РАНХиГС
РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

ЭМИТ
ИНСТИТУТ ЭКОНОМИКИ, МАТЕМАТИКИ И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И
ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Методы снижения размерности в данных в макроэкономике

Отчёт по научно-исследовательской работе

2019

Михаил Гареев

ЭО-15-01

mkhlgrv@gmail.com

Научный руководитель: к.э.н. Полбин А.В.

Актуальность исследования

- При оценке моделей из макроэкономики часто можно столкнуться с тем, что параметров относительно много, а наблюдений - мало. Иногда эту проблему решается использованием методов снижения размерности в данных.

Цели и задачи

Цель:

- Проверка целесообразности использования методов снижения размерности в данных/

Задачи:

1. Обзор методов снижения размерности (LASSO, Post-LASSO, Ridge, Elastic Net, Random Forest, Spike-and-Slab variable selection).
2. Применение этих методов для оценки макроэкономических зависимостей в России (оценка безработицы), анализ результатов, сравнение с традиционными методами оценивания временных рядов.

Методы снижения размерности

Разреженная линейная модель с высокой размерностью в данных

Модель:

$$\beta_0 + \varepsilon_i, \epsilon_i \sim N(0, \sigma^2), \beta_0 \in \mathbb{R}^p, i = 1, \ldots, n,$$

где:

- y_i — это значения объясняемой переменной,
- x_i — это значения p-размерной объясняющей переменной,
- ε_i — значения независимых случайных ошибок в каждом наблюдении i,

при этом возможно, что $p \geq n$, но только $s < n$ компонентов вектораβ_0 не равны 0.

Можно ли уменьшить размерность модели?

Задача (Oracle Problem):

$$\min_{\beta \in \mathbb{R}^p} \mathbb{E}_n \left[(y_i - x_i{}' \beta)^2 \right] + \sigma^2 \frac{\|\beta\|_0}{n}, \tag{1}$$

где $\|\beta\|_0$ — это количество ненулевых компонентов в векторе β, обобщение понятия нормы для степени 0.

Гёльдерова норма для вектора x степени p:

$$\|x\|_p = \sqrt[p]{\sum_i |x_i|^p},$$

где обычно $p \geq 1$.

Решение (1) — это баланс между ошибкой регрессии и количеством ненулевых коэффициентов из вектора β.
Методы снижения размерности оптимизируют эмпирические аналоги задачи (1).

AIC/ BIC

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \left[(y_i - x_i{}' \beta)^2 \right] + \frac{\lambda}{n} \|\beta\|_0 ,$$

где λ — параметр штрафа.

LASSO

$$\hat{\beta}^{\text{LASSO}} \in \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \left[(y_i - x_i{}' \beta)^2 \right] + \frac{\lambda}{n} \|\beta\|_1 ,$$

где λ — параметр штрафа.

Ridge Regression

$$\hat{\beta}^{\text{Ridge}} \in \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \left[(y_i - x_i{}' \beta)^2 \right] + \frac{\lambda}{n} \|\beta\|_2 ,$$

где λ — параметр штрафа.

Elastic Net Regression

$$\hat{\beta}^{\text{EN}} \in \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \left[(y_i - x_i{}' \beta)^2 \right] + \frac{\lambda}{n} \left(\frac{1-\alpha}{2} \|\beta\|_1 + \alpha \|\beta\|_2 \right) ,$$

где λ — параметр штрафа, α — параметр регуляризации, равен 1 для Ridge и 2 для LASSO.

Post-LASSO

1. Использовать метода LASSO, найти $\hat{\beta}^{\text{LASSO}}$.
2. Применить МНК-регрессию, оценивая только неисключенные элементы $\hat{\beta}^{\text{LASSO}}$:

$$\hat{\beta}^{\text{Post-LASSO}} \in \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \left[(y_i - x_i{}' \beta)^2 \right], \text{ где } \beta_j = 0, \text{ если } \hat{\beta}_j = 0.$$

Random Forest

Двухэтапное получение оценок:

1. На разных подвыборках данных строится множество решающих деревьев,
2. в качестве предсказанного значения $\hat{y}_i$ выбираются усреднённые значения показаний по всем деревьям.

