Nonparametric Modelling of High Dimensional Time Series

Oliver Linton
University of Cambridge

Yale University, 2016

We are concerned with the nonparametric prediction/estimation problem

$$\mathbb{E}(Y_{t+k}|\mathcal{F}_t),$$

where the information set $\mathcal{F}_t$ may be infinite dimensional and the conditional expectation is not restricted to be linear or even parametric. There are two leading cases:

- Autoregression - $\mathcal{F}_t = \sigma\left(Y_s; s \leq t\right)$ is the sigma algebra generated by the sequence $\left(Y_s\right)_{s \leq t}$
- Static regression - $\mathcal{F}_t = \sigma\left(X_{jt},\ j = 1, 2, \ldots\right).$
- In practice there is usually a combination of both types of conditioning information.
- Data could be high or low frequency
- Most econometric predictions are within parametric and even linear models.

Outline

1. Survey some work on flexible nonparametric models for time series with estimation and forecasting applications
2. Some themes are compromise between curse of dimensionality versus generality
3. Discuss in detail proposal to use marginal regression functions in a model averaging approach for estimation, forecasting, and portfolio choice.

Let (Y_t, X_t) be a stationary time series process, where $X_t \in \mathbb{R}^d$.
Interested in the regression/forecasting function

$$\mathrm{E}(Y_t|X_t = x) \quad ; \quad E(Y_t|Y_{t-1}, \ldots, Y_{t-d})$$

Nonparametric methods are consistent

- m can be well estimated when the dimension d is small, but very poorly if the dimension d is high (say larger than 3) owing to the "curse of dimensionality": Stone (1980, 1982) and Ibragimov & Hasminskii (1980). Silverman (1986). Smoothness r is equally important for rates of convergence.
- Main methods include Kernel methods, sieve/series methods, splines (penalization) etc.

For standard kernel estimators with bandwidth h and kernel K

$$MSE(\widehat{m}_d(x)) \approx \overbrace{\mu_r(K) h^{2r} \left(\sum_{|u|=r} D^u m(x) \right)^2}^{\textit{bias terms}} + \overbrace{v_r(K) \frac{1}{Th^d} \frac{\sigma^2(x)}{f(x)}}^{\textit{variance terms}}.$$

The optimal bandwidth is $h \sim T^{-1/(2r+d)}$ which delivers the pointwise rate $T^{-2r/(2r+d)}$ where r is smoothness. CLT, LFCLT, etc.

- Optimality in a limited sense. Tibshirani (1984) Local Cramer-Rao bound. Fan (1993). Local linear (kernel) is Best Linear Minimax. Extends to a number of nonparametric "models" like additive.
- We are particularly interested in the case where d is large, which is not covered by this theory. My interest goes back to Pagan and Ullah (1988) and Hong and Pagan (1991) who discuss the case where $d \to \infty$

- Functional regression. Masry (2005), Ferraty and Vieu (2002,2006,2012), Ramsay and Silverman (2005),

$$Y \in \mathbb{R} \quad ; \quad X \in \mathbf{B} = L_2$$

$$Y, X \in \mathbf{B}$$

  Infill framework, e.g., $X : [0,1] \to \mathbb{R}$ with observations $X(t_1), \ldots, X(t_T)$ with $\{t_1, \ldots, t_T\}$ becoming dense in $[0,1]$.
- Functional autoregression, linear and nonlinear. Bosq (2000). Kargin and Onatski (2008).
- Kernel methods with metric $\varrho(X, X')$; Assume infill type of asymptotics with functional smoothness, Frechet/Hadamard differentiability. No covariate joint density. Consistency at slow rate.

We consider the case where $X = (X_1, X_2, \ldots) \in \ell_2$, i.e., the set of sequences for which say $\sum_{k=1}^{\infty} x_k^2$ may not be well defined.
We will weight the X_i to get back to normed space.

Linton and Sancetta (2009) considered estimation of the predictor

$$\mathrm{E}(Y_0|Y_{-1}, Y_{-2}\ldots).$$

They established (strong) consistency of kernel estimator under weak conditions (stationarity and ergodicity and moment).
Hong and Linton (2016) consider the problem of estimating

$$E(Y_t|\mathbf{X}_t = x),$$

where $\mathbf{X}_t$ could be lagged $Y's$ or $(X_t, X_{t-1}, \ldots)$. In either case there is an ordering of the covariate which is exploited in estimation.

Kernel estimator aggregates covariate (lags) through metric between sequences inside the kernel.

$$\widehat{\mathrm{E}}(Y_0|Y_{-1}, Y_{-2}, \ldots) = \frac{\sum_{s=1}^{T} Y_{-s} K_h \left(\overbrace{\sum_{j=1}^{T-s} \phi_j^{-1} \rho(Y_{-j}, Y_{-s-j})}^{\varrho(Y_{-T}^{-s-1}, Y_{-(T-s)}^{-1})} \right)}{\sum_{s=1}^{T} K_h \left(\sum_{j=1}^{T-s} \phi_j^{-1} \rho(Y_{-j}, Y_{-s-j}) \right)}$$

where $K(\cdot)$ is a kernel function, $K_h(\cdot/h)/h$ and h is a bandwidth. Downweights the past through parameter $\phi_j \longrightarrow \infty$. Looks for sequences $(Y_{-s}, Y_{-s-1}, \ldots)$ close to the most recent history $(Y_{-1}, Y_{-2}, \ldots,)$

Can show that in the special case where $K(\sum u_i) = \prod K(u_i)$ (e.g., exponential) and $\rho(t,s) = (t-s)^2$

$$\widehat{\mathrm{E}}(Y_0|Y_{-1}, Y_{-2}\ldots) = \frac{\sum_{s=1}^{T} Y_{-s} \prod_{j=1}^{T-s} K_{h_j}(Y_{-j} - Y_{-s-j})}{\sum_{s=1}^{T} \prod_{j=1}^{T-s} K_{h_j}(Y_{-j} - Y_{-s-j})}, \quad h_j = \phi_j h,$$

where $h \to 0$ as $T \to \infty$ and $\phi_j \to \infty$ as $j \to \infty$.
Method is multivariate kernel regression with bandwidths expanding with lag order. Achieves shrinkage indirectly. No explicit truncation.
Application to Risk return relationship.

Some Heuristics for high dimensional np regression

Consider np regression with $d(T) \to \infty$. Then for standard kernel estimators with bandwidth h

$$MSE(\widehat{m}_d(x)) \approx \overbrace{\mu_r(K) h^{2r} \left(\sum_{|u|=r} D^u m(x) \right)^2}^{\textit{bias} \text{ terms}} + \overbrace{v_r(K) \frac{1}{Th^d} \frac{\sigma^2(x)}{f(x)}}^{\textit{variance} \text{ terms}}$$

Suppose that $d = d(T) \to \infty$ then log of the optimal MSE is

$$\log O(T^{-2r/(2r+d(T))}) = \frac{-2r}{2r + d(T)} \log T \to -\infty$$

provided $\log T / d(T) \to \infty$ (rules out $d(T) = c \log T$). Heuristics not quite right.

