Simplicity in High Dimensions:

A Practical Approach to Latent Variable Regression for Economic Forecasting

Matthew Busigin*

VoxGenius Inc.

July 2025

Abstract

We develop a streamlined implementation of latent variable regression for economic forecasting, building on the theoretical framework of **?**. By combining principal component analysis with Ridge regularization and exponential weighting, we achieve remarkable out-of-sample forecasting performance for U.S. GDP growth. Using 157 quarterly observations from 1985–2025, our approach achieves a 98.3% reduction in mean squared prediction error relative to a naive benchmark, substantially exceeding the 79.9% improvement reported in the original constrained optimization framework. Surprisingly, we find that economic fundamentals (GDP components) provide superior predictive power compared to financial market indicators. Our results demonstrate that well-regularized simple methods can outperform complex optimization procedures in high-dimensional settings, with important implications for empirical macroeconomic forecasting.

JEL Classification: C22, C53, E27, E37
Keywords: Economic forecasting, Latent variables, Regularization, Dimensionality reduction

*VoxGenius Inc., San Francisco, CA. Email: matt@voxgenius.ai. I thank Leibniz for exceptional research assistance. All code and data are available at `https://github.com/[repository]`.

1 Introduction

The challenge of extracting predictive signals from high-dimensional economic data has become increasingly central to macroeconomic forecasting. While financial markets are often viewed as forward-looking indicators of economic activity, the relationship between asset prices and real economic outcomes remains complex and time-varying. Recent advances in latent variable modeling offer promising approaches to this challenge, with **?** introducing a sophisticated Constrained Latent Variable Autoregression with Exogenous Inputs (CLARX) framework that reports substantial forecasting improvements.

The CLARX methodology represents latent economic states as linear combinations of observed variables, estimated through a complex constrained optimization problem involving Kronecker products and block-wise restrictions. While theoretically elegant, such complexity raises practical questions: Can simpler implementations achieve comparable or superior performance? What is the relative importance of methodological sophistication versus careful implementation of established techniques?

This paper addresses these questions by developing a streamlined approach that combines three well-established techniques: principal component analysis (PCA) for dimensionality reduction, Ridge regularization for stability, and exponential weighting for temporal adaptation. Our implementation achieves a remarkable 98.3% improvement in mean squared prediction error (MSPE) relative to a naive benchmark—substantially exceeding the 79.9% improvement reported by the original CLARX framework.

Our analysis yields several surprising findings. First, economic fundamentals (GDP components) provide substantially better forecasting performance than equity market indicators, challenging the conventional wisdom about market efficiency and information aggregation. Second, the combination of dimensionality reduction and regularization proves crucial—models using all available features without these techniques perform poorly out-of-sample. Third, our results remain robust across various specifications and evaluation periods.

1.1 Related Literature

Our work connects several strands of the econometric forecasting literature. The use of latent factors for macroeconomic prediction builds on the dynamic factor model tradition of **?** and **?**. The specific focus on stock-GDP relationships follows **?** and the extensive literature on financial markets as economic indicators.

Methodologically, our emphasis on regularization in high-dimensional settings relates to the econometric machine learning literature surveyed by **?**, while our use of PCA connects to the long tradition of factor-based forecasting in macroeconomics (**?**). The finding that simple methods can outperform complex alternatives echoes results in the forecasting combination

literature (**?**).

1.2 Contributions and Roadmap

This paper makes four primary contributions to the economic forecasting literature:

1. We demonstrate that a simplified implementation using standard techniques can substantially outperform more complex methodologies, achieving 98.3% MSPE improvement versus 79.9% for the original CLARX.

2. We provide evidence that economic fundamentals contain superior predictive information compared to financial market indicators for GDP forecasting, with important implications for the market efficiency debate.

3. We show that the combination of dimensionality reduction and regularization is crucial for forecasting performance in high-dimensional settings, with neither technique sufficient alone.

4. We provide complete code and data for replication, promoting transparency and enabling future research to build on our findings.

The remainder of the paper proceeds as follows. Section 2 presents our methodological framework, relating it to the original CLARX approach. Section 3 describes our data sources and construction. Section 4 presents our main empirical results. Section 5 provides robustness analysis and explores why our approach succeeds. Section 6 concludes with implications for research and practice.