Регрессия пик-плато (Spike-and-slab)

- $\beta_j|\tau_j, r_j^2 \sim N(0, \tau_j \cdot r_j^2)$
- $$\tau_j = \begin{cases} 0 & \text{se } \omega \in A \\ 1 & \text{se } \omega \in A^c \end{cases}$$
- $r_j^2 \sim \text{Exp}(\lambda)$

Прогнозирование безработицы

Описание данных

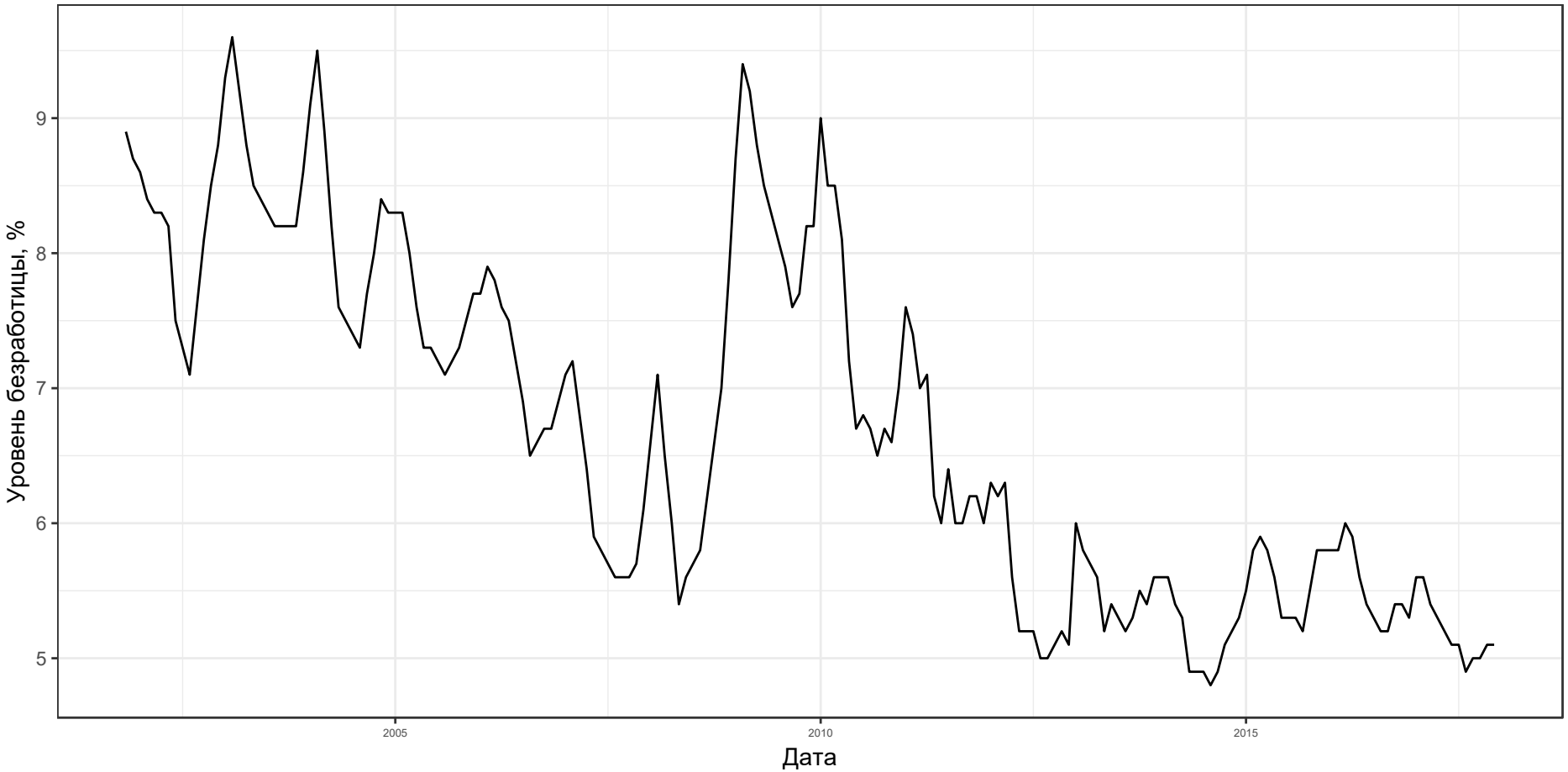
1. Прогнозируемая переменная: уровень безработицы в России (ноябрь 2001 – декарь 2017),
2. Объясняющие переменные: 83 ряда данных, отражающие различные макроэкономические показатели в России, уровень деловой активности и др. (январь 2001 – декарь 2016).

Обучение моделей ведется на десятилетнем движущемся окне, проверка качества моделей ведется на однолетнем окне для изменения безработицы в период от 1 до 24 месяцев.
Все ряды были очищены от сезонных и календарных эффектов и приведены к стационарному виду.