Hong and Linton (2016)

$$Y_t = m(X_t, X_{t-1}, \ldots) + \varepsilon_t,$$

where $E(\varepsilon_t | X_t, X_{t-1}, \ldots) = 0$, and:

1. *The sample observations* $\{Y_t, X_t\}_{t=1}^T$ *are stationary and (jointly) near epoch dependent with some rate not specified here and* $E(|Y_t|^{2+\delta}) < \infty$.
2. *The real-valued stochastic process* $\{X_s\}_{s=1}^\infty$ *admits a moving average representation:*

$$X_s = \sum_{j=-\infty}^{\infty} a_j \epsilon_{s-j},$$

where a_j *is a square summable sequence, and* $\{\epsilon_j\}_j$ *is an independent and identically distributed standard Gaussian sequence.*

3. *We have for* $m : \mathbb{R}^\infty \to \mathbb{R}$,

$$\left| m(x) - m(x') \right| \leq \sum_{j=1}^{\infty} c_j \left| x_j - x_j' \right|^\beta$$

where $\{c_j\}$ *is a sequence with* $\sum_{j=1}^\infty c_j j^{p\beta} < \infty$ *and* $\beta \in (0, 1]$.

- A major issue concerns the small ball (or deviation) probability

$$\varphi_z(\delta) := \Pr\left(\|z - Z\| \leq \delta\right) \quad \text{as} \quad \delta \to 0$$

- When Z is a d-dimensional continuous random vector with Lebesgue density $p(\cdot) > 0$, it can be readily shown that the shifted small ball (with respect to the usual Euclidean norm) is given by

$$\varphi_z(h) = V_d h^d p(z) \propto h^d,$$

  where $V_d = \pi^{d/2}/\Gamma(d/2+1)$ is the volume of d-dimensional unit sphere. However when Z takes values in an infinite-dimensional space, it is generally difficult to specify the exact form of the small ball probability, and its behaviour varies depending heavily on the nature of the associated space and its topological structure.
- Hong, Lifshits and Nazarov (2016, Electron Comm. Probab) establish results for weakly dependent Gaussian sequences.

Suppose that K is supported on (and bounded away form zero on) $[0, \overline{K}]$, that $h_j = j^p h$, where $p > 1/2$ and $h = h(T) \to 0$

$$
\begin{aligned}
C_G^* &= \frac{(2\pi)^{(1-p)}(2p-1)}{2 \cdot (2p)^{\frac{3p-1}{2p-1}}} \cdot \left[\frac{\pi 2^{(1-2p)/2p}}{\sin(\pi/2p)} \right]^{\frac{-p}{2p-1}}, \\
C_G^{**} &= \frac{2p-1}{2} \left(\frac{\pi}{2p \sin \frac{\pi}{2p}} \right)^{\frac{2p}{2p-1}} \\
C_\ell &= \lim_{h \to 0} \left[\ell^{-1/2} \left(h^{-\frac{4p}{2p-1}} \right) \right] = \sqrt{\frac{2}{\pi}} \\
C_{\mathcal{A}} &= \left[\frac{1}{2\pi} \int_0^{2\pi} \left| \sum_{j=0}^{\infty} a_j \exp(\mathrm{i} j u) \right|^{1/p} du \right]^p \\
\mathcal{V}_2(x) &= \frac{C_G^* C_\ell}{\overline{K}^{\frac{1-p}{2p-1}}} \frac{\sigma^2(x)}{e^{-\frac{1}{2}\|\Gamma^{-1/2} z\|_2^2} C_{\mathcal{A}}^{\frac{1-p}{2p-1}}} = v(x; p, \overline{K}, a),
\end{aligned}
$$

where $\sigma^2(\cdot)$ is the conditional variance, and $z = (z_j) = (j^{-p} x_j)$. $(1^{-p} X_1, 2^{-p} X_2, \ldots)^\intercal =: Z$ is $(\ell_2, \|.\|_2)$-valued

Theorem *Suppose A1-A8 and B1-B4 hold. Then,*

$$\Delta_T^{1/2}(x)\left(\widehat{m}(x) - m(x) - \mathcal{B}_T(x)\right) \Longrightarrow N\left(0,1\right),$$

where $\mathcal{B}_T(x) = O(h^{\beta}) \longrightarrow 0$ *is the bias component and*

$$\Delta_T(x) = Th^{\frac{1-p}{2p-1}} \exp\left\{-C_G^{**}(C_{\mathcal{A}}\overline{K})^{-\frac{2}{2p-1}} h^{-\frac{2}{2p-1}}\right\} \mathcal{V}_2(x)^{-1}.$$

Limiting variance depends on dependence structure of covariates

Bias Variance Trade-off

Suppose that

$$h = (\log T)^a$$

Then the following bandwidth sequence constant balances squared bias against variance

$$a_{opt} = \frac{\vartheta \cdot \mathcal{W}(\frac{\chi}{\vartheta} T^{\chi/\vartheta}) - \chi \log T}{\vartheta \times \chi \times \log\log T} \in \left(-\frac{2p-1}{2}, 0\right) \downarrow \frac{1}{2} - p,$$

where $\mathcal{W}(y)$ is the lambert W function, which returns the solution x of $y = x \cdot e^x$, while $\vartheta = [2\beta + (1-p)/(2p-1)]$ and $\chi = 2/(2p-1)$. The upper bound on p is obtained from the bias condition $p_{\max}(\beta, c)$, hence optimal mse is bounded by

$$(\log T)^{\beta\left(\frac{1}{2} - p_{\max}\right)}.$$

Inference

Self normalized CLT allows usual inference. Let

$$\widehat{V}(x) = \left(\sum_{t=1}^{T} K_t \right)^{-2} \sum_{t=1}^{T} \left\{ K_t (Y_t - \widehat{m}(x)) - \frac{1}{T} \sum_{t=1}^{T} K_t (Y_t - \widehat{m}(x)) \right\}^2$$

$$\widehat{V}(x)^{-1/2} \left(\widehat{m}(x) - m(x) - \mathcal{B}_T(x) \right) \Longrightarrow N(0,1)$$

$$K_t = K\left(\| H^{-1}(x - X_t) \| \right) \quad ; \quad H = (h_1, h_2, \ldots).$$