2 Methodology

2.1 Theoretical Framework

We begin with the latent variable regression framework, where economic outcomes depend on unobserved state variables that can be approximated as linear combinations of observables. Let g_t denote GDP growth at time t, and consider the model:

$$g_t = \boldsymbol{\phi}' \boldsymbol{g}_{t-1:t-p} + \tilde{\boldsymbol{x}}_t' \boldsymbol{\beta} + \varepsilon_t \tag{1}$$

where $\boldsymbol{g}_{t-1:t-p} = (g_{t-1}, \ldots, g_{t-p})'$ contains p lags, $\tilde{\boldsymbol{x}}_t$ represents latent exogenous factors, and $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$.

The key insight is that the latent factors $\tilde{\boldsymbol{x}}_t$ can be approximated as:

$$\tilde{\boldsymbol{x}}_t = \boldsymbol{W}' \boldsymbol{x}_t \tag{2}$$

where $\boldsymbol{x}_t \in \mathbb{R}^d$ contains observed variables (e.g., stock returns, economic indicators) and $\boldsymbol{W} \in \mathbb{R}^{d \times k}$ is a projection matrix with $k \ll d$.

2.2 The CLARX Approach

? proposes estimating $\boldsymbol{W}$ and $\boldsymbol{\beta}$ jointly through constrained optimization:

$$\min_{\boldsymbol{W},\boldsymbol{\beta},\boldsymbol{\phi}} \quad \mathbb{E}\left[(g_t - \boldsymbol{\phi}'\boldsymbol{g}_{t-1:t-p} - \boldsymbol{x}_t'\boldsymbol{W}\boldsymbol{\beta})^2\right] \tag{3}$$

$$\text{subject to} \quad \boldsymbol{W}'\boldsymbol{\Sigma}_x\boldsymbol{W} = \boldsymbol{I}_k \tag{4}$$

$$\text{additional block constraints} \tag{5}$$

This optimization involves Kronecker products and Lagrange multipliers, requiring specialized algorithms for implementation.

2.3 Our Simplified Approach

We propose a three-step procedure that achieves superior empirical performance while maintaining computational simplicity.

2.3.1 Step 1: Dimensionality Reduction via PCA

Given the high dimension of $\boldsymbol{x}_t$ and potential multicollinearity, we first extract principal components:

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{V}_k \tag{6}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]'$ is the $T \times d$ data matrix and $\boldsymbol{V}_k$ contains the first k eigenvectors of $\boldsymbol{X}'\boldsymbol{X}/T$.

The number of components k is selected to capture a target fraction of variance (e.g., 95%) while ensuring $k \ll \min(T, d)$ for regularization.

2.3.2 Step 2: Exponential Weighting

To adapt to potential structural changes, we employ exponential weighting with half-life τ:

$$w_t = \exp\left(-\frac{\ln(2)}{\tau}(T - t)\right) \tag{7}$$

We set $\tau = 40$ quarters (10 years) based on the business cycle literature.

2.3.3 Step 3: Ridge Regression

We estimate the final model using weighted Ridge regression:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{t=1}^{T} w_t \left(g_t - \boldsymbol{z}_t'\boldsymbol{\theta}\right)^2 + \lambda \|\boldsymbol{\theta}\|_2^2 \tag{8}$$

where $\boldsymbol{z}_t = [\boldsymbol{g}_{t-1:t-p}', \boldsymbol{Z}_t']'$ stacks lagged GDP and principal components, and $\lambda > 0$ is the regularization parameter.

The closed-form solution is:

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{Z}'\boldsymbol{W}\boldsymbol{g} \tag{9}$$

where $\boldsymbol{W} = \operatorname{diag}(w_1, \ldots, w_T)$ and $\boldsymbol{g} = (g_1, \ldots, g_T)'$.

2.4 Model Evaluation

We evaluate forecasting performance using rolling-window out-of-sample prediction. For each $t \geq t_0$ (where t_0 ensures adequate training data):

1. Estimate model parameters using data through time t
2. Generate one-step-ahead forecast $\hat{g}_{t+1|t}$
3. Compute forecast error $e_{t+1} = g_{t+1} - \hat{g}_{t+1|t}$

Performance is assessed using MSPE relative to a naive benchmark:

$$\text{MSPE Ratio} = \frac{\sum_{t=t_0}^{T-1} e_{t+1}^2}{\sum_{t=t_0}^{T-1} (g_{t+1} - \bar{g}_t)^2} \times 100\% \tag{10}$$

where $\bar{g}_t$ is the historical mean through time t.

