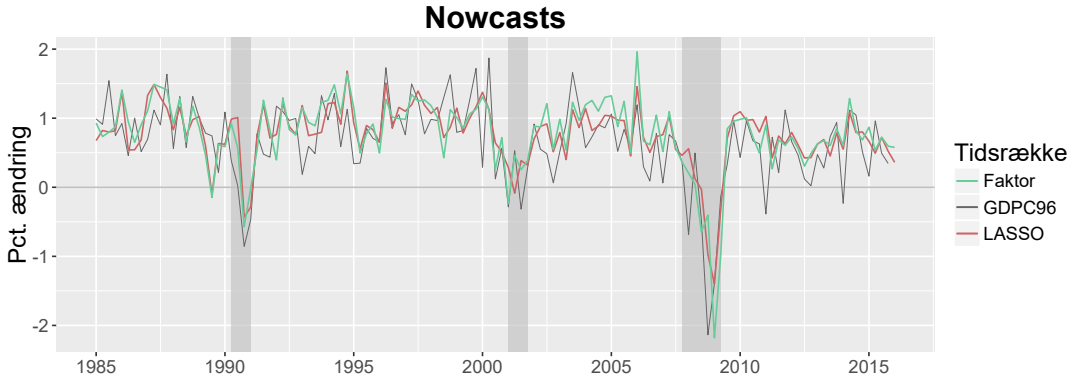
Tabel 10.2: Relative forbedringer ved nowcasting og Diebold-Mariano testresultater

now- og forecasts for LASSO-modellen er signifikant forskellige ved niveau 5 %, mens de for faktor-modellen næsten er signifikant forskellige.

Til sammenligning af de to modellers nowcasts går vi tilbage og ser på værdien af deres MSNE over for hinanden:

$$\frac{MSNE_{\text{LASSO}}}{MSNE_{\text{FAKTOR}}} = \frac{1.986\,\text{E-}05}{2.177\,\text{E-}05} = 0.912. \tag{10.8}$$

Modsat ved forecasting ser vi her, at LASSO-modellen får mindre nowcast-fejl end faktor-modellen. En Diebold-Mariano-test kan ikke afvise, at de to modellers nowcasts ikke er signifikant forskellige, men fordi vi trods alt har en 9 % forskel i MSNE, vil vi stadig fortolke det således, at LASSO-modellen nowcaster bedre.

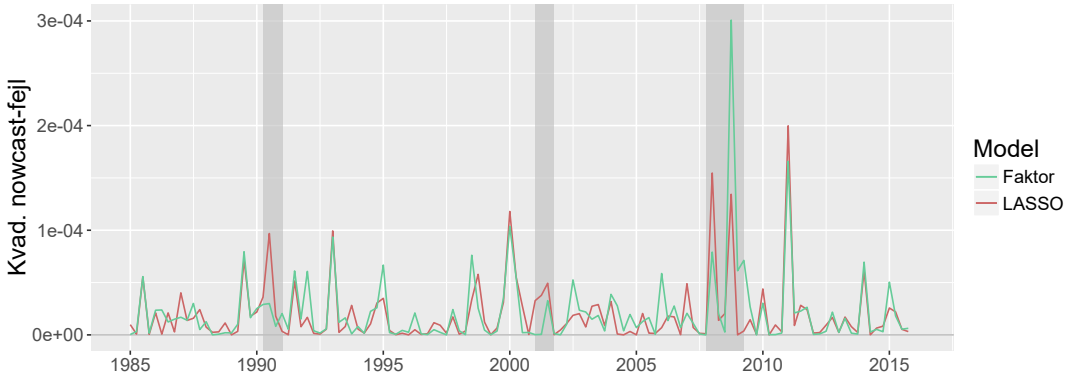


Figur 10.5: Nowcasts for LASSO- og faktor-modellen vist sammen med observerede værdier

På Figur 10.5 er nowcasts for de to modeller vist sammen. Vi kan se, at modellerne følger hinanden lidt tættere ved nowcasting, end de gjorde ved forecasting. Dette var også at forvente, idet de begge nowcaster bedre, end de forecaster.