Прогнозирование безработицы

Описание данных

Безработица в России



Прогнозирование безработицы

Метод главных компонент

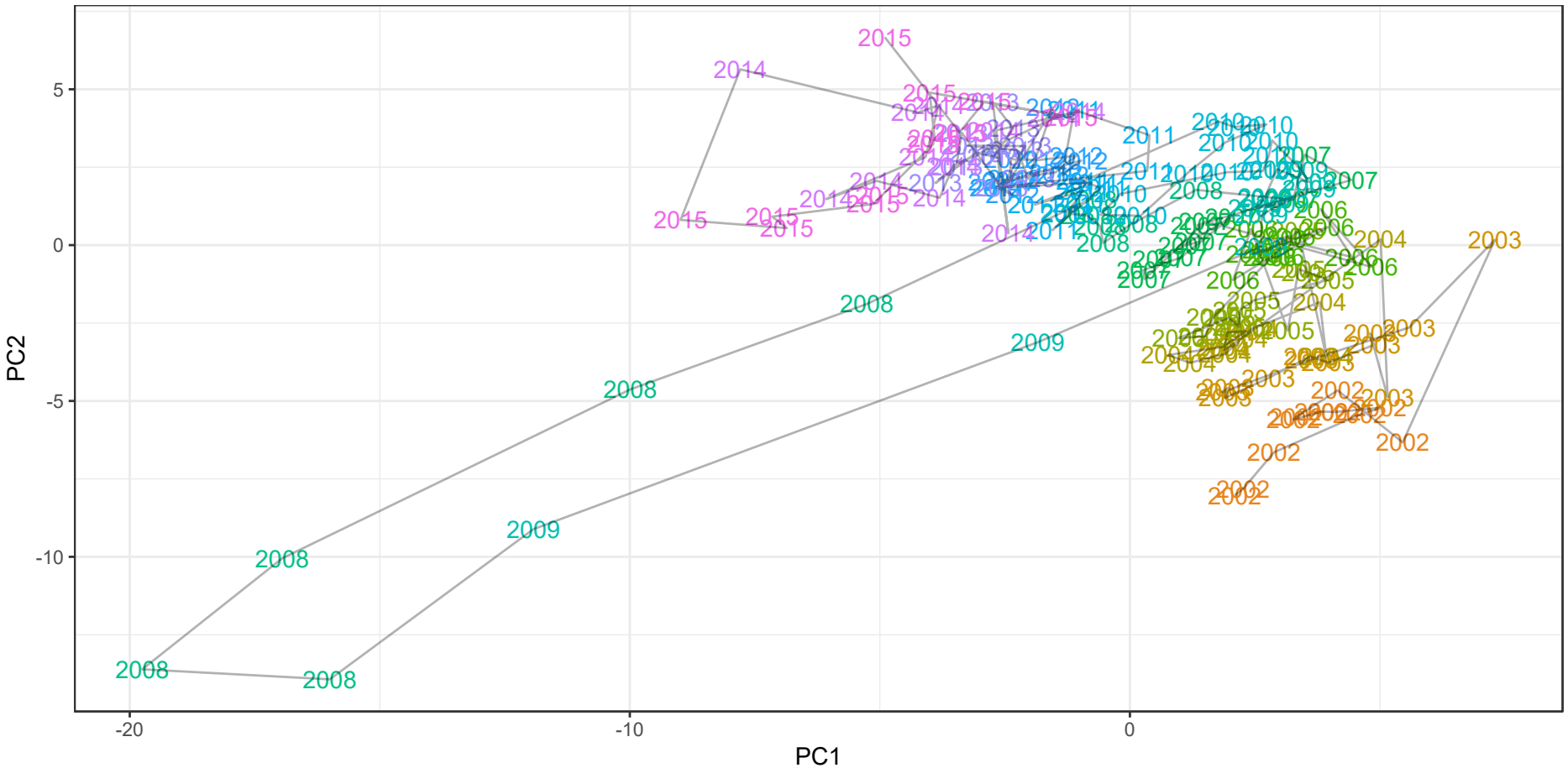
PCA: последовательная минимизация суммы квадратов отклонений старых значений от новых или замена матрицы $X_{n\times k}$ на матрицу $n \hat{\times} k$ ранга $p < k$, так, чтобы:

$$\min \sum_{j=1}^{k} \sum_{i=1}^{n} (x_{ij} - \hat{x_{ij}})^2$$

Прогнозирование безработицы

Метод главных компонент

Экономика России в двумерном пространстве



Прогнозирование безработицы

Базовый бенчмарк

Модель ARMA(p,q)

Для сравнения качества используется модель ARMA(p,q), где p и q выбираются при помощи AIC.

