HERIOT WATT UNIVERSITY

Structural Economic Models with Large Number of Potential Explanatory Variables and Inclusion/Exclusion Restrictions

Arnab Bhattacharjee, SEEC, Heriot-Watt Univ, UK

Conference on

Big Data, Machine Learning and the Macroeconomy

Norges Bank, October 2017

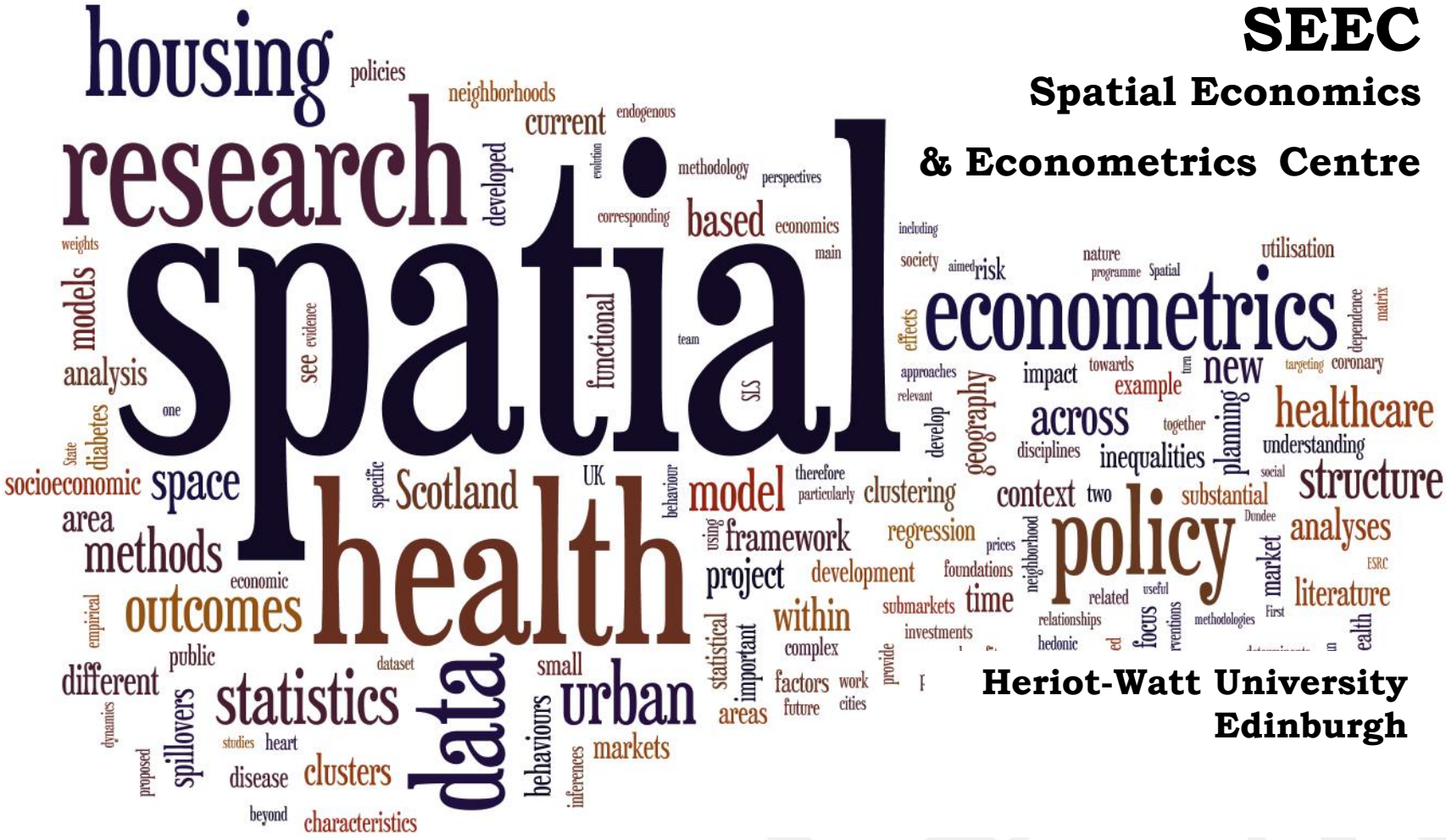
Collaborators: Maiti (Michigan State Univ) and Hewings (Univ Illinois)