På Figur 10.6 er vist de kvadrerede nowcast-fejl for de to modeller. Generelt er niveauet for fejlene nogenlunde det samme som ved forecasting uden for recessionsperioder. Under recessioner ser vi markante forbedringer i præcision i forhold til forecasting for begge modeller — pånær i 2008-Q4, hvor faktor-modellen faktisk har en lavere forcast-fejl end nowcast-fejl. Dette virker overraskende, da den i de andre recessioner får endnu mindre nowcast-fejl end LASSO-modellen.

I Tabel 10.3 er prædiktionerne for de forskellige modeller angivet samlet. LASSO-modellens forecasts er mere optimistiske end benchmark og faktor. Europa-Kommissionen forecaster til sammenligning den årlige vækst for USA's BNP i 2016 til 2.7 % [European Commission, 2016, s. 1].



Figur 10.6: Kvadrerede nowcast-fejl for LASSO- og faktor-modellen vist sammen med recessionsperioder

	Forecast, 2016-Q1	Forecast, hele 2016	Nowcast, 2016-Q1
Benchmark	0.587 %	2.685 %	—
LASSO	0.840 %	3.401 %	0.361 %
Faktor	0.629 %	2.492 %	0.577 %

Tabel 10.3: Forecasts og nowcasts for de forskellige modeller

11

Konklusion

Vi undersøgte, hvorvidt LASSO- og faktor-modellerne kunne forbedre på forecasts af vækst i amerikansk BNP i forhold til en simpel autoregressionsmodel. Det viste sig, at vores LASSO-model ikke kunne forecaste bedre end AR-modellen, for begge de forecast-horisonter, vi betragtede. Den havde omtrent lige så store forecast-fejl som AR-modellen ved kvartalsmæssige forecasts og ved forecasts af årlig vækst havde den signifikant større forecast-fejl.

Faktor-modellen udviste til gengæld mindre forecast-fejl for de kvartalsmæssige forecasts, og marginalt mindre fejl for årlige forecasts. Denne model forecastede desuden recessionsperioder bedre end de to andre modeller, hvilken især var tydeligt ved forecasts under finanskrisen i 2008.

Vi undersøgte også LASSO- og faktor-modellernes evne til at nowcaste, og her så vi en betydelig forbedring i præcision ved blot at betinge på én yderligere måneds data i forhold til ved forecasting. Overraskende var det her LASSO-modellen, der klarede sig bedst, idet faktor-modellen fik en højere nowcast-fejl under finanskrisen i 2008, end den gjorde ved forecasting. Baseret på datasættet `FRED-MD` opgjort for marts 2016 og vores analyse, er vores bedste prædiktioner således hhv.

- Kvartalsmæssig vækst, 2016-Q1: **0.361 %** (nowcastet med LASSO-modellen)
- Årlig vækst, hele 2016: **2.492 %** (forecastet med faktor-modellen)

11.1 Perspektivering

Der findes mange metoder til at udvide disse modeller, og de kan alle muligvis være med til at forbedre på præcisionen af både forecasts og nowcasts. De mest kendte udvidelser er *adaptive LASSO*, hvor prædiktorene ikke straffes ens, og *dynamisk faktor-model*, hvor lags af faktorene også medtages.

I henhold til problemafgrænsningen (Afsnit 1.1) aggregerede vi alle de forklarende variable til kvartalsbasis på samme måde. En oplagt udvidelse her ville være at undersøge, hvilke typer af data, de forskellige variable repræsenterer, og aggregere data af de forskellige typer på måder, som bedst reflekterer observationen for et helt kvartal. En anden måde at gribe problemstillingen med forskellig frekvens af observationer an på, ville være at inddrage en model, som direkte kan regressere BNP på højere-frekvent data, f.eks. *Mixed Data Sampling* (MIDAS).

Vi kunne også udvide datasættet med forklarende variable ved at medtage én eller flere laggede værdier af hver enkelt variabel. Endeligt kunne vi betragte forecasting med *rullende vindue*, hvor antallet af tidsperioder, som en given model tilpasser over, er fast, hvormed ældre data løbende "glemmes". Dette kunne forhindre, at historiske tendenser, som ikke længere var relevante, havde indflydelse på forecastet.

Del III

Appendicer

A

Central R-kode

Herunder er vist de centrale funktioner, vi har anvendt i rapporten. De udgør kun en lille del af den samlede kode, og de resterende filer (inkl. anvendte datasæt) kan findes i den medfølgende `zip`-fil.