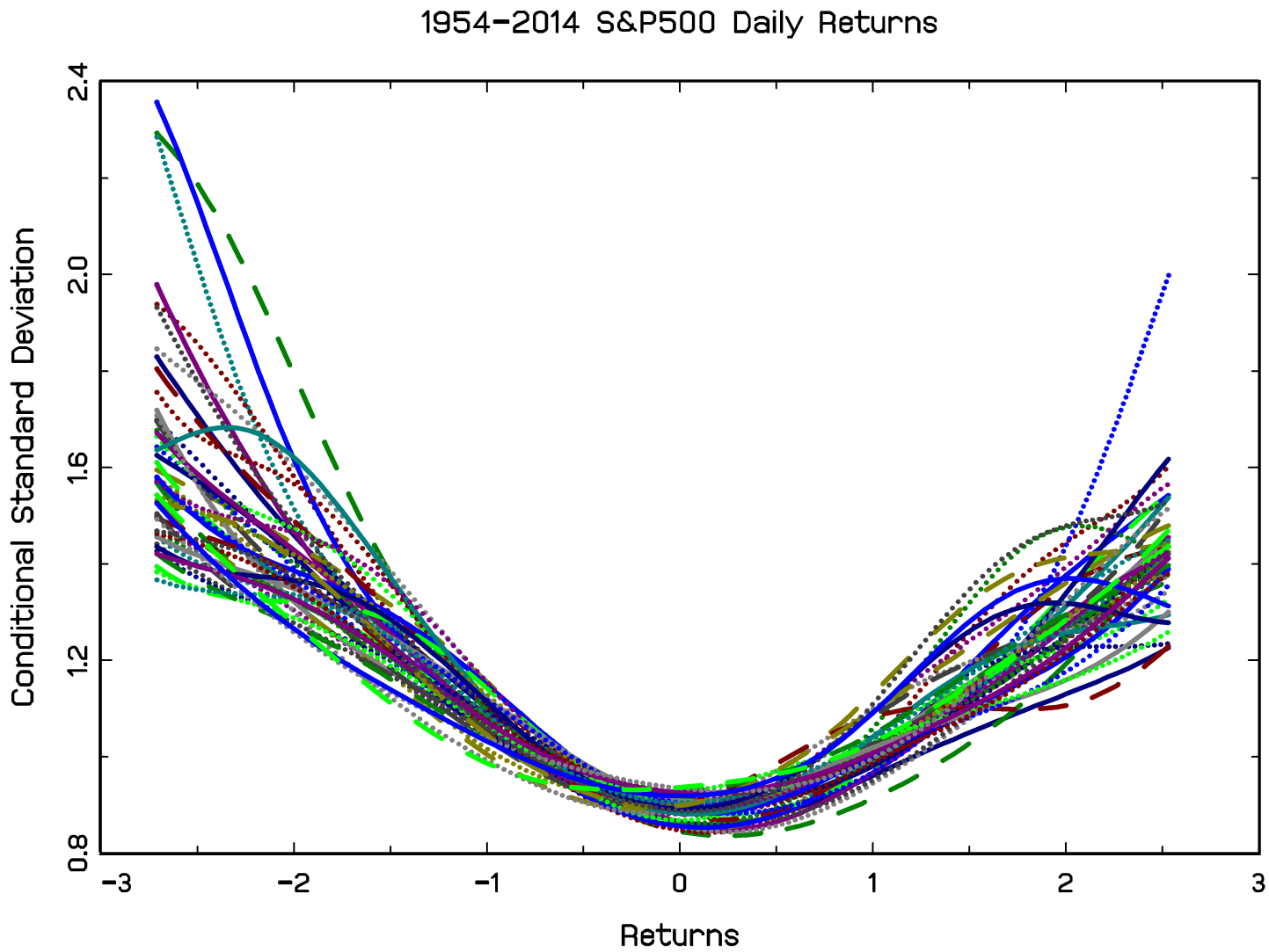
How to get faster rates of convergence?

- Change the function class to impose additional structure implicitly, e.g., Neural networks.
- Assume that m is infinitely smooth or the smoothness increases with sample size. For example, if smoothness of order $r(T) = \gamma \log T$ and $d(T) = c \log T$, we get optimal MSE of order

$$O(T^{-2\gamma \log T/(2\gamma \log T + c \log T)}) = O(T^{-2\gamma/(2\gamma + c)})$$

  Hard to justify.
- Impose some structure

Estimates of $\sqrt{E(Y_{t+j}^2 | Y_t = y)}$, $j = 1, \ldots, 50$. $y \in [0.01\%, 0.99\%]$



Additive Nonparametric Models

Generalized Additive models (Hastie and Tibshirani, 1990)). For known G

$$G(m(x)) = \sum_{j=1}^{d} m_j(x_j)$$

Stone (1985, 1986), free from curse of dimensionality.
Estimation by Marginal Integration (averaging out of high dimensional estimator) or Backfitting (Iterative one dimensional smooths of partial residuals) $G = I$

$$m_j(X_{jt}) \longleftarrow E\left[\overbrace{\left\{Y_t - \sum_{k\neq j}^{d} m_k(X_{kt})\right\}}^{\text{partial residuals}} | X_{jt}\right]$$

Hastie and Tibshirani (1990). Mammen, Linton, and Nielsen (1999). Smooth backfitting
Identification issue. Rule out concurvity (i.e., singularity in joint distribution of $(X_1, \ldots, X_d)$).

- In the time series case the conditioning information may consist of an infinite number of lags of Y, i.e., $d = \infty$.
- Sparse additive models (SpAM). Meier, van de Geer, and Buhlman (2009). Potential components $d >> T$ but true model d_0 finite. iid data
- We now consider semiparametric models that use all the lags but in some special structure

Semiparametric Models

- Linton and Mammen (2005) considered the Engle and Ng (1993) semiparametric (volatility) regression model

$$\mathrm{E}(Y_t^2 | Y_{t-1}, Y_{t-2} \ldots) = \sum_{j=1}^{\infty} \psi_j(\theta) m(Y_{t-j}),$$

  where $m(\cdot)$ is an unknown function and the parametric family $\{\psi_j(\theta), \theta \in \Theta, j = 1, \ldots, \infty\}$ are parametric weights that decline sufficiently fast. This model includes the GARCH(1,1) as a special case and includes an infinite set of lags.
- Linton and Mammen (2008) generalized this class of models to allow for exogenous regressors and more complicated dynamics

$$B(L) Y_t = A(L) m(X_t) + \varepsilon_t$$

  including some unit roots (in Y)
- Chen and Ghyssels (2010, RFS) model. Application of these methods to volatility forecasting and news impact curve estimation using high frequency data.

Estimation method is based on a characterization of the function m as the solution of a linear integral equation

$$m = m^* + \mathcal{H}m$$

with operator $\mathcal{H}$ and intercept of the form

$$m^*_\theta(x) = \sum_{j=1}^{\infty} \psi_j(\theta) m_j(x), \;\; m_j(x) = \mathsf{E}(Y_t^2 | Y_{t-j} = x)$$

They proposed an estimation strategy for the unknown quantities, which requires as input the estimation of $m_j(x)$ for $j = 1, 2, \ldots, d(T)$, where $d(T) = c \log T$ for some $c > 0$. Solve the inverse problem numerically. Essentially, iterative one dimensional smoothing.

- This general approach to modelling is promising but quite computationally demanding.
- In addition, the models considered thus far all have a finite number of unknown functions (for example, in Linton and Mammen (2005) only one unknown function was allowed), and so appear to be heavily over identified. Not flexible enough for forecasting.
- Other approaches (Fan and Yao 2003) :
  - varying coefficient models
  - Partial linear
  - Single index models, for example

$$\mathsf{E}(Y_t^2 | Y_{t-1}, Y_{t-2} \ldots) = m\left(\sum_{j=1}^{\infty} \psi_j v(Y_{t-j}; \theta)\right)$$

The infinite order additive model with decay may have some advantages

$$Y_t = \sum_{j=1}^{\infty} m_j(Y_{t-j}) + \varepsilon_t, \quad m_j \in S(r_j, c_j),$$

where smoothness r_j increases with lag j (i.e., number of derivatives increases or global bound on second derivatives shrinks)
Maybe backfitting with increasing order can be plausible

$$m_j(Y_{t-j}) \longleftarrow E\left[\overbrace{\left\{Y_t - \sum_{k\neq j}^{d(T)} m_k(Y_{t-k})\right\}}^{\text{partial residuals}} | Y_{t-j}\right]$$

Take bandwidth $h_j = c_j h(T)$ such that

$$h_j(T) \to \infty \text{ as } j \to \infty \text{ and } h_j \to 0 \text{ as } T \to \infty$$

MAMA: Model Averaging MArginal Regression

Li, Linton, and Lu (2014). Suppose that Y_t, $t = 1, \ldots, T$, are T observations collected from a stationary time series process and $\mathbf{X}_t = (\mathbf{Z}_t^\intercal, \mathbf{Y}_{t-1}^\intercal)^\intercal$ with $\mathbf{Z}_t = \left(Z_{t1}, Z_{t2}, \ldots, Z_{tp_T}\right)^\intercal$ and $\mathbf{Y}_{t-1} = \left(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-d_T}\right)^\intercal$, where $\mathbf{Z}_t$ are exogenous regressors.
Here, p_T, d_T are large i.e., $p_T, d_T \longrightarrow \infty$.
The main interest is to study the multivariate regression function:

$$m(\mathbf{x}) = \mathrm{E}\left(Y_t \middle| \mathbf{X}_t = \mathbf{x}\right).$$