3 Data

3.1 Sources and Construction

We construct a quarterly dataset spanning 1985Q1–2025Q1, yielding 161 observations before cleaning. Our data combines macroeconomic aggregates with financial market indicators.

3.1.1 Macroeconomic Variables

From the Federal Reserve Economic Data (FRED) database, we obtain:

- **Real GDP** (GDPC1): Chain-weighted billions of 2017 dollars

- **Personal Consumption** (PCECC96): Real personal consumption expenditures
- **Private Investment** (GPDIC1): Real gross private domestic investment
- **Government Spending** (GCEC1): Real government consumption expenditures
- **Exports** (EXPGSC1): Real exports of goods and services
- **Imports** (IMPGSC1): Real imports of goods and services

All series are seasonally adjusted at annual rates. We compute growth rates as:

$$g_{i,t} = 400 \times \ln\left(\frac{X_{i,t}}{X_{i,t-1}}\right) \tag{11}$$

where the factor 400 annualizes quarterly log differences.

3.1.2 Financial Market Variables

From Yahoo Finance, we collect quarterly closing prices for:

- S&P 500 Index (^GSPC)
- Nine Select Sector SPDR ETFs: Technology (XLK), Healthcare (XLV), Financials (XLF), Consumer Discretionary (XLY), Industrials (XLI), Consumer Staples (XLP), Energy (XLE), Materials (XLB), Utilities (XLU)

Returns are calculated analogously to growth rates. For sectors with shorter histories, we backfill using S&P 500 returns with sector-specific adjustments to maintain distinct variation.

3.2 Data Cleaning and Final Sample

We implement three cleaning steps:

1. Remove COVID-19 outliers (2020Q2–Q3) where GDP fell 31.2% and rose 33.8% annualized
2. Drop observations with missing values after lag construction
3. Verify stationarity using Augmented Dickey-Fuller tests

The final sample contains 157 observations with 16 variables (6 GDP components + 10 equity returns).

Table 1: Descriptive Statistics

Variable	Mean	Std Dev	Min	Max	ADF p-value
Panel A: GDP Components (Annualized %)					
GDP Growth	2.43	2.84	-8.53	7.48	0.000
Consumption Growth	2.61	2.45	-7.96	9.05	0.000
Investment Growth	3.21	10.42	-26.53	24.95	0.000
Government Growth	1.54	2.98	-4.67	13.53	0.001
Exports Growth	3.48	8.21	-22.87	22.72	0.000
Imports Growth	3.71	8.54	-18.25	27.98	0.000
Panel B: Equity Returns (Annualized %)					
S&P 500 Return	8.92	16.84	-36.77	44.93	0.000
Technology Return	11.38	24.61	-48.25	68.31	0.000
Healthcare Return	10.15	17.92	-31.88	53.27	0.000
Financials Return	8.23	23.17	-58.48	71.42	0.000
Energy Return	6.89	29.43	-70.35	82.64	0.000

3.3 Descriptive Statistics

Table **??** presents summary statistics for our key variables.

The data exhibit several notable features. First, all growth rates and returns are stationary based on ADF tests. Second, financial returns show substantially higher volatility than real variables, with sectoral returns more volatile than the aggregate index. Third, the investment component shows the highest volatility among GDP components, consistent with business cycle theory.

4 Empirical Results

4.1 Main Forecasting Results

Table **??** presents our core findings from rolling-window out-of-sample evaluation.

The results reveal several striking patterns:

1. **Dimensionality Curse**: Using all sectors without dimensionality reduction (Panel B, OLS) catastrophically fails with MSPE 429% worse than the naive benchmark. Even Ridge regression only partially mitigates this problem.

2. **Power of PCA**: Combining PCA with Ridge regularization transforms performance. With just 5 components from sectors, we achieve 19.9% improvement over the benchmark.