A.1 Databehandlings-scripts

```r
# ./functions/misc/M2Q.R

Monthly2Quarterly <- function(dat, agg.method) {
  # Aggregates monthly data into quarterly data
  # based on selected aggregation method
  #
  # Args:
  #   dat: tbl_df with the structure [DATE, X], where
  #     $DATE: vector representing time (as.Date)
  #     X: matrix of time series
  #   agg.method: method used to aggregate the data
  #
  # Returns:
  #   df.out: tbl_df with the same structure as `dat`

  if (is.null(dat)) return(NULL)
  if (is.null(agg.method)) agg.method <- "mean"

  del.dates <- function(x) {
    for (i in seq_along(x)) {
      if (i %% 3 != 1) {
        x[i] <- NA
      }
    }
    return(x)
  }

  dates <- dat %>%
    select(DATE) %>%
    mutate_each(funs(del.dates))

  agg <- function(x) {
    n.month <- length(x)
    x.out <- rep(NA, n.month)

    if (agg.method == "mean") {
      n.quart <- ceiling(n.month / 3)
      for (i in 1:n.quart) {
        t <- i * 3
```

```
40          if (!all(is.na(x[(t-2):t]))) {
41            x.out[t-2] <- mean(x[(t-2):t], na.rm = TRUE)
42          }
43        }
44      } else if (agg.method == "first") {
45        for(i in 1:n.month) {
46          if (i %% 3 == 1) {
47            x.out[i] <- x[i]
48          }
49        }
50      }
51
52      return(x.out)
53    }
54
55    series <- dat %>%
56      select(-DATE) %>%
57      mutate_each(funs(agg))
58
59    df.out <- bind_cols(dates, series) %>%
60      filter(!is.na(DATE))
61
62    return(df.out)
63  }
```

R-script A.1: Funktion til at aggregere månedlige data til kvartalsdata

A.2 Benchmark-scripts

```
1  # ./functions/benchmark/benchmark_forecast.R
2
3  BenchmarkAIC <- function(y, X) {
4    # Calculate AIC for some different values of p
5    #
6    # Args:
7    #   y: response vector
8    #   X: matrix with lagged values of y
9    #
10   # Returns:
11   #   AICs: vector of AIC values
12
13   AIC <- function(y, X, p) {
14     n.obs <- length(y)
15     X.p <- X[, 1:(p+1)]
16     beta.p <- solve(crossprod(X.p), crossprod(X.p, y))
17     sigma2est <- mean((y - X.p %*% beta.p)^2)
18     AIC <- log(sigma2est) + (n.obs + 2 * p) / n.obs
19     return(AIC)
20   }
21
22   p.max <- 4L
23   AICs <- rep(NA, p.max)
```

```r
  for (p in 1:p.max) {
    AICs[p] <- AIC(y, X, p)
  }
  return(AICs)
}


BenchmarkFc <- function(y, X.df, h) {
  # forecast h periods ahead with the benchmark model
  #
  # Args:
  #   y: response variable to time t, depends on h
  #   X.df: predictor matrix with lags as data_frame
  #   h: forecast horizon
  #
  # Returns:
  #   list.out: list containing
  #     $forecast: value of forecast of y[t+h]
  #     $ar.order: number of lags included by the model
  #     $sigma2: variance estimate

  n.obs <- length(y)

  fc.resp <- y[(1+h):n.obs]
  fc.vars <- as.matrix(X.df[1:(n.obs-h), ])

  AICs <- BenchmarkAIC(fc.resp, fc.vars)
  p.opt <- which.min(AICs)
  X.opt <- fc.vars[, 1:(p.opt+1)]
  beta.opt <- solve(crossprod(X.opt), crossprod(X.opt, fc.resp))
  sigma2est <- mean((fc.resp - X.opt %*% beta.opt)^2)

  fc.vars.T <- X.df %>% tail(n = 1) %>%
    select(1:(p.opt+1)) %>% collect %>% as.matrix
  y.fc <- drop(fc.vars.T %*% beta.opt)

  list.out <- list(
    "forecast" = y.fc,
    "ar.order" = p.opt,
    "sigma2" = sigma2est
  )
  return(list.out)
}
```