Прогнозирование безработицы

Метрика качества моделей

RMSE

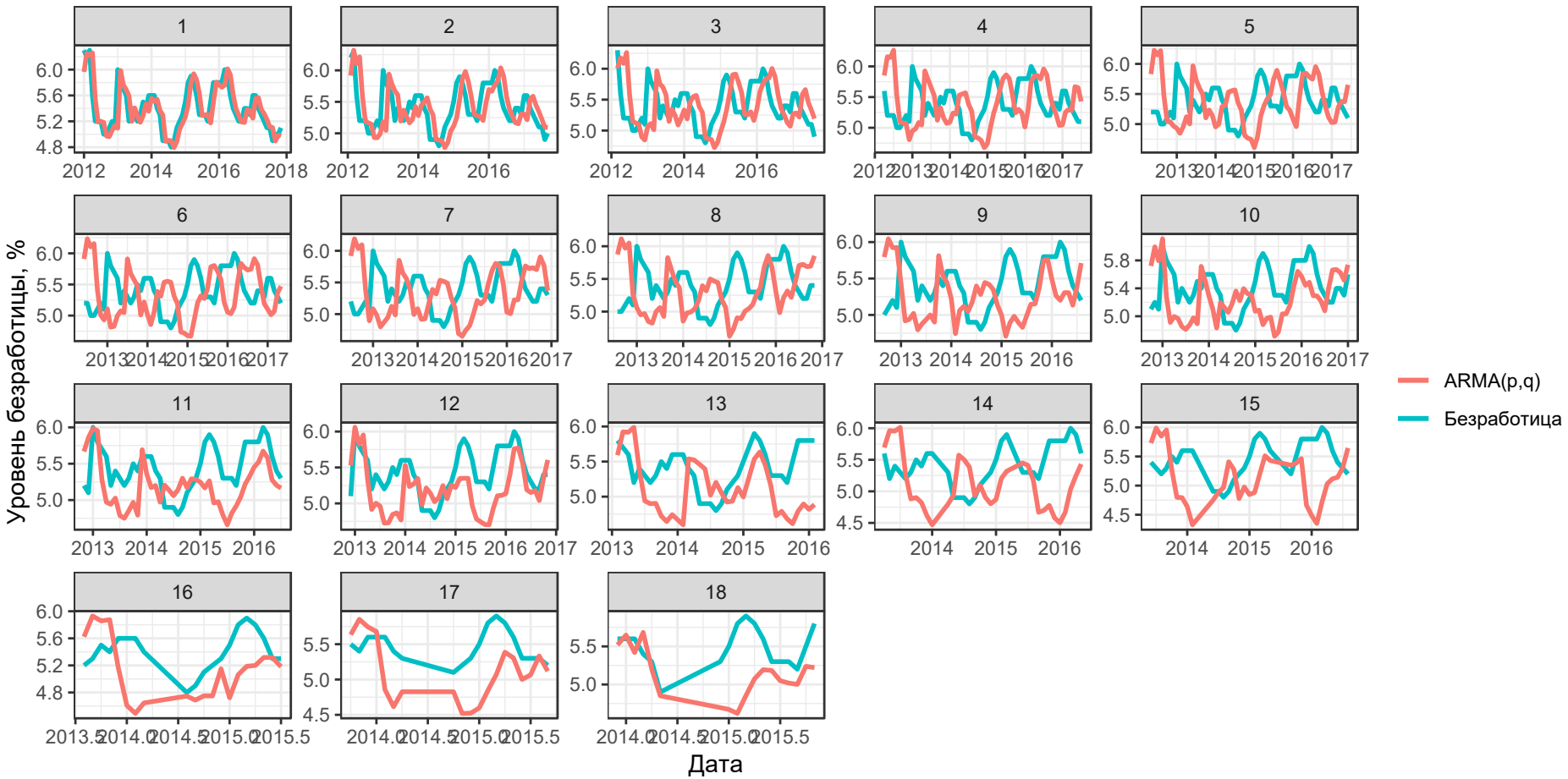
Для сравнения качества используется метрика RMSE (Root-mean-square error):