We propose methods to approximate this by one dimensional smoothing operations

We model or approximate $m(x) = \mathsf{E}(Y|X = x)$ by

$$m_w(x) = w_0 + \sum_{j=1}^{J} w_j \mathsf{E}(Y|X_{(j)} = x_{(j)})$$

for some weights w_j, $j = 0, 1, \ldots, J$, where $X_{(j)} = (X_{i_1}, \ldots, X_{i_{k_j}})^\intercal$ is a subset of X and $x_{(j)} = (x_{i_1}, \ldots, x_{i_{k_j}})^\intercal$.
In general, $X_{(j)}$ and $X_{(k)}$ could have different dimensions and of course overlapping members. The union of $X_{(j)}$ may exhaust one or more of S_ℓ (the set of all subsets of $S = \{1, 2, \ldots, d\}$ of ℓ components, and this has cardinality $J_\ell = \binom{d}{\ell}$) or it may not. In a time series setting we may be combining

$$E(Y_t|Y_{t-1}), \ldots, E(Y_t|Y_{t-d}), E(Y_t|Y_{t-1}, Y_{t-2}), \ldots, E(Y_t|Y_{t-2}, \ldots, Y_{t-d})$$

A simple special case that we focus on is where $J = d$ and $X_{(j)} = X_j$ is just the j^{th} component and the covariates are non overlapping.

KSIS+PMAMAR

Chen, Li, Linton, Lu (2016). We assume that the dimension the potential explanatory variables $\mathbf{X}_t$ can be diverging at an exponential rate, i.e.,

$$p_T + d_T = O(\exp\{T^{\delta_0}\})$$

for some positive constant δ_0.

- We introduce screening method to reduce the number of possible variables to less than T
- We then apply penalization technique to select down even further to $k(T)$ with $k(T)/T \to 0$ and apply the aggregation
- A practically important refinement involves iterating between the screening step and the PMAMAr step.
- We examine the nonlinear forecasting performance after dimension reduction.

Linear models: Fan and Lv (2008); Additive models: Fan, Feng and Song (2011); Varying coefficient models: Fan, Ma and Dai (2014) and Liu, Li and Wu (2014); Linear SIS+model averaging: Ando and Li (2014).

For notational simplicity, we let

$$X_{tj} = \begin{cases} Z_{tj}, & j = 1, 2, \ldots, p_T, \\ Y_{t-(j-p_T)}, & j = p_T + 1, p_T + 2, \ldots, p_T + d_T. \end{cases}$$

For $j = 1, \ldots, p_T + d_T$, the kernel smoother of marginal regression

$$m_j(x_j) := \mathrm{E}(Y_t | X_{tj} = x_j)$$

is

$$\hat{m}_j(x_j) = \frac{\sum_{t=1}^{T} Y_t K_{tj}(x_j)}{\sum_{t=1}^{T} K_{tj}(x_j)}, \quad K_{tj}(x_j) = K\Big(\frac{X_{tj} - x_j}{h_1}\Big).$$

Screening Step

We consider ranking the importance of the covariates by calculating the correlation between the response variable and marginal regression:

$$\text{cor}(j) = \frac{\text{cov}(j)}{\sqrt{\text{v}(Y) \cdot \text{v}(j)}} = \left[\frac{\text{v}(j)}{\text{v}(Y)}\right]^{1/2},$$

where $\text{v}(Y) = \text{var}(Y_t)$, $\text{v}(j) = \text{var}(m_j(X_{tj}))$ and

$$\text{cov}(j) = \text{cov}(Y_t, m_j(X_{tj})) = \text{var}(m_j(X_{tj})) = \text{v}(j).$$

The value of $\text{cor}(j)$ is non-negative for all j and the ranking of $\text{cor}(j)$ is equivalent to the ranking of $\text{v}(j)$ as $\text{v}(Y)$ is positive and invariant across j.

The sample version of $\mathrm{cor}(j)$ can be constructed as

$$\hat{\mathrm{cor}}(j) = \frac{\hat{\mathrm{cov}}(j)}{\sqrt{\hat{\mathrm{v}}(Y) \cdot \hat{\mathrm{v}}(j)}} = \left[\frac{\hat{\mathrm{v}}(j)}{\hat{\mathrm{v}}(Y)}\right]^{1/2},$$

where

$$\hat{\mathrm{v}}(Y) = \frac{1}{T}\sum_{t=1}^{T} Y_t^2 - \left(\frac{1}{T}\sum_{t=1}^{T} Y_t\right)^2,$$

$$\hat{\mathrm{cov}}(j) = \hat{\mathrm{v}}(j) = \frac{1}{T}\sum_{t=1}^{T} \hat{m}_j^2(X_{tj}) - \left[\frac{1}{T}\sum_{t=1}^{T} \hat{m}_j(X_{tj})\right]^2,$$

The screened sub-model can be determined by,

$$\hat{\mathcal{S}} = \left\{j = 1, 2, \ldots, p_T + d_T : \ \hat{\mathrm{v}}(j) \geq \rho_T\right\},$$

where ρ_T is a pre-determined positive number.

The above criterion is equivalent to

$$\hat{\mathcal{S}} = \left\{ j = 1, 2, \ldots, p_T + d_T : \ \hat{\text{cor}}(j) \geq \rho_T^{\diamond} \right\},$$

where $\rho_T^{\diamond} = \rho_T^{1/2} / \sqrt{\hat{\text{v}}(Y)}$.
Define the index set of "true" candidate models as

$$\mathcal{S} = \left\{ j = 1, 2, \ldots, p_T + d_T : \ \text{v}(j) \neq 0 \right\}.$$

Theorem *Suppose that the conditions A1–A5 in the paper are satisfied. For any small* $\delta_1 > 0$*, there exists a positive constant* δ_2 *such that*

$$\begin{aligned} &\Pr\left(\max_{1 \leq j \leq p_T + d_T} |\hat{\text{v}}(j) - \text{v}(j)| > \delta_1 T^{-2(1-\theta_1)/5} \right) \\ &= O\left(M(T) \exp\left\{ -\delta_2 T^{(1-\theta_1)/5} \right\} \right), \end{aligned}$$

where $M(T) = (p_T + d_T) T^{(17+18\theta_1)/10}$ *and* $1/6 < \theta_1 < 1$ *is defined by* $h_1 = T^{-\theta_1}$.

Theorem *If we choose the pre-determined tuning parameter* $\rho_T = \delta_1 T^{-2(1-\theta_1)/5}$ *and assume*

$$\min_{j \in \mathcal{S}} \mathrm{v}(j) \geq 2\delta_1 T^{-2(1-\theta_1)/5},$$

then we have

$$\Pr\left(\mathcal{S} \subset \hat{\mathcal{S}}\right) \geq 1 - O\left(M_{\mathcal{S}}(T) \exp\left\{-\delta_2 T^{(1-\theta_1)/5}\right\}\right),$$

where $M_{\mathcal{S}}(T) = |\mathcal{S}| T^{(17+18\theta_1)/10}$ *with* $|\mathcal{S}|$ *being the cardinality of* $\mathcal{S}$.
As $p_T + d_T = O(\exp\{T^{\delta_0}\})$, in order to ensure the validity of Theorem 1(i), we need to impose the restriction $\delta_0 < (1-\theta_1)/5$, which reduces to $\delta_0 < 4/25$ if the order of the optimal bandwidth in kernel smoothing (i.e., $\theta_1 = 1/5$) is used.