LEADERS IN IDEAS AND SOLUTIONS

HERIOT WATT UNIVERSITY

SEEC

Spatial Economics & Econometrics Centre



Heriot-Watt University
Edinburgh

LEADERS IN IDEAS AND SOLUTIONS

Background

- Current generation of macroeconomists have at their disposal data on hundreds (sometimes thousands) of macro time series variables
- How do you make effective and efficient use of such BIG data to build macro models?
  - Good prediction, for example, Vector Autoregression (VARs)
  - Effective counterfactual analysis and policy – structural VARs
- The most popular methods/ models are various types of VARs endowed with (latent) factor structure
  - Factor-Augmented VAR (FAVAR) (Bernanke, Boivin & Eliasz, 2005); Infinite dimensional VARs (Chudik & Pesaran, 2009); Infinite dimensional VARs with dominant units (Pesaran & Chudik, 2010)

HERIOT WATT UNIVERSITY

Challenges

- Traditional approaches not very satisfactory. Why?
  - VARs offer excellent prediction, but they are atheoretic. Without structural modeling of contemporaneous dependence, they are relatively useless for policy
  - Structural macro models are the opposite – good for policy, but poor predictions or fit to data; for example Dynamic Stochastic General Equilibrium (DSGE) models
  - Structural VARs strike a useful balance; for example, DSGE-VARs (Del-Negro & Schorfheide, 2004; Fernández-Villaverde, Rubio-Ramírez, Sargent & Watson, 2007)
  - But do not use fully data available in BIG data contexts

LEADERS IN IDEAS AND SOLUTIONS

What do we do?

- VARs with factor structure try to bridge the gap
  - Large number of variables aggregated into latent factors
  - Combined with other variables into VARs or structural VARs
  - But statistical factor analysis loosens link with theory, often breaking the structural links that bind theoretical models together; see, for example, Bhattacharjee & Christev (2016)
  - Double ML? (Chernozhukov et al., 2016)
- What we do in this paper
  - Take a very specific application context
  - Place a simple structural model at the base – in our case just a simple VECM based consumption function
  - Use current generation model selection methods to add variables from a large collection – IIS/OCMT and LASSO
  - Evaluate forecast performance, but importantly structural interpretations of augmented models

Model selection methods

- Impulse Indicator Saturation (IIS) & friends (OCMT)
  - Hendry & Santos (2005); Hendry, Johansen & Santos (2008); Castle, Doornik & Hendry (2012)
  - An extension of general-to-specific modelling
  - Detects and models location shifts, traditionally over time
  - A variety of shifts using a 'split-half' analysis, the simplest specialization of a multiple-block search algorithm
  - We turn the problem round to use for model selection – similar to OCMT (Chudik, Kapetanios & Pesaran, 2016)
- LASSO & friends (Tibshirani, 1996; Varian, 2014)
  - No introduction necessary for this audience
  - We use a specific implementation that allows for cross-section dependence (Cai, Bhattacharjee, Calantone & Maiti, 2016)

Application context

- Basic idea:
  - Endow VARs with economic structure, using structural VARs (SVARs) or vector error correction models (VECMs)
  - Plus, selection of additional variables restricted by inclusion and exclusion constraints in such a way that they offer structural interpretation
  - Can IIS/OCMT or LASSO then be useful for (a) improved prediction , together with (b) better structural interpretation & policy?
- In our application:
  - In the model for consumption in Illinois, wages for the counties in Illinois can be additional regressors, but not counties in Michigan, and also data only at far enough lags that make the model suitable for prediction.