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)}{T}}$$

Прогнозирование безработицы

Результаты

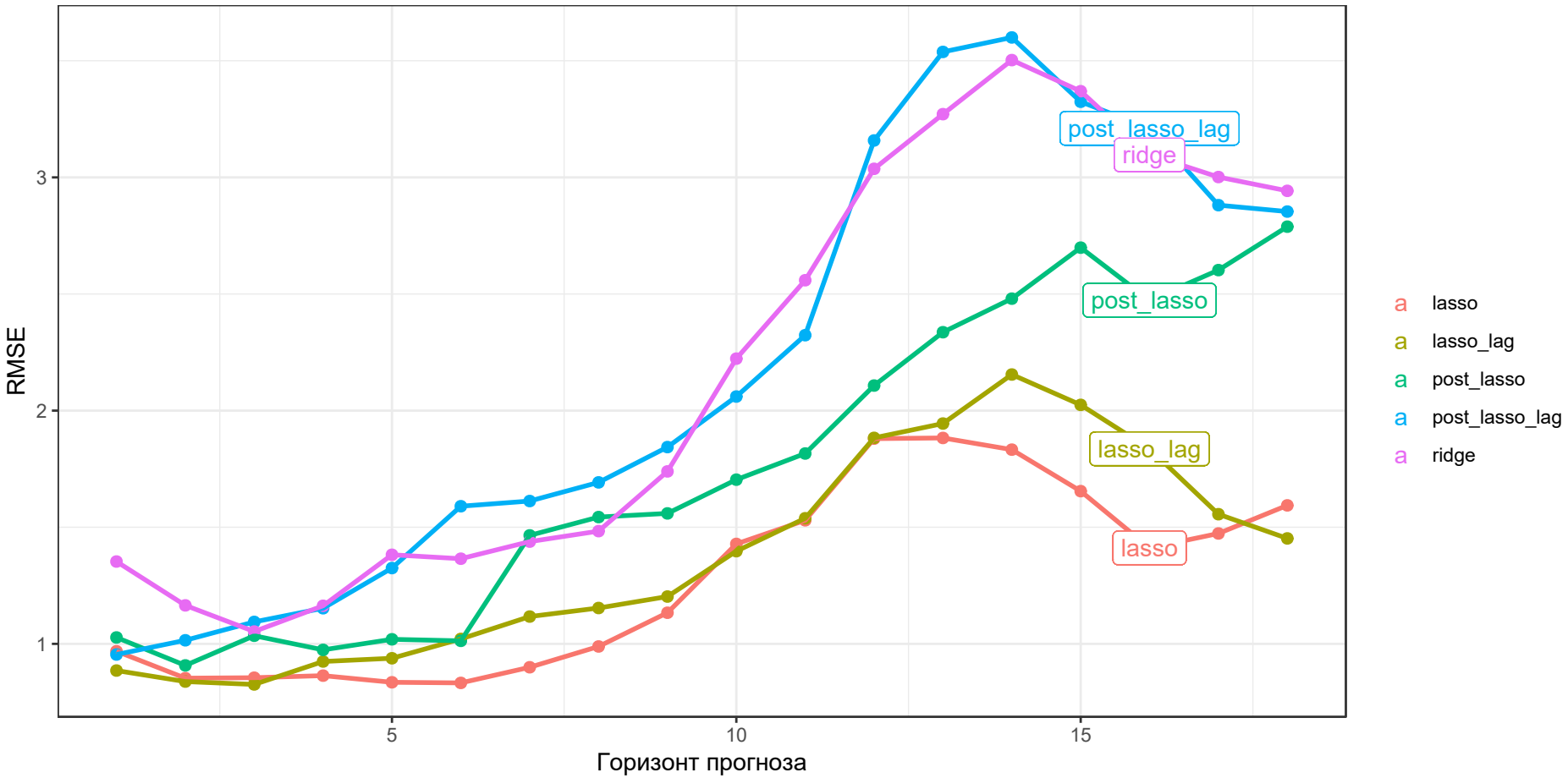
Базовый прогноз



Прогнозирование безработицы

Результаты