R-script A.2: Funktioner til forecasting med benchmark-modellen

A.3 LASSO-scripts

```r
# ./functions/lasso/lasso_forecast.R

LassoBIC <- function(y, X, fit) {
  # Helper function for computing BIC
  #
  # Args:
  #   y: response vector
  #   X: predictor matrix
  #   fit: `glmnet` fit
  #
  # Returns:
  #   BICs: vector of BICs corresponding to each value of lambda

  BIC <- function(y, X, fit, i) {
    n.obs <- length(y)
    sigma2est <- sum((y - fit$a0[i] - X %*% fit$beta[,i])^2) / n.obs
    BIC <- log(sigma2est) + fit$df[i] * log(n.obs) / n.obs
    return(BIC)
  }

  n.lambda <- length(fit$lambda)
  BICs <- rep(NA, n.lambda)
  for (i in 1:n.lambda) {
    BICs[i] <- BIC(y, X, fit, i)
  }
  return(BICs)
}


LassoFc <- function(y, X.df, h) {
  # Forecasts y[t+h] using the variables in X
  #
  # Args:
  #   y: response vector
  #   X.df: predictor matrix as a data_frame
  #   h: forecast horizon
  #
  # Returns:
  #   list.out: list containing
  #     $forecast: forecast of y[t+h]
  #     $n.pred: number of nonzero betas
  #     $lambda: lambda value for the chosen model

  n.obs <- length(y)

  fc.resp <- y[(1+h):n.obs]
  fc.vars <- as.matrix(X.df[1:(n.obs-h), ])
  fc.fit <- glmnet(fc.vars, fc.resp)

  BICs <- LassoBIC(fc.resp, fc.vars, fc.fit)
  idx.opt <- which.min(BICs)
  lambda.opt <- fc.fit$lambda[idx.opt]
```

```
  beta.opt <- as.vector(fc.fit$beta[, idx.opt])
  n.pred <- fc.fit$df[idx.opt]
  nonzeros <- which(!isZero(beta.opt))
  lasso.vars <- colnames(X.df)[nonzeros]
  beta.hat <- beta.opt[nonzeros]

  colNums <- match(lasso.vars, names(X.df))

  fc.vars.T <- X.df %>%
    tail(n = 1) %>%
    select(colNums)

  y.fc <- drop(as.matrix(fc.vars.T) %*% beta.hat + fc.fit$a0[idx.opt])

  list.out = list(
    "forecast" = y.fc,
    "n.pred" = n.pred,
    "lambda" = lambda.opt
  )
  return(list.out)
}
```

R-script A.3: Funktioner til forecasting med LASSO-modellen

```
# ./functions/lasso/lasso_nowcast.R

LassoNc <- function(y, X.df) {
  # Nowcasts y[t] using the variables in X
  #
  # Args:
  #   y: response vector
  #   X.df: predictor matrix as a data_frame
  #
  # Returns:
  #   list.out: list containing
  #     $nowcast: nowcasted value of y[t]
  #     $n.pred: number of nonzero betas
  #     $lambda: lambda value for chosen model
  #     $names: vector with names of the significant variables

  n.obs <- length(y)

  resp <- y
  vars <- as.matrix(X.df[1:n.obs, ])
  nc.fit <- glmnet(vars, resp)

  BICs <- LassoBIC(resp, vars, nc.fit)

  idx.opt <- which.min(BICs)
  lambda.opt <- nc.fit$lambda[idx.opt]
  n.pred <- nc.fit$df[idx.opt]

  beta.opt <- as.vector(nc.fit$beta[, idx.opt])
  nonzeros <- which(!isZero(beta.opt))
```

```
  lasso.vars <- colnames(X.df)[nonzeros]
  beta.hat <- beta.opt[nonzeros]

  nc.vars <- as.matrix(tail(X.df, n = 1))[, nonzeros]

  y.nc <- drop(nc.vars %*% beta.hat + nc.fit$a0[idx.opt])

  list.out = list(
    "nowcast" = y.nc,
    "n.pred" = n.pred,
    "lambda" = lambda.opt,
    "names" = lasso.vars
  )
  return(list.out)
}
```