Application context

- Our implementation:
  - We implement this idea based on monthly data on consumption and disposable income for the core states of the US MidWest (Illinois, Indiana, Iowa, Michigan and Wisconsin) – Source: Chicago Fed
  - Together with a collection of a large number of potential covariates – Source: FRED database, but imputed to monthly frequency
  - At the base is an error correction model (ECM) with one potential cointegrating relationship between consumption, income and prices, as well as short run dynamics.
  - Model follows Pesaran, Shin & Smith (1999) (i: state, t: time)

$$\Delta \ln c_{it} = \alpha_i + \sum_p \beta_{1pi} \Delta \ln y_{it} + \sum_q \beta_{2qi} \Delta \pi_{it} - \gamma_i (\ln c_{i,t-1} - \delta_{1i} \ln y_{i,t-1} - \delta_{2i} \pi_{i,t-1}) + \varepsilon_{it}$$

Implementation (contd.)

- Our implementation:
  - This above model is structural in the sense that the effect of permanent and transitory income shocks on consumption are clearly emphasized.
  - The model is then augmented with a large number of potential additional covariates, with inclusion-exclusion restrictions, chosen by model selection.
  - 5 different regression models, one each with dependent variables il_dlc, in_dlc, ia_dlc, mi_dlc and wi_dlc. Each model has a different set of regressors, including:
    - regressors that are guaranteed inclusion;
    - regressors that need to be selected by IIS/ LASSO; and
    - regressors that must be excluded.

HERIOT WATT UNIVERSITY

Data

- Observations and variables:
  - Total of 483 monthly observations for each of the 5 states
  - We fit the model using the first 470 (potentially reduced to 457 using lags), Feb 1976 to Mar 2015
  - Remaining 13 (Apr 2015 to Apr 2016) are retained for evaluation of out-of-sample predictions.
  - Two prediction exercises:
    - First, evaluate one-step-ahead forecasts only. Estimate model using data for Feb 1976 to Mar 2015, and obtain forecasts for Apr 2015. Then use real data for Feb 1976 to Apr 2015 and obtain forecasts for May 2015. And so on.
    - Second, 'dynamic forecasts' are obtained for the final 13 observations, using one-step-ahead forecasts to obtain forecasts for the following period.

Data (contd.)

- BIG data context:
  - Model for Illinois
    - Regressors that are guaranteed inclusion (9 variables)
    - Regressors that need to be selected by IIS/ LASSO (4173 variables)
    - Regressors that must be excluded (430 variables)
  - Model for Indiana
    - Regressors: 9, 4056 and 439 variables, respectively
  - Model for Iowa
    - Regressors: 9, 4082 and 437 variables, respectively
  - Model for Michigan
    - Regressors: 9, 3952 and 447 variables, respectively
  - Model for Wisconsin
    - Regressors: 9, 3796 and 459 variables, respectively

Results (preliminary)

- The model selection exercise using Impulse Indicator Saturation (IIS) reflects an overall surprising finding
- **It is very hard to beat a simple error correction model for DLC with common correlated effects and lag selection (ECM)**
- Main problem is that the IIS selects many variables which are irrelevant for forecasting (overfitting), and hence forecast error is large.
- Then, we apply a multiple testing correction using a p-value (penalty or smoothing parameter) chosen by forecast performance – note the connection to OCMT
- Likewise, for LASSO, we use cross validation.

Results (contd.)

- Thus, we estimate the model using data upto March 2015, then obtain predictions/ forecasts for the 13 months April 2015 to April 2016. For IIS, we use a multiple-testing-corrected modified p-value that provides the minimum RMSE for the forecast period.
- Having done this, we find that the IIS beats the ECM in out-of-sample forecast performance, but only just.
  - RMSE for the IIS is 0.001784, which is 99.38% of the ECM's RMSE at 0.001795.
  - We also compared performance against the Pesaran panel ECM; this has a slightly worse performance with RMSE of 0.001868.