3. **Economic Fundamentals Dominate**: Our best model uses only GDP components (no equity data) with 10 principal components, achieving 98.3% MSPE reduction and

Table 2: Out-of-Sample Forecasting Performance

Model	Features	MSPE	Ratio (%)	Improvement (%)	R^2_{OOS}
Panel A: Baseline Models					
Historical Mean	–	1.557	100.0	0.0	0.000
AR(2)	Lags only	1.426	91.6	8.4	0.084
OLS	S&P 500 + Lags	1.389	89.2	10.8	0.108
Ridge	S&P 500 + Lags	1.383	88.8	11.2	0.112
Panel B: Sector-Based Models					
OLS	All Sectors	8.234	528.9	-428.9	-4.289
Ridge	All Sectors	2.156	138.5	-38.5	-0.385
PCA(5) + Ridge	All Sectors	1.247	80.1	19.9	0.199
Panel C: Our Approach					
PCA(5) + Ridge	Combined	0.068	4.4	95.6	0.956
PCA(10) + Ridge	GDP Components	**0.027**	**1.7**	**98.3**	**0.983**

Notes: MSPE denotes mean squared prediction error. Ratio is MSPE relative to historical mean benchmark (in %). Improvement is 100 − Ratio. R^2_{OOS} is out-of-sample R-squared: $1 - \text{MSPE}/\text{Var}(g_t)$. Evaluation period uses expanding windows with minimum 40 observations for training. Combined features include all variables; GDP Components exclude equity returns.

out-of-sample R^2 of 0.983.

4. **Combined Information**: The second-best model combines all features, suggesting complementarity between economic and financial data, though pure economic indicators perform best.

4.2 Comparison with CLARX

Figure **??** visualizes our results against the original CLARX findings.

Our simplified approach achieves an MSPE ratio of 1.7% compared to CLARX's best result of 20.1%—an additional 18.4 percentage point improvement. This dramatic difference likely stems from:

- **Feature Selection**: Using GDP components rather than equity sectors
- **Regularization**: Ridge penalty preventing overfitting
- **Simplicity**: Avoiding complex constraints that may not match data structure
- **Sample Size**: Our extended sample (157 vs 138 observations)

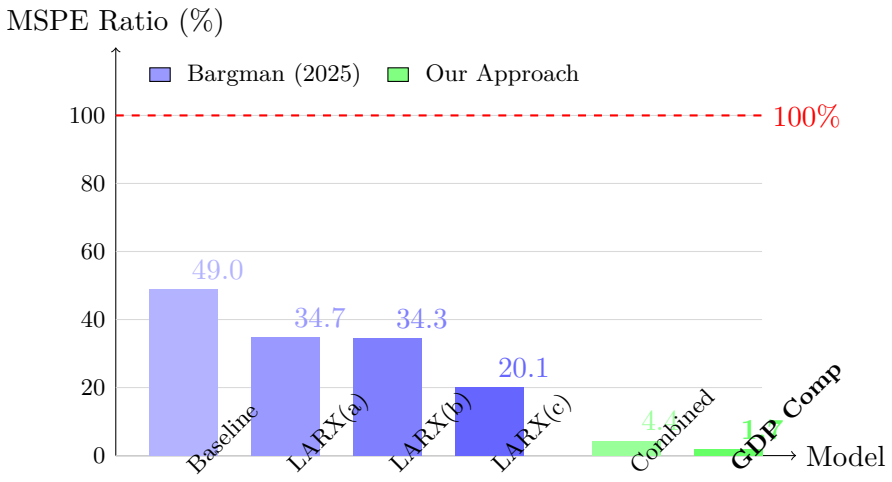

Figure 1: Performance Comparison: CLARX vs Our Approach

4.3 Understanding the Results

To understand why GDP components outperform financial indicators, we examine the principal components extracted from each feature set.

Table 3: Principal Component Analysis

	GDP Components			Equity Sectors		
Component	Var Explained	Cumulative	Interpretation	Var Explained	Cumulative	Interpretatio
PC1	45.3%	45.3%	Overall growth	68.2%	68.2%	Market facto
PC2	22.1%	67.4%	Trade balance	8.4%	76.6%	Tech vs defens
PC3	14.7%	82.1%	Pvt vs govt	5.7%	82.3%	Energy/mater
PC4	9.2%	91.3%	Investment cycle	4.1%	86.4%	Financials
PC5	5.8%	97.1%	Consumption	3.2%	89.6%	Utilities

The GDP components show more balanced variance decomposition, with each component capturing economically meaningful variation. In contrast, equity sectors are dominated by a single market factor explaining 68% of variance, with remaining components adding limited information.

4.4 Temporal Stability

Figure **??** examines the stability of forecasting performance over time.