R-script A.4: Funktioner til nowcasting med LASSO-modellen

A.4 Faktor-scripts

```
# ./functions/factor/est_factor

GetFactors <- function(X, r) {
  # determine factors in factor model given r
  #
  # Args:
  #   X: matrix of regressor time series (T * N)
  #   r: number of factors
  #
  # Returns:
  #   list.out: list containing
  #     $factors: (T * r) matrix containing factors
  #     $loadings: (N * r) matrix containing loadings

  n.var <- dim(X)[2]
  XTX <- crossprod(X)

  X.eig <- eigen(XTX, symmetric = TRUE)
  eig.vec <- X.eig$vectors

  loadings <- eig.vec[, 1:r] * sqrt(n.var)
  factors <- (X %*% loadings) / n.var

  list.out <- list(
    "factors" = factors,
    "loadings" = loadings
  )
  return(list.out)
}


EstFactors <- function(X.df, ic = 1, trace = FALSE) {
```

```r
  # estimate optimal factors in factor model
  #
  # Args:
  #   X.df: tbl_df that has the format [DATE, X], where
  #     $DATE: vector representing time (as.Date)
  #     X: matrix of regressor time series (T * N)
  #   ic: information criterion flag
  #
  # Returns:
  #   df.out: tbl_df of format [DATE, F], where
  #     $DATE: vector representing time (as.Date)
  #     F: matrix of factors (T * r)

  X <- X.df[, -1] %>% as.matrix
  X <- scale(X)
  n.obs <- dim(X)[1]
  n.var <- dim(X)[2]

  r.max <- floor(10 * log10(n.var))
  ics <- rep(NA, r.max)

  for (r in 1:r.max) {
    est.r <- GetFactors(X, r)
    F.r <- est.r$factors
    L.r <- est.r$loadings

    if (ic == 1) {
      penalty <- r * (n.var + n.obs) / (n.var * n.obs) * log((n.var * n.obs) / n.var + n.obs)
    } else if (ic == 2) {
      penalty <- r * (n.var + n.obs) / (n.var * n.obs) * log(min(n.var, n.obs))
    } else if (ic == 3) {
      penalty <- r * log(min(n.var, n.obs)) / min(n.var, n.obs)
    } else {
      stop("Invalid information criterion argument")
    }

    V.r <- matrixcalc::matrix.trace(crossprod(X - tcrossprod(F.r, L.r))) / (n.obs * n.var)
    ics[r] <- log(V.r) + penalty
    if (trace) cat("r =", r, "\tV =", V.r, "\tIC =", ics[r], "\n")
  }

  r.opt <- which.min(ics)
  est.opt <- GetFactors(X, r.opt)
  factors.opt <- est.opt$factors
  colnames(factors.opt) <- paste("F", 1:r.opt, sep = "")

  df.out <- data.frame(X.df[, 1], factors.opt) %>% tbl_df
  return(df.out)
}
```

R-script A.5: Funktioner til estimering af faktorer i faktor-modellen

```r
FactorFc <- function(y, X.df, lags.df, h, ic = 1) {
  # forecasts y[t+h] using the factor model
  #
  # Args:
  #   y: response vector
  #   X.df: tbl_df that has the format [DATE, X], where
  #     $DATE: vector representing time (as.Date)
  #     X: matrix of regressor time series (T * N)
  #   lags.df: tbl_df with the format [DATE, lags], where
  #     $DATE: vector representing time (as.Date)
  #     lags: lagged values of y
  #   h: forecast horizon
  #   ic: index between 1 and 3 selecting information criterion
  #
  # Returns:
  #   list.out: list containing
  #     $forecast: forecasted value of y[t+h]
  #     $r: number of factors
  #     $p: number of lags

  n.obs <- X.df %>% dim %>% .[1]
  p.max <- lags.df %>% dim %>% .[2] - 1
  F.df <- EstFactors(X.df, ic)
  r <- dim(F.df)[2] - 1

  ones.df <- data.frame(X.df$DATE, 1) %>% tbl_df
  names(ones.df) <- c("DATE", "intercept")

  fit.resp <- y[(1+h):n.obs]
  AICs <- rep(NA, (p.max+1))

  for (p in 0:p.max) {
    lags.p <- lags.df[, 1:(1+p)]
    model.mat <- ones.df %>%
      left_join(F.df, by = "DATE") %>%
      left_join(lags.p, by = "DATE") %>%
      select(-DATE) %>% as.matrix
    fit.mat <- model.mat[1:(n.obs-h), ]

    beta <- solve(crossprod(fit.mat), crossprod(fit.mat, fit.resp))
    resids <- fit.resp - fit.mat %*% beta
    sigma2 <- mean(resids^2)
    AICs[p+1] <- log(sigma2) + (n.obs + 2 * p) / n.obs
  }

  p.opt <- which.min(AICs) - 1
  model.mat <- ones.df %>%
    left_join(F.df, by = "DATE") %>%
    left_join(lags.df[, 1:(1+p.opt)], by = "DATE") %>%
    select(-DATE) %>%
    as.matrix
  fit.mat <- model.mat[1:(n.obs-h), ]
  beta <- solve(crossprod(fit.mat), crossprod(fit.mat, fit.resp))

  fc.vars <- drop(model.mat %>% tail(n = 1))
```

```
  y.fc <- drop(fc.vars %*% beta)

  list.out <- list(
    "forecast" = y.fc,
    "r" = r,
    "p" = p.opt
  )
  return(list.out)
}
```