HERIOT WATT UNIVERSITY

Results (contd.)

- In sample, IIS beats the ECM and panel ECM handsomely
  - The RMSEs are 0.004055 (ECM), 0.004779 (Panel ECM) and 0.003098 (IIS).
  - The RMSEs here are larger, because the estimation sample includes periods of very high volatility.
- Results with LASSO are mixed
  - In sample, lower RMSEs than IIS by 5 to 10 percent
  - Out of sample, mixed results – overfitting?
  - This is a serious issue in dependent data settings; see Nandy, Lim & Maiti (2017) and Cai et al. (2016)
  - Work in progress: Careful choice of penalty required – new Stata program (Ahrens, Hansen & Schaffer 2017)

LEADERS IN IDEAS AND SOLUTIONS

Flavour of estimated models

- Illinois: $\Delta \ln c_{it} = -0.040 + 0.133\, \Delta \ln y_{it} + (0)\, \Delta \pi_{it} - 0.015\, (\ln c_{i,t-1} - 1.419\, \ln y_{i,t-1} + 1.212\, \pi_{i,t-1})$
  - Factor: avg US consumption growth (strong dependence)
  - VAR: lags of $\Delta \ln c$ (IL:1,3,4,6; IN:1,4; IA:1; MI:1,2,6; WI:1)
  - Spatial Durbin – popn wtd spat lag of MidWest State unemp
  - State popn growth, 3m lag (Missouri, Nebraska)
  - State income growth, 3m lag (Hawaii, Washington)
  - IL County income growth, 3m lag
    (Washington, McHenry, Adams, Rock Island, Cumberland)
- Wisconsin: $\Delta \ln c_{it} = (0) + 0.488\, \Delta \ln y_{it} + (0)\, \Delta \pi_{it}$
  - No evidence of cointegration or factors/strong dependence
  - VAR: lags of $\Delta \ln c$ (WI:1,3; IN:1; IA:1; MI:6)

Flavour of estimated models

- Indiana: $\Delta \ln c_{it} = -0.067 + (0)\ \Delta \ln y_{it} + (0)\ \Delta \pi_{it} - 0.029\ (\ln c_{i,t-1} - 1.374\ \ln y_{i,t-1} + 0.762\ \pi_{i,t-1})$
  - Factor: average US consumption growth (strong dependence)
  - VAR: lags of $\Delta \ln c$ (IN:3,6; IA:1; MI:1)
  - Spatial Durbin – popn wtd spatial lag of MidWest State unemp
  - State unemployment rate, 3m lag (New Jersey)
  - State income growth, 3m lag (Philadelphia)
  - IL City per capita personal income growth, 3m lag (Elkhart-Goshen)
- Iowa: $\Delta \ln c_{it} = 0.010 + 0.123\ \Delta \ln y_{it} + (0)\ \Delta \pi_{it}$
  - No evidence of cointegration or factors/strong dependence
  - VAR: lags of $\Delta \ln c$ (IA:1,3,4;WI:3)
- Michigan: $\Delta \ln c_{it} = (0) + (0)\ \Delta \ln y_{it} + (0)\ \Delta \pi_{it}$
  - No evidence of short run dynamics, cointegration or factors
  - VAR: lags of $\Delta \ln c$ (MI:1,3,6;IN:1;IA:1,4)