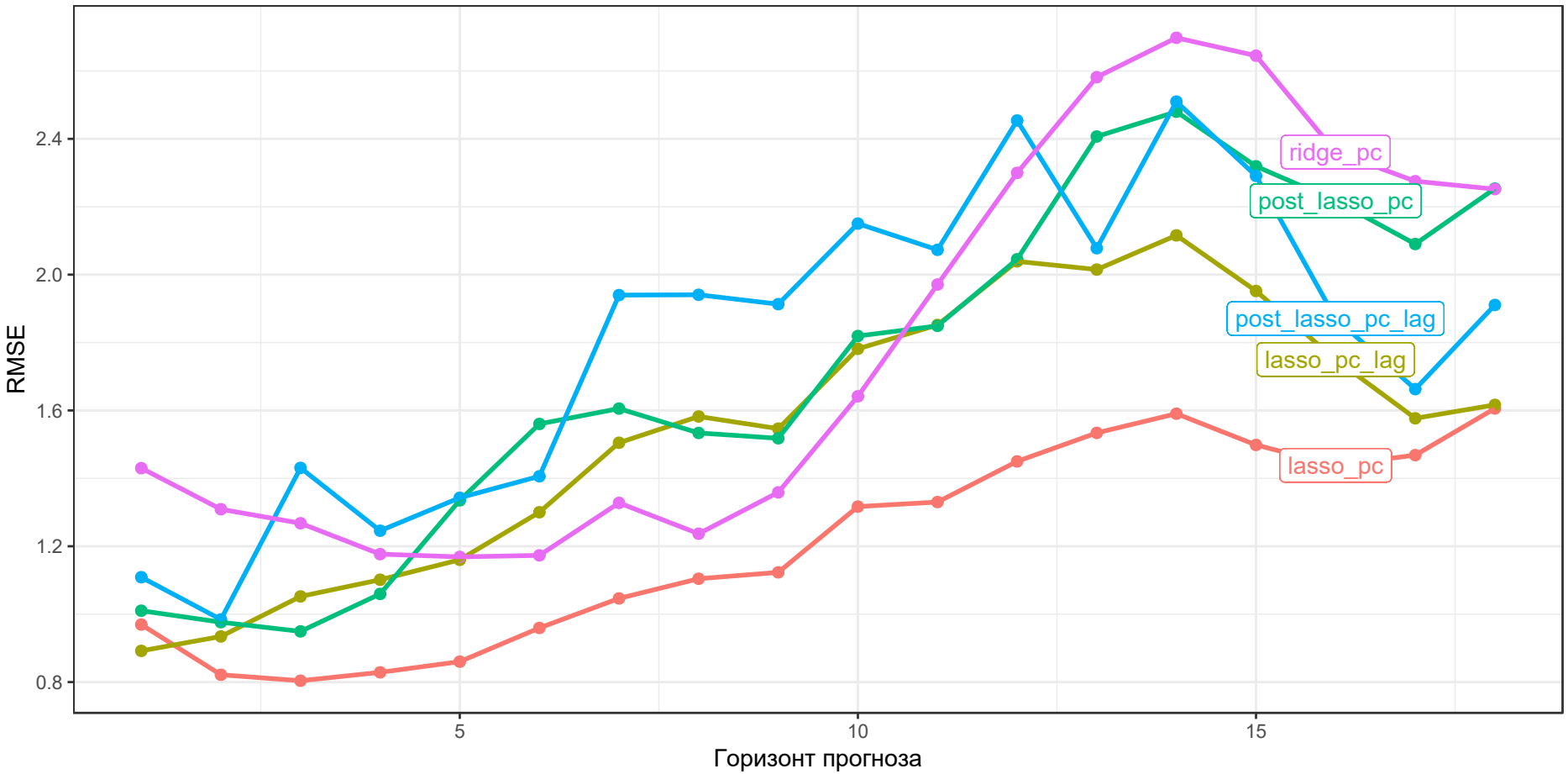
Модели с регуляризацией



Прогнозирование безработицы

Результаты

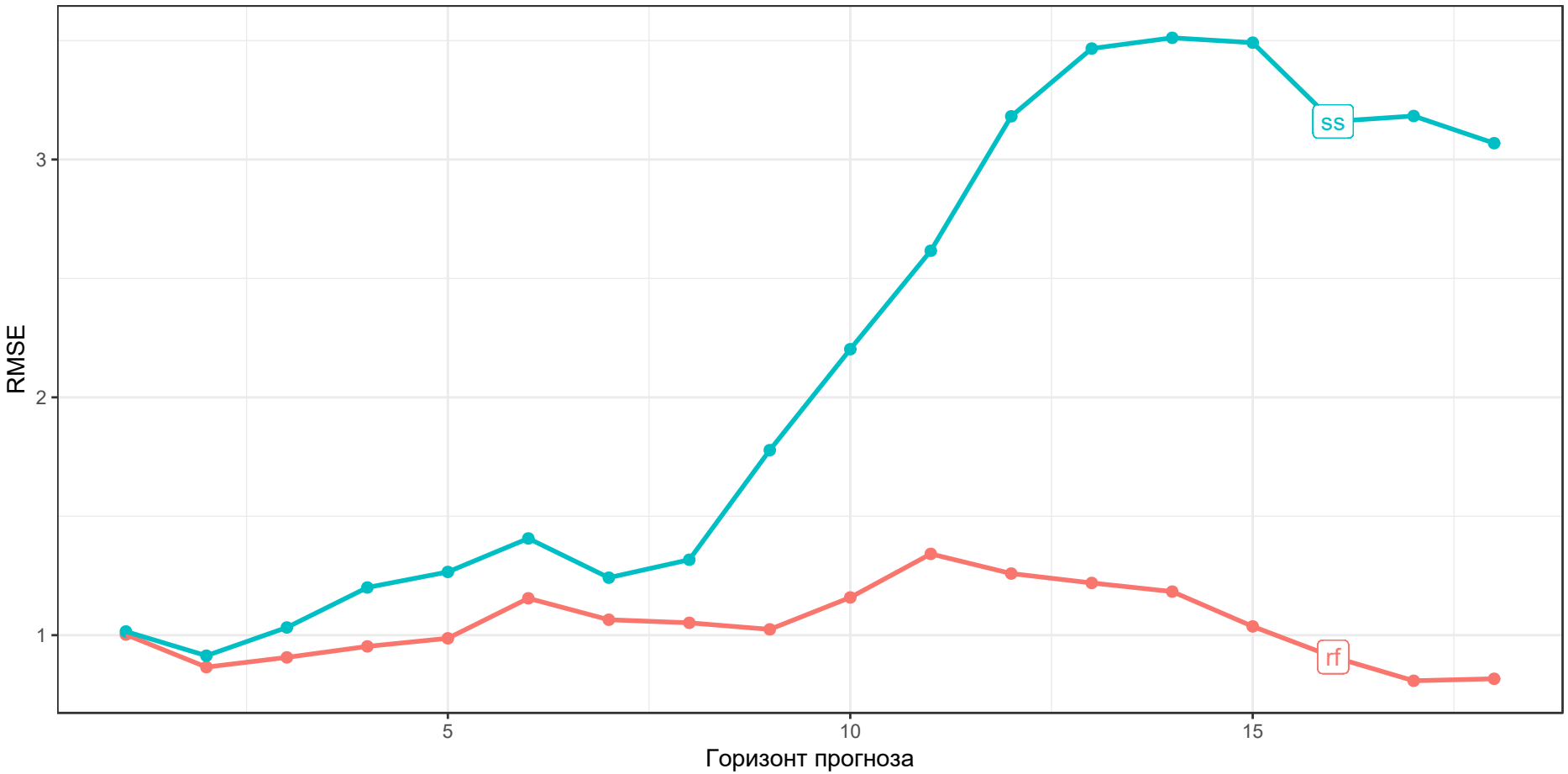
Модели с регуляризацией и трансформацией данных через главные компоненты