R-script A.6: Funktioner til forecasting med faktor-modellen

```
FactorNc <- function(y, X.df, lags.df, ic = 1) {
  # nowcasts y[t] with the factor model
  #
  # Args:
  #   y: response vector
  #   X.df: tbl_df that has the format [DATE, X], where
  #     $DATE: vector representing time (as.Date)
  #     X: matrix of regressor time series (T * N)
  #   lags.df: tbl_df with the format [DATE, lags], where
  #     $DATE: vector representing time (as.Date)
  #     lags: lagged values of y corresponding to forecast type
  #   ic: index between 1 and 3 selecting information criterion
  #
  # Returns:
  #   list.out: list containing:
  #     $nowcast: estimated value for nowcast
  #     $r: number of factors
  #     $p: number of lags

  n.obs <- length(y)
  p.max <- lags.df %>% dim %>% .[2] - 1
  F.df <- EstFactors(X.df, ic)
  r <- dim(F.df)[2] - 1

  ones.df <- data.frame(X.df[, 1], 1) %>% tbl_df
  names(ones.df)[2] <- "intercept"

  fit.resp <- y
  AICs <- rep(NA, (p.max+1))

  for (p in 0:p.max) {
    lags.p <- lags.df[, 1:(1+p)]
    model.mat <- ones.df %>%
      left_join(F.df, by = "DATE") %>%
      left_join(lags.p, by = "DATE") %>%
      select(-DATE) %>% as.matrix
    fit.mat <- model.mat[1:n.obs, ]

    beta <- solve(crossprod(fit.mat), crossprod(fit.mat, fit.resp))
    resids <- fit.resp - fit.mat %*% beta
    sigma2 <- mean(resids^2)
    AICs[p+1] <- log(sigma2) + (n.obs + 2 * p) / n.obs
```

```
  }

  p.opt <- which.min(AICs) - 1
  model.mat <- ones.df %>%
    left_join(F.df, by = "DATE") %>%
    left_join(lags.df[, 1:(1+p.opt)], by = "DATE") %>%
    select(-DATE) %>%
    as.matrix
  fit.mat <- model.mat[1:n.obs, ]
  beta <- solve(crossprod(fit.mat), crossprod(fit.mat, fit.resp))

  nc.vars <- drop(model.mat %>% tail(n = 1))
  y.nc <- drop(nc.vars %*% beta)

  list.out <- list(
    "nowcast" = y.nc,
    "r" = r,
    "p" = p.opt
  )
  return(list.out)
}
```

R-script A.7: Funktioner til nowcasting med faktor-modellen

B

Korrelation mellem nowcast-variable i LASSO-modellen

	GDPC96	ISRATIOx	DPCERA...	INDPRO	CLAIMSx	RPI
GDPC96	1	-0.588	0.563	0.712	-0.563	0.525
ISRATIOx	-0.588	1	-0.463	-0.614	0.596	-0.242
DPCERA...	0.563	-0.463	1	0.388	-0.372	0.481
INDPRO	0.712	-0.614	0.388	1	-0.581	0.558
CLAIMSx	-0.563	0.596	-0.372	-0.581	1	-0.335
RPI	0.525	-0.242	0.481	0.558	-0.335	1
TB3SMFFM	0.301	-0.242	0.237	0.182	-0.294	0.164
UEMPLT5	-0.36	0.342	-0.231	-0.301	0.388	-0.195
S.P.500	0.293	-0.282	0.279	0.086	-0.271	0.212
IPFPNSS	0.708	-0.578	0.426	0.944	-0.536	0.563
AWOTMAN	0.454	-0.439	0.224	0.517	-0.495	0.28