PMAMA step

We denote the chosen covariates (after KSIS in the first step) by $\mathbf{X}_t^* = \left(X_{t1}^*, X_{t2}^*, \ldots, X_{tq_T}^*\right)^\intercal$ which may include both exogenous variables and lags of Y_t, where q_T might be divergent but is smaller than the sample size T.

We approximate the conditional regression function $m^*(\mathbf{x}) = \mathrm{E}(Y_t|\mathbf{X}_t^* = \mathbf{x})$ by an affine combination of one-dimensional conditional component regressions

$$m_j^*(x_j) = \mathrm{E}(Y_t|X_{tj}^* = x_j), \quad j = 1, \ldots, q_T.$$

Each marginal regression $m_j^*(\cdot)$ can be treated as a "nonlinear candidate model". That is,

$$m^*(\mathbf{x}) \approx w_0 + \sum_{j=1}^{q_T} w_j m_j^*(x_j),$$

where w_j, $j = 0, 1, \ldots, q_T$, are to be determined later and can be seen as the weights for different "candidate models".

For $j = 1, \ldots, q_T$, we estimate the marginal regression functions $m_j^*(\cdot)$ by the kernel smoothing method:

$$\hat{m}_j^*(x_j) = \frac{\sum_{t=1}^T Y_t \overline{K}_{tj}(x_j)}{\sum_{t=1}^T \overline{K}_{tj}(x_j)}, \quad \overline{K}_{tj}(x_j) = K\Big(\frac{X_{tj}^* - x_j}{h_2}\Big).$$

Then, for $j = 1, \ldots, q_T$, we let

$$\hat{\mathcal{M}}(j) = \big[\hat{m}_j^*(X_{1j}^*), \ldots, \hat{m}_j^*(X_{Tj}^*)\big]^\intercal =: \mathcal{S}_T(j)\mathcal{Y}_T$$

be the estimated values of

$$\mathcal{M}(j) = \big[m_j^*(X_{1j}^*), \ldots, m_j(X_{Tj}^*)\big]^\intercal,$$

where $\mathcal{S}_T(j)$ is the $T \times T$ smoothing matrix whose (k, l)-component is $\overline{K}_{lj}(X_{kj}^*) / \big[\sum_{t=1}^T \overline{K}_{tj}(X_{kj}^*)\big]$, and $\mathcal{Y}_T = (Y_1, \ldots, Y_T)^\intercal$.

We define the objective function by

$$\mathcal{Q}_T(\mathbf{w}_T) = \left[\mathcal{Y}_T - \hat{\mathcal{M}}(\mathbf{w}_T)\right]^{\intercal}\left[\mathcal{Y}_T - \hat{\mathcal{M}}(\mathbf{w}_T)\right] + T\sum_{j=1}^{q_T} p_\lambda(|w_j|),$$

where

$$\hat{\mathcal{M}}(\mathbf{w}_T) = \left[w_1\mathcal{S}_T(1) + \ldots + w_{q_T}\mathcal{S}_T(q_T)\right]\mathcal{Y}_T = \mathcal{S}_T(\mathcal{Y})\mathbf{w}_T,$$

$\mathcal{S}_T(\mathcal{Y}) = \left[\mathcal{S}_T(1)\mathcal{Y}_T, \ldots, \mathcal{S}_T(q_T)\mathcal{Y}_T\right]$, and $p_\lambda(\cdot)$ is a penalty function with a tuning parameter λ.
Our semiparametric estimator of the optimal weights $\mathbf{w}_o$ can be obtained through minimising the objective function $\mathcal{Q}_T(\mathbf{w}_T)$:

$$\hat{\mathbf{w}}_T = \arg\min_{\mathbf{w}_T} \mathcal{Q}_T(\mathbf{w}_T).$$

AIC and BIC: $p_\lambda(|z|) = 0.5\lambda^2 I(|z| \neq 0)$ with different values of λ;
LASSO: $p_\lambda(|z|) = \lambda|z|$;
SCAD: $p'_\lambda(z) = \lambda\left[I(z \leq \lambda) + \frac{a_0\lambda - z}{(a_0-1)\lambda} I(z > \lambda)\right]$ with $p_\lambda(0) = 0$, where $a_0 > 2$, $\lambda > 0$ and $I(\cdot)$ is the indicator function.
In our numerical studies, we use the SCAD penalty in PMAMAR.

Theorem *Suppose that the conditions A1–A8 are satisfied. There exists a local minimizer* $\hat{\mathbf{w}}_T$ *of the objective function* $\mathcal{Q}_T(\cdot)$ *such that*

$$\|\hat{\mathbf{w}}_T - \mathbf{w}_o\| = O_P\Big(\sqrt{q_T}(T^{-1/2} + a_T)\Big),$$

where $\|\cdot\|$ *denotes the Euclidean norm and*

$$a_T = \max_{1 \le j \le q_T} \big\{|p'_\lambda(|w_{oj}|)|,\ |w_{oj}| \neq 0\big\}.$$

Theorem *Let* $\hat{\mathbf{w}}_T(2)$ *be the estimator of* $\mathbf{w}_o(2)$ *which is composed of all the zero weights and further assume that*

$$\lambda \to 0, \quad \frac{\sqrt{T}\lambda}{\sqrt{q_T}} \to \infty, \quad \liminf_{T\to\infty} \liminf_{\vartheta \to 0+} \frac{p'_\lambda(\vartheta)}{\lambda} > 0.$$

Then, the local minimizer $\hat{\mathbf{w}}_T$ *of the objective function* $\mathcal{Q}_T(\cdot)$ *satisfies* $\hat{\mathbf{w}}_T(2) = \mathbf{0}$ *with probability approaching one.*

Theorem *If we further assume that the eigenvalues of* $\mathbf{\Lambda}_{T1}$ *are bounded away from zero and infinity,*

$$\sqrt{T}\mathbf{A}_T\mathbf{\Sigma}_T^{-1/2}\left(\mathbf{\Lambda}_{T1}+\mathbf{\Omega}_T\right)\left[\hat{\mathbf{w}}_T(1)-\mathbf{w}_o(1)-\left(\mathbf{\Lambda}_{T1}+\mathbf{\Omega}_T\right)^{-1}\boldsymbol{\omega}_T\right] \xrightarrow{d} \mathrm{N}\left(\mathbf{0},\mathbf{A}_0\right),$$

where $\mathbf{0}$ *is a null vector whose dimension may change from line to line,* $\mathbf{A}_T$ *is an* $s \times s_T$ *matrix such that* $\mathbf{A}_T\mathbf{A}_T^{\intercal} \longrightarrow \mathbf{A}_0$ *and* $\mathbf{A}_0$ *is an* $s \times s$ *symmetric and non-negative definite matrix,* s *is a fixed positive integer. The definitions of* $\mathbf{\Sigma}_T$, $\mathbf{\Lambda}_{T1}$, $\mathbf{\Omega}_T$ *and* $\boldsymbol{\omega}_T$ *are given in the paper.*