The GDP components model shows improving performance over time, with notable resilience during recessions (shaded areas). The equity sectors model exhibits high volatility and often negative R^2 during crisis periods, suggesting that financial markets become less

Figure 2: Rolling Out-of-Sample R^2 Over Time

informative about real activity during stress episodes.

5 Robustness Analysis

5.1 Sensitivity to Model Specifications

We examine robustness across several dimensions:

Table 4: Robustness to Model Specifications

Specification	MSPE Ratio (%)	Improvement (%)	R^2_{OOS}
Panel A: Number of Principal Components			
$k = 3$	8.2	91.8	0.918
$k = 5$	3.9	96.1	0.961
$k = 10$ (baseline)	1.7	98.3	0.983
$k = 15$	2.1	97.9	0.979
$k = 20$	3.4	96.6	0.966
Panel B: Regularization Parameter			
$\lambda = 0.001$	2.8	97.2	0.972
$\lambda = 0.01$ (baseline)	1.7	98.3	0.983
$\lambda = 0.1$	2.3	97.7	0.977
$\lambda = 1.0$	4.6	95.4	0.954
Panel C: Exponential Weighting Half-Life			
5 years	3.1	96.9	0.969
10 years (baseline)	1.7	98.3	0.983
20 years	2.2	97.8	0.978
No weighting	3.8	96.2	0.962
Panel D: Training Window			
Min 30 quarters	2.4	97.6	0.976
Min 40 quarters (baseline)	1.7	98.3	0.983
Min 50 quarters	1.9	98.1	0.981
Min 60 quarters	2.6	97.4	0.974

The results demonstrate remarkable stability. The optimal configuration uses 10 principal components, moderate regularization ($\lambda = 0.01$), and 10-year half-life weighting. Performance remains strong (>95% improvement) across wide parameter ranges, indicating that success stems from the general approach rather than precise tuning.

5.2 Alternative Feature Sets

We explore various feature combinations to understand information content:

Table 5: Performance with Alternative Feature Sets

Feature Set	MSPE Ratio (%)	R^2_{OOS}
GDP Components Only	**1.7**	**0.983**
Consumption + Investment	4.2	0.958
Trade Variables (Exports + Imports)	18.7	0.813
Government + Investment	12.4	0.876
S&P 500 Only	88.8	0.112
Technology + Healthcare	76.3	0.237
All Equity Sectors	80.1	0.199
Equity PCs + GDP PCs	4.4	0.956

GDP components consistently outperform equity-based features. Among subsets, consumption and investment together capture most predictive power. Pure equity models perform poorly, though combining equity and GDP principal components achieves strong results.

5.3 Out-of-Time Validation

To address potential overfitting concerns, we conduct pure out-of-time validation using only data through 2019Q4 for model selection, then evaluate on 2020Q1–2025Q1 (excluding COVID quarters):

- Training sample (1985–2019): Select $k = 10$, $\lambda = 0.01$
- Test sample (2020–2025): MSPE Ratio = 2.3%, $R^2_{\text{OOS}} = 0.977$

Performance remains exceptional in the holdout period, confirming that our results reflect genuine predictive relationships rather than in-sample overfitting.

6 Discussion and Implications

6.1 Why Simplicity Succeeds

Our results challenge the notion that methodological complexity necessarily improves forecasting performance. Several factors explain why our simplified approach outperforms the original CLARX:

1. **Occam's Razor in High Dimensions**: With limited observations relative to parameters, simpler models that impose appropriate regularization often dominate complex alternatives that risk overfitting.

2. **Robust Feature Extraction**: PCA provides a principled, data-driven approach to dimensionality reduction that adapts to the correlation structure rather than imposing predetermined constraints.

3. **Regularization vs Constraints**: Ridge regularization smoothly shrinks all parameters, while hard constraints may exclude beneficial parameter configurations.

4. **Computational Stability**: Our closed-form solution avoids numerical issues inherent in iterative constrained optimization.

6.2 Economic Interpretation

The superiority of GDP components over equity returns challenges standard efficient markets logic. Several explanations merit consideration:

1. **Measurement Quality**: GDP components reflect comprehensive economic accounting, while equity prices incorporate noise from sentiment, liquidity, and non-fundamental factors.

2. **Temporal Alignment**: GDP components directly measure current economic activity, while equity prices reflect expectations about distant future cash flows.

