# Ecce Signum: An R Software Package for Analyzing Multivariate Time Series

Tucker McElroy[1] James Livsey[2]

[1]US Census Bureau

[2]US Census Bureau

## Abstract

The package provides multivariate time series models for structural analysis, allowing one to extract latent signals such as trends or seasonality. Models are fitted using maximum likelihood estimation, allowing for non-stationarity, fixed regression effects, and ragged-edge missing values. Extracted signals are produced with uncertainty measures that account for sample edge effects and missing values, and the signals (as well as the original time series) can be forecasted.

## Disclaimer

This presentation is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the author and not those of the U.S. Census Bureau. All time series analyzed in this presentation are from public or external data sources.

# Outline

- Introduction
- Stages of Analysis
- Non-Defense Capitalization
- Housing Starts

# Introduction

## What is Ecce Signum?

- *Ecce Signum* (Latin for "Behold the Proof") is an R package for analyzing multiple time series.
- The original intent was to do multivariate signal extraction (i.e., estimate trends, compute seasonal adjustments, etc.).
- More models and methodology have been added, allowing forecasting, missing values, extreme-value adjustment, and more.

# Introduction

## What Type of Data?

Time series data with some or all of the following attributes can be analyzed with *Ecce Signum*:

- Multivariate time series, or univariate time series with embedding multivariate structure
- Fixed effects, such as holiday patterns
- Aberrancies, such as additive outliers
- Complex dynamics, including trend, seasonality, and cycles
- Seasonal cycles of diverse period
- Linked dynamics, such as common trends and common seasonality
- Ragged edge missing values
- Changes in sampling frequency

# Introduction

## What Types of Analyses?

We have in mind several possible objectives of analysis, which the methodological tools can address:

- Modeling of the data via reduction to a maximum entropy residual
- Identification and estimation of fixed effects
- Linearization of the data via maximum entropy transformations
- Extraction of underlying signals, which may be common, such as trends, seasonals, and cycles
- Imputation of missing values
- Forecasting and aftcasting
- Quantification of uncertainty in extractions

# Introduction

## Why Ecce Signum?

- The modeling philosophy is informed by the concept of entropy.
- This allows us to combine extreme-value adjustment and missing values together with model comparison statistics.
- A unified paradigm for modeling time series is the ultimate goal.
- Efficient algorithms allow for models not embeddable in state space, and can handle ragged edge missing values.

# Introduction

## Framework

- The framework of *Ecce Signum* consists of an $N$-dimensional multivariate time series, which may be observed with missing values at various times for various components.
- After the possible application of a log transform, the model is written additively in terms of fixed effects (given through user-specified time series regressors) and latent stochastic processes (which the user specifies during model declaration).
- The class of models includes ARIMA and SARIMA, structural time series models (such as Local Level and Smooth Trend), VAR, and VARMA.
- After model fitting, the analyst can proceed to applications such as forecasting or signal extraction.

# Stages of Analysis

## Loading, Transforming, and Exploring

- First the data is enetered as a $T \times N$-dimensional matrix, involving $T$ time points and $N$ series.
- Missing values can be encoded with **NA**
- Metadata (start date, frequency, names, etc.) are entered by the user.
- A log transformation can be applied; one can select a restricted temporal range or subset of component series for analysis.
- Spectral density plots can be viewed to assist with model identification.

# Stages of Analysis

## Model Declaration

- Models can have stochastic effects and fixed effects.
- Fixed effects determine the mean of the time series $X_t$ through specified regressors $z_t$, one for each component series:

$$\mathbb{E}[X_t] = z_t' \beta.$$

- These can be holiday effects or outlier regressors.
- Latent processes are stochastic effects, and correspond to trend, seasonality, business cycle, etc.

# Stages of Analysis

## Model Declaration

The observed time series $X_t$ is modeled via

$$X_t = z_t' \beta + Y_t. \tag{1}$$

The stochastic process $\{Y_t\}$ has mean zero, and is modeled by

$$Y_t = \sum_{k=1}^{K} S_t^{(k)}, \tag{2}$$

where there are $K$ vector latent processes $\{S_t^{(k)}\}$ (for $1 \leq k \leq K$).

# Stages of Analysis

## Model Declaration

- Each latent process $S_t^{(k)}$ is added to the model by specifying a scalar differencing operator $\delta^{(k)}(z)$, the rank configuration for the driving white noise, and the time series model type (e.g., VARMA or SARMA, etc.).
- $\delta(z)$ defines the type of non-stationarity in the latent process.
- Rank configuration allows for common signals (i.e., collinearity).
- Holiday effects can be constructed with windows of activity around each date.

# Stages of Analysis

## Model Fitting

- Parameter estimates can be obtained by Method of Moments (MOM) and Maximum Likelihood (ML).
- Parameterization is always through pre-parameters in Euclidean space, that are homeomorphically mapped to the parameters, which have certain implicit constraints (e.g., variances are non-negative).
- Additional linear constraints on regression parameters can be applied (e.g., all the coefficients add up to zero, or one of the parameters equals zero).
- Additive outliers are handled by maximum entropy principle, replacing them by imputed values when this significantly improves the likelihood.