Simulations

The sample size is set to be $T = 100$, and the numbers of candidate exogenous covariates and lagged terms are $(p_T, d_T) = (30, 10)$ and $(p_T, d_T) = (150, 50)$. The model is defined by

$$\begin{aligned} Y_t &= m_1(Z_{t1}) + m_2(Z_{t2}) + m_3(Z_{t3}) + m_4(Z_{t4}) + m_5(Y_{t-1}) \\ &\quad + m_6(Y_{t-2}) + m_7(Y_{t-3}) + \varepsilon_t \end{aligned}$$

for $t \geq 1$, where, following Meier, van de Geer and Bühlmann (2009), we set

$$m_i(x) = \sin(0.5\pi x), \qquad i = 1, 2, \ldots, 7.$$

The exogenous covariates

$$\mathbf{Z}_t = (Z_{t1}, Z_{t2}, \ldots, Z_{tp_T})^{\intercal}$$

are independently drawn from p_T-dimensional Gaussian distribution with zero mean and covariance matrix $\mathrm{cov}(\mathbf{Z}) = I_{p_T}$ or $C_{\mathbf{Z}}$, whose the main-diagonal entries are 1 and off-diagonal entries are $1/2$. The error term ε_t are independently generated from the $\mathrm{N}(0, 0.7^2)$ distribution. The real size of exogenous regressors is 4 and the real lag length is 3.

We generate $100 + T$ observations from the process with initial states $Y_{-2} = Y_{-1} = Y_0 = 0$ and discard the first $100 - d_T$ observations.
The iterative version of KSIS+PMAMAR performs better in both estimation and prediction than the KSIS+PMAMAR.
The penGAM is the most conservative in variable selection and on average selects the least number of variables.
The ISIS suffers from the model misspecification problem.
When the correlation among the exogenous variables increases, the performance of all approaches worsens.

PCA+PMAMAR

- Impose an approximate factor modelling structure on the ultra-high dimensional exogenous regressors and use the well-known principal component analysis to estimate the latent common factors;
- Apply the PMAMAR method to select the estimated common factors and lags of the response variable which are significant.

Letting

$$\mathbf{B}_T^0 = (\mathbf{b}_1^0, \ldots, \mathbf{b}_{p_T}^0)^\intercal \text{ and } \mathbf{U}_t = (u_{t1}, \ldots, u_{tp_T})^\intercal,$$

we assume the approximate factor model:

$$\mathbf{Z}_t = \mathbf{B}_T^0 \mathbf{f}_t^0 + \mathbf{U}_t,$$

where $\mathbf{b}_k^0$ is an r-dimensional vector of factor loadings, $\mathbf{f}_t^0$ is an r-dimensional vector of common factors, and u_{tk} is called an idiosyncratic error.
Denote $\mathcal{Z}_T = (\mathbf{Z}_1, \ldots, \mathbf{Z}_T)^\intercal$, the $T \times p_T$ matrix of the observations of the exogenous variables. We then construct

$$\hat{\mathcal{F}}_T = (\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_T)^\intercal$$

as the $T \times r$ matrix consisting of the r eigenvectors (multiplied by $\sqrt{T}$) associated with the r largest eigenvalues of the $T \times T$ matrix

$$\mathcal{Z}_T \mathcal{Z}_T^\intercal / (T p_T).$$

Define

$$\mathbf{H} = \hat{\mathbf{V}}^{-1}\left(\hat{\mathcal{F}}_T^{\intercal}\mathcal{F}_T^0/T\right)\left[(\mathbf{B}_T^0)^{\intercal}\mathbf{B}_T^0/p_T\right], \quad \mathcal{F}_T^0 = \left(\mathbf{f}_1^0, \ldots, \mathbf{f}_T^0\right)^{\intercal},$$

and $\hat{\mathbf{V}}$ is the $r \times r$ diagonal matrix of the first r largest eigenvalues of $\mathcal{Z}_T\mathcal{Z}_T^{\intercal}/(Tp_T)$ arranged in descending order.

Theorem. *Suppose that the conditions B1–B4 are satisfied, and*

$$T = o(p_T^2), \quad p_T = O\left(\exp\{T^{\delta_*}\}\right), \quad 0 \le \delta_* < 1/3.$$

For the PCA estimation $\hat{\mathbf{f}}_t$*, we have*

$$\max_t \left\|\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t^0\right\| = O_P\left(T^{-1/2} + T^{1/4}p_T^{-1/2}\right).$$

Consider the following multivariate regression function with rotated latent factors and lags of response:

$$m_f^*(\mathbf{x}_1, \mathbf{x}_2) = \mathrm{E}\left(Y_t | \mathbf{H}\mathbf{f}_t^0 = \mathbf{x}_1, \mathbf{Y}_{t-1} = \mathbf{x}_2\right).$$

Apply the PMAMAR with

$$\hat{\mathbf{X}}_{t,f}^* = \left(\hat{\mathbf{f}}_t^\intercal, \mathbf{Y}_{t-1}^\intercal\right)^\intercal = \left(\hat{f}_{t1}, \ldots, \hat{f}_{tr}, \mathbf{Y}_{t-1}^\intercal\right)^\intercal.$$

For $k = 1, \ldots, r$, define

$$m_{k,f}^*(z_k) = \mathrm{E}\left[Y_t | \tilde{f}_{tk}^0 = z_k\right], \quad \tilde{f}_{tk}^0 = e_r^\intercal(k)\mathbf{H}\mathbf{f}_t^0,$$

where $e_r(k)$ is an r-dimensional column vector with the k-th element being one and zeros elsewhere, $k = 1, \ldots, r$. We estimate $m_{k,f}^*(z_k)$ by the kernel smoothing method:

$$\hat{m}_{k,f}^*(z_k) = \frac{\sum_{t=1}^T Y_t \widetilde{K}_{tk}(z_k)}{\sum_{t=1}^T \widetilde{K}_{tk}(z_k)}, \quad \widetilde{K}_{tk}(z_k) = K\left(\frac{\hat{f}_{tk} - z_k}{h_3}\right), \quad j = 1, \ldots r,$$

where h_3 is a bandwidth and $\hat{f}_{tk}$ is the k-th element of $\hat{\mathbf{f}}_t$.

Theorem *Suppose that the conditions A5 and B1–B5 are satisfied, and the latent factor* $\mathbf{f}_t^0$ *has a compact support. Then we have*

$$\max_{1\leq k\leq r} \sup_{z_k\in\mathcal{F}_k^*} \left|\hat{m}_{k,f}^*(z_k) - \tilde{m}_{k,f}^*(z_k)\right| = o_P\left(T^{-1/2}\right),$$

where $\mathcal{F}_k^*$ *is the compact support of* $\tilde{f}_{tk}^0$, $\tilde{m}_{k,f}^*(z_k)$ *is the infeasible kernel estimation defined as* $\hat{m}_{k,f}^*(z_k)$ *but with* $\hat{f}_{tk}$ *replaced by* $\tilde{f}_{tk}^0$.

Factor-augmented linear regression and autoregression: Stock and Watson (2002), Bernanke, Boivin and Eliasz, (2005) Bai and Ng (2006), Pesaran, Pick and Timmermann (2011) and Cheng and Hansen (2015).

Additive regression on principal components: Härdle and Tsybakov (1995).

Set a maximum number, say $r_{\max}$ (which is usually not too large), for the factors. Since the factors extracted from the eigenanalysis are orthogonal to each other, the over-extracted insignificant factors will be discarded in the PMAMAR step.

Select the first few eigenvectors (corresponding to the first few largest eigenvalues) of $\mathcal{Z}_T \mathcal{Z}_T^{\intercal}/(Tp_T)$ so that a pre-determined amount, say 95%, of the total variation is accounted for.