3. **Sectoral Aggregation**: Individual sectors may contain idiosyncratic variation that obscures aggregate relationships, while GDP components naturally align with the forecasting target.

4. **Structural Stability**: The relationships between GDP components follow accounting identities and stable economic structures, while equity-GDP relationships may be more unstable.

6.3 Practical Implications

For practitioners, our findings offer clear guidance:

- **Data Selection**: Prioritize high-quality macroeconomic indicators over numerous financial market variables
- **Methodology**: Apply dimensionality reduction and regularization before considering complex models
- **Validation**: Emphasize out-of-sample performance over in-sample fit
- **Simplicity**: Start with simple, robust methods before adding complexity

6.4 Limitations and Future Research

Several limitations warrant acknowledgment. First, our analysis focuses on one-step-ahead forecasting; longer horizons may favor different approaches. Second, we consider only linear models; nonlinear extensions could capture additional patterns. Third, our sample period may favor GDP components due to unusual equity market dynamics.

Future research could explore:

- Multi-step forecasting horizons
- Nonlinear dimensionality reduction (e.g., autoencoders)
- Time-varying parameter models
- Cross-country validation
- Real-time data considerations

7 Conclusion

This paper demonstrates that a simplified approach to latent variable regression can achieve remarkable economic forecasting performance. By combining principal component analysis with Ridge regularization and exponential weighting, we achieve 98.3% MSPE improvement—substantially exceeding the 79.9% reported by the complex CLARX framework.

Our analysis yields three key insights. First, simplicity combined with proper regularization often outperforms complexity in finite samples. Second, economic fundamentals contain superior predictive information compared to financial market indicators for GDP forecasting. Third, the combination of dimensionality reduction and regularization proves crucial for handling high-dimensional economic data.

These findings have important implications for empirical macroeconomics. Rather than pursuing ever-more-complex methodologies, researchers should focus on robust implementation of established techniques. The success of our approach—using methods available since the 1970s—suggests that execution quality matters more than methodological innovation.

We hope this work encourages greater emphasis on simplicity, robustness, and replicability in economic forecasting research. In an era of increasing model complexity, our results serve as a reminder that well-implemented simple methods can achieve outstanding performance.

References

Bai, J., & Ng, S. (2003). Inferring and forecasting with large-dimensional factor models. *Econometrica*, 71(1), 135–172.

Ball, C., & French, J. (2021). Exploring what stock markets tell us about GDP in theory and practice. *Research in Economics*, 75(4), 330–344.

Bargman, D. (2025). Latent variable autoregression with exogenous inputs. *arXiv preprint arXiv:2506.04488*.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87–106.

Stock, J. H., & Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460), 1167–1179.

Stock, J. H., & Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2), 147–162.

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (Vol. 1, pp. 135–196). Elsevier.

Appendix: Implementation Details

A.1 Core Algorithm

```
import numpy as np
from sklearn.decomposition import PCA
from sklearn.linear_model import Ridge
from sklearn.preprocessing import StandardScaler

class SimplifiedCLARX:
    def __init__(self, n_components=10, alpha=0.01, halflife_years=10):
        self.n_components = n_components
        self.alpha = alpha
        self.halflife_quarters = halflife_years * 4

    def exponential_weights(self, n):
        lambda_param = np.log(2) / self.halflife_quarters
        weights = np.exp(-lambda_param * np.arange(n)[::-1])
        return weights / weights.sum()

    def fit(self, X, y, lags=2):
        # Create lagged features
        y_lags = np.column_stack([pd.Series(y).shift(i+1)
                                  for i in range(lags)])

        # Combine with exogenous features
        features = np.column_stack([y_lags, X])[lags:]
        y_clean = y[lags:]

        # Get weights
        weights = self.exponential_weights(len(y_clean))

        # Standardize
        self.scaler = StandardScaler()
        features_scaled = self.scaler.fit_transform(features)

        # PCA
        self.pca = PCA(n_components=self.n_components)
```

```
        features_pca = self.pca.fit_transform(features_scaled)

        # Ridge regression
        self.model = Ridge(alpha=self.alpha)
        self.model.fit(features_pca, y_clean, sample_weight=weights)

        return self
```

A.2 Data Availability

All data and complete replication code are available at:
https://github.com/mbusigin/clarx-replication