	TB3SMFFM	UEMPLT5	S.P.500	IPFPNSS	AWOTMAN
GDPC96	0.301	-0.36	0.293	0.708	0.454
ISRATIOx	-0.242	0.342	-0.282	-0.578	-0.439
DPCERA...	0.237	-0.231	0.279	0.426	0.224
INDPRO	0.182	-0.301	0.086	0.944	0.517
CLAIMSx	-0.294	0.388	-0.271	-0.536	-0.495
RPI	0.164	-0.195	0.212	0.563	0.28
TB3SMFFM	1	-0.199	0.221	0.15	0.108
UEMPLT5	-0.199	1	-0.255	-0.311	-0.286
S.P.500	0.221	-0.255	1	0.08	0.084
IPFPNSS	0.15	-0.311	0.08	1	0.504
AWOTMAN	0.108	-0.286	0.084	0.504	1

Tabel B.1: Korrelationsmatrix for væksten i `GDPC96` og de 10 oftest udvalgte prædiktorer i LASSO-modellen

Litteratur

Bache, Stefan Milton, & Wickham, Hadley. 2014. *magrittr: A Forward-Pipe Operator for R.* R package version 1.5.

Bai, Jushan, & Ng, Serena. 2002. Determining the Number of Factors in Approimate Factor Models. *Econometrica*, **70**(1), 191–221.

Bengtsson, Henrik. 2016. *R.utils: Various Programming Utilities.* R package version 2.3.0.

Diebold, Francis X. 2013 (December). *Comparing Predictive Accuracy, Twenty Years Later.* `http://www.ssc.upenn.edu/~fdiebold/papers/paper113/Diebold_DM%20Test.pdf`.

European Commission. 2016 (February). *European Economic Forecast, Winter 2016.* `http://ec.europa.eu/economy_finance/publications/eeip/pdf/ip020_en.pdf`.

Friedman, Jerome, Hastie, Trevor, & Tibshirani, Robert. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**(1), 1–22. `http://www.jstatsoft.org/v33/i01/`.

Hastie, Trevor, Tibshirani, Robert, & Wainwright, Martin. 2015. *Statistical Learning with Sparsity.* CRC Press.

Hyndman, Rob J. 2015. *forecast: Forecasting functions for time series and linear models.* R package version 6.2.

Kasparis, Ioannis. 2008 (June). *A Simple Proof For The Invertibility Of The Lag Polynomial Operator.*

McCracken, Michael W. 2016 (March). *FRED-MD and FRED-QD: Monthly and Quarterly Databases for Macroeconomic Research.* `https://research.stlouisfed.org/econ/mccracken/fred-databases`. version '2016-03'.

McCracken, Michael W., & Ng, Serena. 2015. *FRED-MD: A Monthly Database for Macroeconomic Research.* `https://research.stlouisfed.org/wp/2015/2015-012.pdf`.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Shumway, Robert H., & Stoffer, David S. 2011. *Time Series Analysis and Its Applications.* 3 edn. Springer.

Stock, James, & Watson, Mark. 2002a. Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association*, **97**(460), 1167–1179.

Stock, James, & Watson, Mark. 2002b. Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business and Economic Statistics*, **20**(2), 147–162.

US. Bureau of Economic Analysis. 2016 (March). *Real Gross Domestic Product, 3 Decimal [GDPC96]*. `https://research.stlouisfed.org/fred2/series/GDPC96`. retrieved from FRED, Federal Reserve Bank of St. Louis.

Wickham, Hadley. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, Hadley. 2016. *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R package version 0.4.1.

Wickham, Hadley, & Francois, Romain. 2015. *dplyr: A Grammar of Data Manipulation*. R package version 0.4.3.