Прогнозирование безработицы

Результаты

Остальные модели



Прогнозирование безработицы

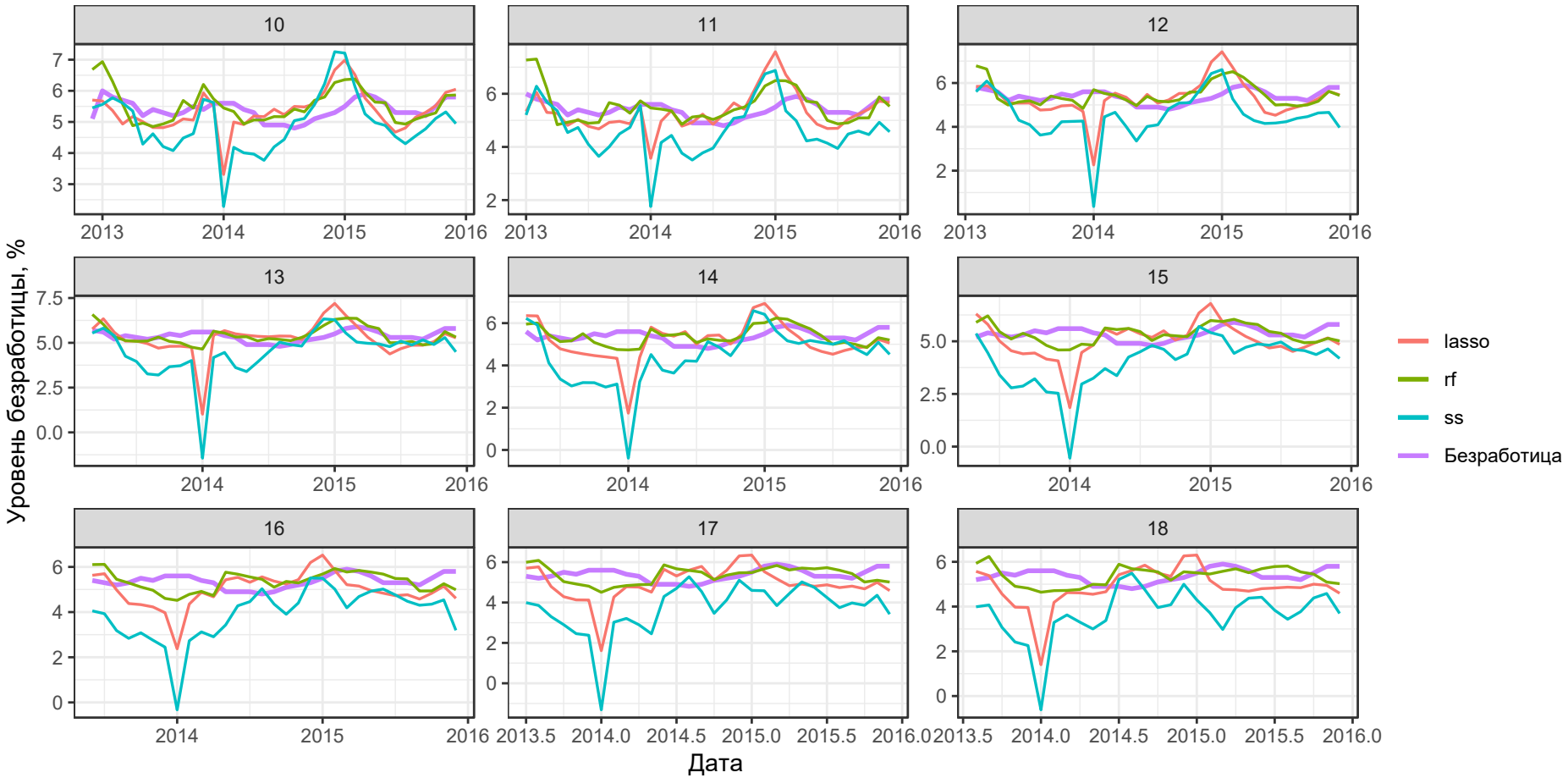
Результаты

Сравнения предсказаний для некоторых моделей (1–9 мес)