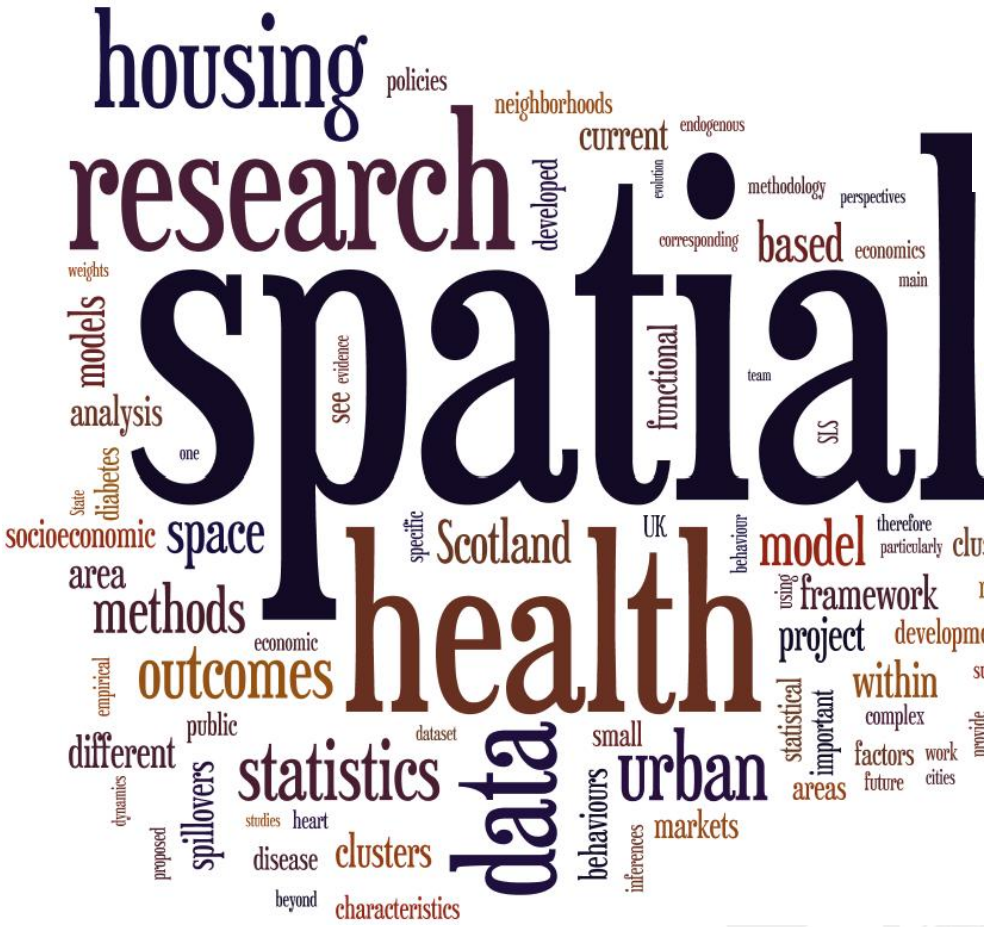
Further thoughts/ work

- The estimated model can be viewed as a structural spatial-Durbin model (LeSage & Pace, 2009).
- However, if there were contemporaneous spillovers between the states, then it can only be interpreted as a reduced form.
- What happens if we can allow for contemporaneous structural (spatial) linkages between the states; see Hewings & Parr (2007) & Chung & Hewings (2015).
- This would require causal identification assumptions such as DAGs.

Further thoughts/ work

- Work in progress:
  - Our reduced form provides a way to infer on the recursive causal structure; see Basak, Bhattacharjee & Das (2017).
  - Potentially, this can be taken to more elaborate causal graphical models as in Chung and Hewings (2015).
  - Once this is done, one can potentially build an SVAR model along the above lines.
- Summary
  - Developing structural VAR/VECMs in BIG data situations is not straightforward.
  - Using well specified structural economic model as the base, improved predictions and structural interpretation can be achieved
  - Then, model selection methods IIS/LASSO can be useful.

Thank you very much!!!



SEEC

Spatial Economics & Econometrics Centre

Heriot-Watt University
Edinburgh

LEADERS IN IDEAS AND SOLUTIONS

Exciting work continues ...