Other commonly-used selection criteria such as BIC can be found in Bai and Ng (2002) and Fan, Liao and Mincheva (2013).

The exogenous variables $\mathbf{Z}_t$ are generated via an approximate factor model:

$$\mathbf{Z}_t = \mathbf{B}\mathbf{f}_t + \mathbf{z}_t,$$

where the rows of the $p_T \times r$ loadings matrix $\mathbf{B}$ and the common factors $\mathbf{f}_t$, $t = 1, \ldots, T$, are independently generated from the multivariate $\mathrm{N}(\mathbf{0}, I_r)$ distribution, and the p_T-dimensional error terms $\mathbf{z}_t$, $t = 1, \ldots, T$, are independently drawn from $0.1\mathrm{N}(\mathbf{0}, I_{p_T})$.
We set $p_T = 30$ or 150, $r = 3$, and generate the response variable via

$$\begin{aligned} Y_t &= m_1(f_{t1}) + m_2(f_{t2}) + m_3(f_{t3}) + m_4(Y_{t-1}) \\ &\quad + m_5(Y_{t-2}) + m_6(Y_{t-3}) + \varepsilon_t, \end{aligned}$$

where f_{ti} is the i-th component of $\mathbf{f}_t$, $m_i(\cdot)$, $i = 1, \ldots, 6$, are the same as in Example 1, and ε_t, $t = 1, \ldots, T$, are independently drawn from the $\mathrm{N}(0, 0.7^2)$ distribution.
In this example, we choose the number of candidate lags of Y as $d_T = 10$.

When $p_T = 30$, the KSIS+PMAMAR outperforms all the other approaches (except the Oracle) in terms of estimation and prediction accuracy.
When p_T becomes larger than T, the PCA based approaches show their advantage in effective dimension reduction of the exogenous variables, which results in their lower EE and PE.
The PCA+PMAMAR has a lower EE but higher PE than the PCA+KSIS+PMAMAR. This is due to the fact that without the KSIS step the PCA+PMAMAR selects more false lags of Y.

Empirical application

We next apply the proposed semiparametric model averaging methods to forecast inflation in the UK. The data were collected from the Office for National Statistics (ONS) and the Bank of England (BoE) websites and included quarterly observations on CPI and some other economics variables over the period Q1 1997 to Q4 2013.

All the variables are seasonally adjusted. We use 53 predictor series measuring aggregate real activity and other economic indicators to forecast CPI. Given the possible time persistence of CPI, we also add its 4 lags as predictors.

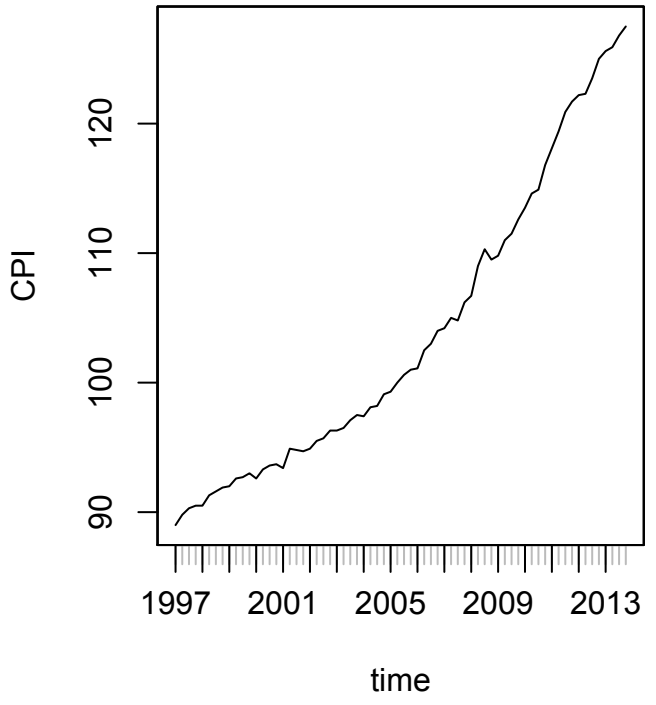
Data from Q1 1997 to Q4 2012 are used as the training set and those between Q1 2013 and Q4 2013 are used for forecasting.

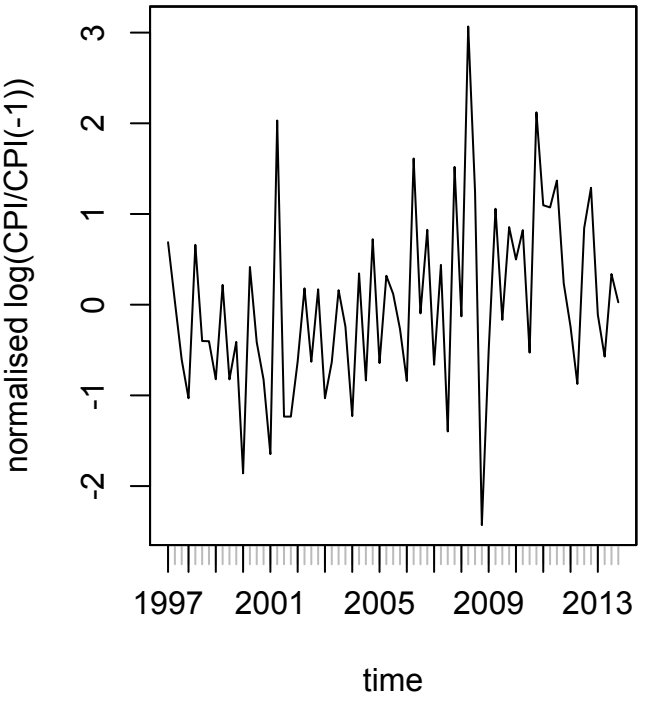
Method	IKSIS+PMAMAR	KSIS+PMAMAR	PCA+PMAMAR
PE	0.0360	0.1130	0.0787
Method	penGAM	ISIS	Phillips curve
PE	0.0865	0.3275	1.1900