Прогнозирование безработицы

Результаты

Сравнения предсказаний для некоторых моделей (10–18 мес)



Модель	1	2	3	4	5	6
Elastic Net	0.97		0.85			
LASSO	0.97	0.85	0.85	0.86	0.84	0.83
LASSO with lag	0.89	0.84	0.83	0.92	0.94	1.02
LASSO with PC	0.97	0.82	0.80	0.83	0.86	0.96
LASSO with PC and lag	0.89	0.93	1.05	1.10	1.16	1.30
Post-LASSO	1.03	0.91	1.03	0.97	1.02	1.01
Post-LASSO with lag	0.95	1.02	1.09	1.15	1.32	1.59
Post-LASSO with PC	1.01	0.98	0.95	1.06	1.34	1.56
Post-LASSO with PC and lag	1.11	0.98	1.43	1.25	1.34	1.41
Random Forest	1.00	0.87	0.91	0.95	0.99	1.16
ridge	1.35	1.16	1.05	1.16	1.38	1.37
ridge_pc	1.43	1.31	1.27	1.18	1.17	1.17
Spike-and-Slab 1.02	0.91	1.03	1.20	1.27	1.41	

Модель	7	8	9	10	11	12
Elastic Net						1.88
LASSO	0.90	0.99	1.13	1.43	1.53	1.88
LASSO with lag	1.12	1.15	1.20	1.40	1.54	1.88
LASSO with PC	1.05	1.10	1.12	1.32	1.33	1.45
LASSO with PC and lag	1.50	1.58	1.55	1.78	1.85	2.04
Post-LASSO	1.47	1.54	1.56	1.70	1.82	2.11
Post-LASSO with lag	1.61	1.69	1.84	2.06	2.32	3.16
Post-LASSO with PC	1.61	1.53	1.52	1.82	1.85	2.05
Post-LASSO with PC and lag	1.94	1.94	1.91	2.15	2.07	2.45
Random Forest	1.06	1.05	1.02	1.16	1.34	1.26
ridge	1.44	1.48	1.74	2.22	2.56	3.04
ridge_pc	1.33	1.24	1.36	1.64	1.97	2.30
Spike-and-Slab 1.24	1.32	1.78	2.20	2.62	3.18	

Модель	13	14	15	16	17	18
Elastic Net						
LASSO	1.88	1.83	1.65	1.41	1.47	1.59
LASSO with lag	1.94	2.15	2.02	1.84	1.56	1.45
LASSO with PC	1.53	1.59	1.50	1.44	1.47	1.61
LASSO with PC and lag	2.02	2.12	1.95	1.75	1.58	1.62
Post-LASSO	2.34	2.48	2.70	2.47	2.60	2.79
Post-LASSO with lag	3.54	3.60	3.32	3.21	2.88	2.85
Post-LASSO with PC	2.41	2.48	2.32	2.22	2.09	2.25
Post-LASSO with PC and lag	2.08	2.51	2.29	1.87	1.66	1.91
Random Forest	1.22	1.18	1.04	0.91	0.81	0.82
ridge	3.27	3.50	3.37	3.10	3.00	2.94
ridge_pc	2.58	2.70	2.64	2.36	2.28	2.25
Spike-and-Slab 3.47	3.51	3.49	3.16	3.18	3.07	

Краткий вывод и планы

- Методы снижения размерности (LASSO, Post-LASSO, Ridge, Elastic Net, Random Forest) потенциально представляют собой мощный инструмент для нахождения и проверки макроэкономических зависимостей.
- Из использованных методов лучшие результаты при прогнозировании инфляции в России показывают модели LASSO и Random Forest. На разных горизонтах планирования (кроме диапазона с 9 до 15 месяцев) хотя бы одна из них показывала лучшие результаты, чем модель-бенчмарк (ARMA).

Спасибо за внимание

Методы снижения размерности в данных в макроэкономике

Михаил Гареев

ЭО-15-01

mkhlgrv@gmail.com

2019

Источники

- Belloni, Alexandre and Chernozhukov, Victor
  High dimensional sparse econometric Модельs: An introduction.
  Springer, 2011
- Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen.
  Lasso methods for gaussian instrumental variables Модельs
  2011
- Barro, Robert J. and Lee, Jong-Wha
  Data Set for a Panel of 138 Countries
  1994
- Candes, Emmanuel, and Terence Tao.
  The Dantzig selector: Statistical estimation when p is much larger than n.
  The Annals of Statistics 35.6 (2007): 2313-2351.
- Akaike, Hirotugu.
  A new look at the statistical Модель identification.
  IEEE transactions on automatic control 19.6 (1974): 716-723.
- Единый архив экономических и социологических данных, статистические ряды
  http://sophist.hse.ru/hse/nindex.shtml