... Will be great to have you on board

Table 1: Estimates of High Dimensional Vector Error Correction Model (Illinois)

Dep.var.: $\Delta \ln C_t$ (p-values in parentheses)	**Panel VECM**	**HD VECM (Lasso)**	**HD VECM (IIS)**
Short run dynamics	$0.1261 \Delta \ln Y_t$ (0.000) $+0.0131 \Delta \pi_t$ (0.412)	$0.0601 \Delta \ln Y_t$ (0.127) $-0.0016 \Delta \pi_t$ (0.932)	$0.1328 \Delta \ln Y_t$ (0.000) $+0.0008 \Delta \pi_t$ (0.961)
ECM Partial adjustment	-0.0005 (0.683)	-0.0007 (0.664)	$(\ln C - 1.419 \ln Y + 1.292 \pi)$ -0.0148 (0.000)
Factor – US consumption	$+0.0037$ (0.000)	$+0.0039$ (0.000)	$+0.0051$ (0.000)
State-level consumption (monthly lags)	Illinois (1, 3, 4, 6, 7, 8, 9, 11, 12); Indiana (1, 2, 4, 5, 6, 7); Iowa (1, 2, 4, 5, 6, 7); Michigan (1, 2, 4, 5); Wisconsin (1, 4, 5)	Illinois (1, 3, 4); Indiana (—); Iowa (1); Michigan (—); Wisconsin (—)	Illinois (1, 3, 4, 6); Indiana (1, 4); Iowa (1); Michigan (1, 2, 6); Wisconsin (1)
State-level income growth, 3 month lag	—	$MD: +0.0998$ (0.002)	$HI: +0.0771$ (0.001) $WA: +0.1284$ (0.000)
State-level population growth, 3 month lag	—	—	$MO: +2.2620$ (0.000) $NE: -1.6200$ (0.001)
Popn.wtd. spatial lag: MidWest state unemp.rate	—	—	$+0.0007\, Wu_t$ (0.000)
IL county-level income growth, 3 month lag	—	Clinton: $+0.1147$ (0.007)	Washington: $+0.1049$ (0.000); McHenry: $+0.2393$ (0.000); Adams: $+0.2314$ (0.002); Rock Island: -0.3949 (0.000); Cumberland: -0.0740 (0.003)
City per capita personal income growth, 3 month lag	—	Davenport-Molin-Rock Island (IL-IA): -0.0170 (0.003)	—
Intercept	0.0063 (0.017)	0.0055 (0.096)	-0.0395 (0.000)
In-sample $\overline{R}^2$, RMSE (1976m2 – 2015m3)	**0.6596, 0.00160**	0.4748, 0.00199	0.6334, 0.00166
Out-of-sample RMSE (2015m4 – 2016m4)	0.001054	0.001222	**0.001028**

Table 2: Estimates of High Dimensional Vector Error Correction Model (Indiana)

Dep.var.: $\Delta \ln C_t$ (p-values in parentheses)	**Panel VECM**	**HD VECM (Lasso)**	**HD VECM (IIS)**
Short run dynamics	$0.0514\,\Delta \ln Y_t$ (0.090) $+0.0039\,\Delta \pi_t$ (0.870)	$-0.0260\,\Delta \ln Y_t$ (0.386) $-0.0113\,\Delta \pi_t$ (0.660)	$-0.0505\,\Delta \ln Y_t$ (0.090) $-0.0074\,\Delta \pi_t$ (0.770)
ECM Partial adjustment	-0.0002 (0.924)	-0.0012 (0.511)	$(\ln C - 1.374 \ln Y + 0.762\pi)$ -0.0292 (0.000)
Factor – US consumption	$+0.0053$ (0.000)	$+0.0084$ (0.000)	$+0.0102$ (0.000)
State-level consumption (monthly lags)	Indiana (1 – 13); Illinois (1, 2, 4, 5, 7); Iowa (1, 2, 4, 5, 6, 7); Michigan (1 – 6); Wisconsin (1, 5, 7)	Indiana (3); Illinois (—); Iowa (1); Michigan (1, 6); Wisconsin (—)	Indiana (3, 6); Illinois (—); Iowa (1); Michigan (1); Wisconsin (—)
State-level income growth, 3 month lag	—	$DE: +0.1519$ (0.000)	$PA: +0.2579$ (0.000)
State-level population growth, 3 month lag	—	$MN: +0.5010$ (0.059)	—
State-level unemployment rate, 3 month lag	—	$DC: +0.0003$ (0.002)	$NJ: -0.0005$ (0.001)
Popn.wtd. spatial lag: MidWest state unemp.rate	—	—	$+0.0021\,Wu_t$ (0.000)
IN county-level income growth, 3 month lag	—	Elkhart: $+0.2155$ (0.001)	—
City per capita personal income growth, 3 month lag	—	—	Elkhart-Goshen (IN): $+0.0289$ (0.000)
Intercept	0.0079 (0.009)	-0.0045 (0.292)	-0.0678 (0.000)
In-sample $\overline{R}^2$, RMSE (1976m2 – 2015m3)	**0.6180, 0.00245**	0.4840, 0.00284	0.5033, 0.00279
Out-of-sample RMSE (2015m4 – 2016m4)	**0.000806**	0.001622	0.001020