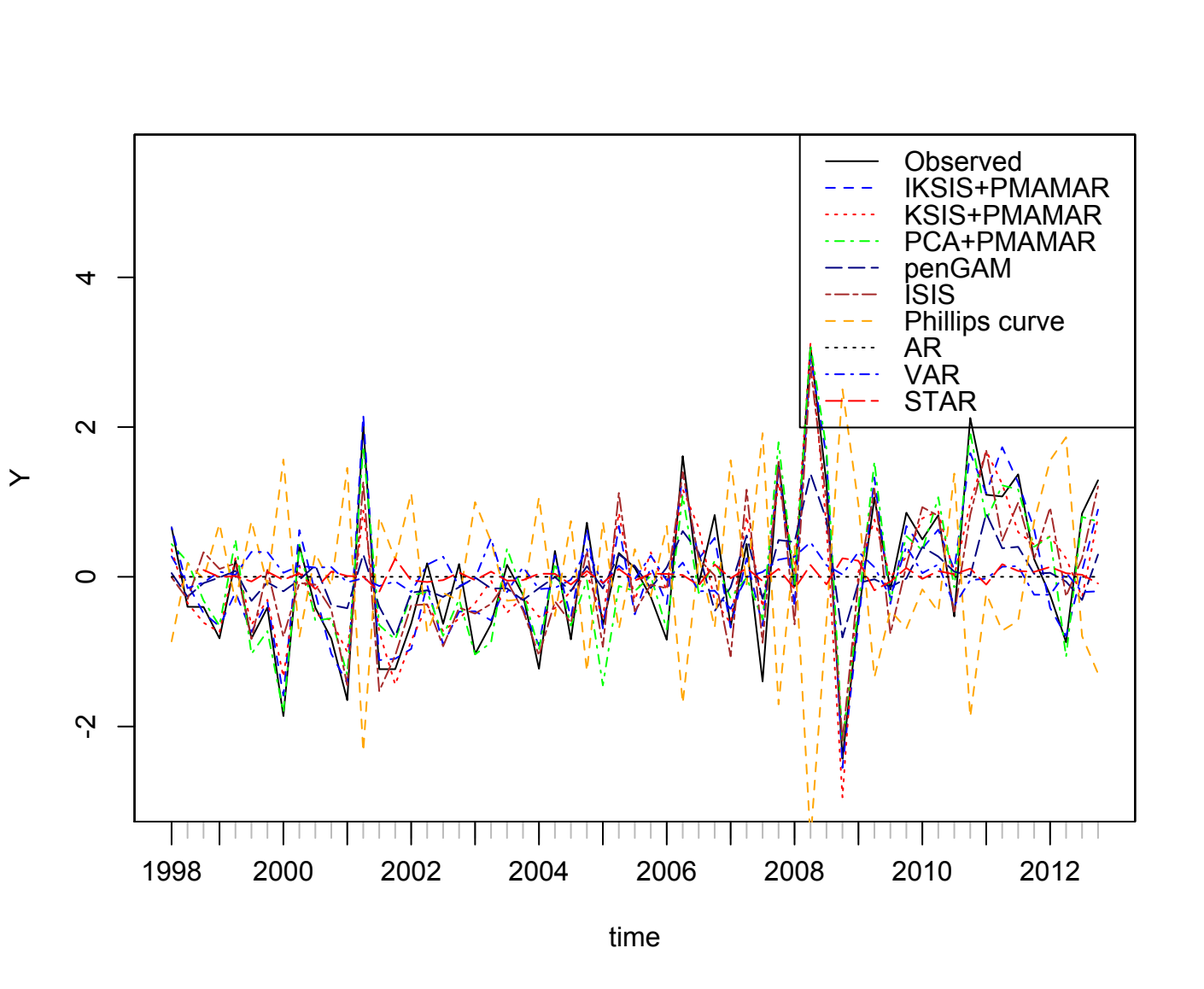
The Phillips curve specification is:

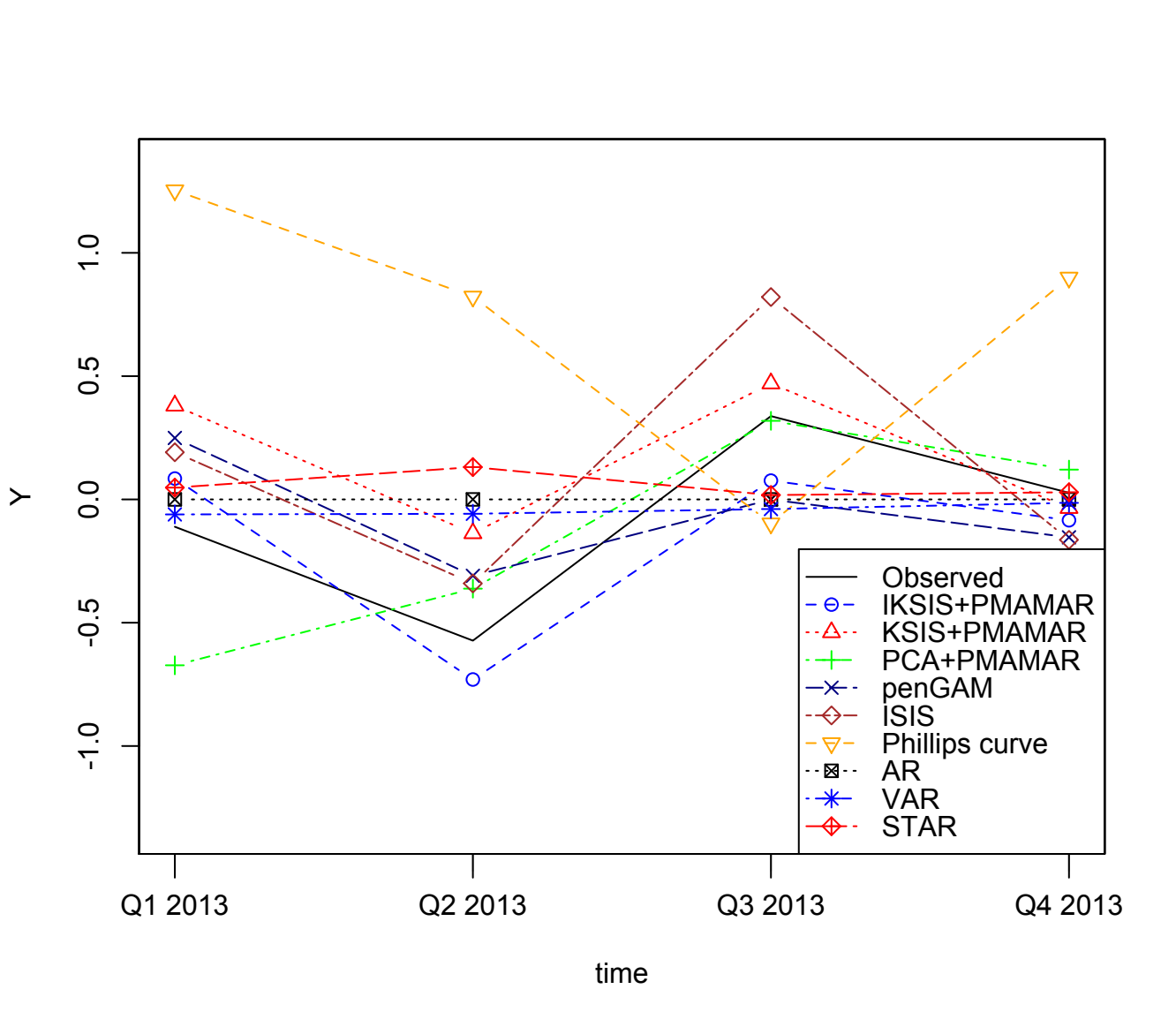
$$I_{t+1} - I_t = \alpha + \beta(L)U_t + \gamma(L)\Delta I_t + \varepsilon_{t+1},$$

where I_t is the CPI in the t-th quarter, $\beta(L) = \beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3$ and $\gamma(L) = \gamma_0 + \gamma_1 L + \gamma_2 L^2 + \gamma_3 L^3$ are lag polynomials with L being the lag operator, U_t is the unemployment rate, and Δ is the first difference operator.


CPI
90
100
110
120
1997
2001
2005
2009
2013
time


normalised log(CPI/CPI(-1))
-2
-1
0
1
2
3
1997
2001
2005
2009
2013
time


Observed
IKSIS+PMAMAR
KSIS+PMAMAR
PCA+PMAMAR
penGAM
ISIS
Phillips curve
AR
VAR
STAR
Y
4
2
0
-2
1998
2000
2002
2004
2006
2008
2010
2012
time


Y
1.0
0.5
0.0
-0.5
-1.0
Q1 2013
Q2 2013
Q3 2013
Q4 2013
time
Observed
IKSIS+PMAMAR
KSIS+PMAMAR
PCA+PMAMAR
penGAM
ISIS
Phillips curve
AR
VAR
STAR