Table 3: Estimates of High Dimensional Vector Error Correction Model (Iowa)

Dep.var.: $\Delta \ln C_t$ (p-values in parentheses)	**Panel VECM**	**HD VECM (Lasso)**	**HD VECM (IIS)**
Short run dynamics	$\underset{(0.007)}{0.0835}\Delta \ln Y_t$ $+\underset{(0.817)}{0.0065}\Delta \pi_t$	$\underset{(0.002)}{0.0994}\Delta \ln Y_t$ $-\underset{(0.199)}{0.0282}\Delta \pi_t$	$\underset{(0.000)}{0.1226}\Delta \ln Y_t$ $-\underset{(0.124)}{0.0380}\Delta \pi_t$
ECM Partial adjustment	$-\underset{(0.825)}{0.0004}$	$(\ln C - 0.699 \ln Y + 3.022\pi)$ $-\underset{(0.028)}{0.0042}$	$(\ln C - 0.461 \ln Y + 7.154\pi)$ $-\underset{(0.067)}{0.0034}$
Factor – US consumption	$+\underset{(0.013)}{0.0024}$	$+\underset{(0.007)}{0.0028}$	—
Factor – US urban inflation	$-\underset{(0.024)}{0.0009}$	—	—
State-level consumption (monthly lags)	Iowa (1, 2, 3, 4, 6, 13); Illinois (4); Indiana (1, 2, 3); Michigan (2, 5); Wisconsin (4)	Iowa (1, 3); Illinois (—); Indiana (—); Michigan (—); Wisconsin (6)	Iowa (1, 3, 4); Illinois (—); Indiana (—); Michigan (—); Wisconsin (3)
State-level income growth, 3 month lag	—	$MD: +\underset{(0.001)}{0.1256}$	—
City per capita personal income growth, 3 month lag	—	Waterloo-Cedar Falls (IA): $-\underset{(0.003)}{0.0182}$	—
Intercept	$\underset{(0.052)}{0.0050}$	$+\underset{(0.006)}{0.0075}$	$+\underset{(0.001)}{0.0096}$
In-sample $\overline{R}^2$, RMSE (1976m2 – 2015m3)	**0.5565, 0.00220**	0.4433, 0.00246	0.3696, 0.00277
Out-of-sample RMSE (2015m4 – 2016m4)	0.001858	0.001829	**0.001765**

Table 4: Estimates of High Dimensional Vector Error Correction Model (Michigan)

Dep.var.: $\Delta \ln C_t$ (p-values in parentheses)	**Panel VECM**	**HD VECM (Lasso)**	**HD VECM (IIS)**
Short run dynamics	$\underset{(0.030)}{0.0935}\Delta \ln Y_t$ $\underset{(0.362)}{-0.0025}\Delta \pi_t$	$\underset{(0.250)}{-0.0501}\Delta \ln Y_t$ $\underset{(0.704)}{+0.0225}\Delta \pi_t$	$\underset{(0.891)}{0.0056}\Delta \ln Y_t$ $\underset{(0.754)}{+0.0180}\Delta \pi_t$
ECM Partial adjustment	$\underset{(0.908)}{+0.0004}$	$\underset{(0.924)}{-0.0003}$	$\underset{(0.065)}{-0.0045}$ $(\ln C - 0.689 \ln Y + 7.267\pi)$
Factor – US consumption	$\underset{(0.006)}{+0.0097}$	$\underset{(0.101)}{+0.0064}$	—
Factor – US urban inflation	—	—	—
State-level consumption (monthly lags)	Michigan (1, 3, 4, 6, 8, 9, 12); Illinois (1); Indiana (1, 2, 4); Iowa (1, 2, 4); Wisconsin (—)	Michigan (1, 3, 6); Illinois (—); Indiana (—); Iowa (1); Wisconsin (—)	Michigan (1, 3, 6); Illinois (—); Indiana (1); Iowa (1, 4); Wisconsin (—)
State-level income growth, 3 month lag	—	$VA: \underset{(0.022)}{+0.3381}$ $VT: \underset{(0.005)}{+0.2903}$	—
Intercept	$\underset{(0.074)}{0.0144}$	$\underset{(0.480)}{+0.0063}$	$\underset{(0.348)}{+0.0080}$
In-sample $\overline{R}^2$, RMSE (1976m2 – 2015m3)	**0.6121, 0.00709**	0.5067, 0.00804	0.5328, 0.00783
Out-of-sample RMSE (2015m4 – 2016m4)	0.002663	0.002736	**0.002639**

Table 5: Estimates of High Dimensional Vector Error Correction Model (Wisconsin)

Dep.var.: $\Delta \ln C_t$ (p-values in parentheses)	**Panel VECM**	**HD VECM (Lasso)**	**HD VECM (IIS)**
Short run dynamics	$\underset{(0.002)}{0.3794}\Delta \ln Y_t$ $\underset{(0.808)}{-0.0156}\Delta \pi_t$	$\underset{(0.746)}{0.0504}\Delta \ln Y_t$ $\underset{(0.141)}{-0.1126}\Delta \pi_t$	$\underset{(0.000)}{0.4884}\Delta \ln Y_t$ $\underset{(0.105)}{-0.1241}\Delta \pi_t$
ECM Partial adjustment	$\underset{(0.800)}{-0.0012}$	$\underset{(0.537)}{-0.0035}$	$\underset{(0.257)}{-0.0057}$
Factor – US consumption	$\underset{(0.006)}{+0.0071}$	$\underset{(0.048)}{+0.0059}$	—
Factor – US income	$\underset{(0.028)}{-0.2801}$	—	—
State-level consumption (monthly lags)	Wisconsin (1, 2, 3, 4, 6, 7, 9, 12); Illinois (1, 2, 4); Indiana (1, 2, 4, 5, 7); Iowa (1, 2, 4, 5, 7); Michigan (1, 2, 4, 5)	Wisconsin (1, 3, 5); Illinois (—); Indiana (3); Iowa (1); Michigan (6)	Wisconsin (1, 3); Illinois (—); Indiana (1); Iowa (1); Michigan (6)
State-level income growth, 3 month lag	—	$OH: \underset{(0.036)}{+0.3425}$ $NC: \underset{(0.069)}{+0.2565}$	—
WI county-level income growth, 3 month lag	—	Burnett: $\underset{(0.002)}{-0.2561}$	—
Intercept	$\underset{(0.138)}{0.0100}$	$\underset{(0.284)}{+0.0087}$	$\underset{(0.292)}{+0.0085}$
In-sample $\overline{R}^2$, RMSE (1976m2 – 2015m3)	**0.6476, 0.00560**	0.4759, 0.00690	0.4610, 0.00700
Out-of-sample RMSE (2015m4 – 2016m4)	**0.002379**	0.002645	0.002527