# Stages of Analysis

## Model Fitting

- MOM is crude but fast; these estimates can be used to initialize the ML routine.
- ML can handle ragged edge missing values (MOM cannot), and regression effects are estimated via GLS.
- ML can be quite slow when the parameter space is high-dimensional.
- Model fit can be assessed with Portmanteau and normality tests on residuals.
- Nested models can be compared using the likelihood ratio statistic.

# Stages of Analysis

## Applications of the Fitted Model

- Applications include casting (forecasting, aftcasting, and midcasting) and signal extraction.
- Midcasts (imputed missing values) are implicitly computed in the likelihood, and can be explicitly obtained with prediction uncertainty.
- Forecasts and aftcasts can also be obtained, and automatically incorporate midcast uncertainty.
- Fixed effects can also be casted and incorporated.

# Stages of Analysis

## Applications of the Fitted Model

- There are two approaches for signal extraction: Wiener-Kolmogorov (WK) filtering or ad hoc (AC) filtering. The former uses latent process modeling to construct signal extraction filters, while the latter applies a user-defined filter to *casted* data (i.e., missing values are imputed and sample boundaries are extended by forecasts and aftcasts).
- For both approaches the extraction uncertainty is quantified, taking account of midcasting.
- We can extract future or past time points of signals.
- We can extract linear combinations of signals (e.g., a growth rate, or the sum accross series).

# Stages of Analysis

## Applications of the Fitted Model

- For the WK approach, we can use direct matrix formulas (not compatible with missing values) or cast-extended data with the bi-infinite WK filter applied.
- Plots of casts and extractions, with uncertainty indicated, is a final step in the analysis.
- Figures of WK filter coefficients (and frequency response functions) can also be viewed.

# Non-Defense Capitalization

## The Data

- This illustration examines monthly Shipments and New Orders from the Manufacturing, Shipments, and Inventories survey. (This is seasonally adjusted monthly data covering the sample period January 1992 to April 2020, downloaded July 21, 2020 (4:45 PM), U.S. Census Bureau, obtained from https://www.census.gov/mtis/index.html by selecting Non-Defense Capital Goods, and either Value of Shipments or New Orders.)
- Call this **ndc** for short.
- The data for New Orders is not available at January 1992, since this series starts at February 1992; so this value is entered as an NA. This is an example of *ragged edge* data.

# Non-Defense Capitalization

## Analysis in Ecce Signum

- We apply Ecce Signum to analyze **ndc**.
- Details in R Markdown notebook, and the working paper.
- A VAR(7) model is fitted using MOM and ML; the missing value is accounted for in the likelihood calculation.
- 50 forecasts and aftcasts are generated, and a midcast for the missing value, with uncertainty measured by confidence intervals.

# Non-Defense Capitalization

## Casting Plot

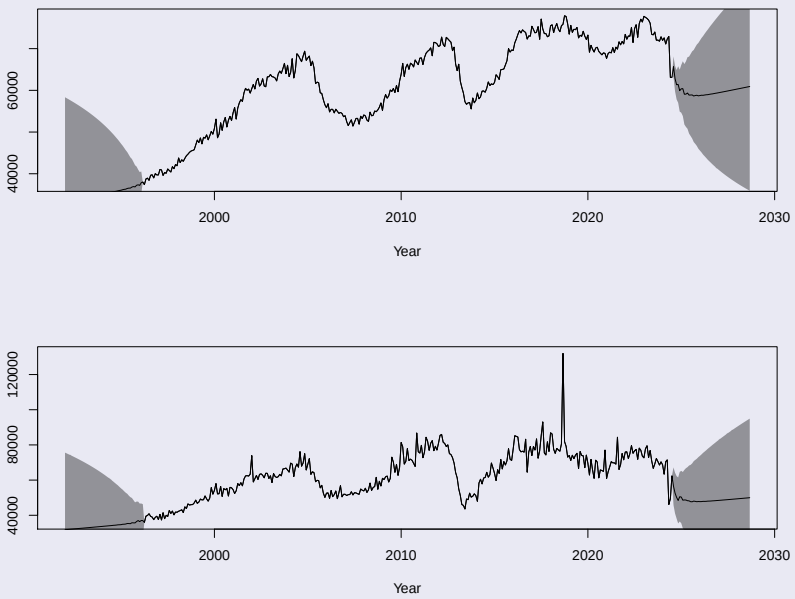


**Figure 1:** Casts of NDC

# Housing Starts

## The Data

- This illustration examines monthly New Residential Construction (1964-2012), Housing Units Started, Single Family Units. (The four series are from the Survey of Construction of the U.S. Census Bureau, available at http://www.census.gov/construction/nrc/how_the_data_are_collected/soc.html.)
- Call this *starts* for short.
- The data corresponds to the four regions of South, West, NorthEast, and MidWest.

# Housing Starts

## Analysis in Ecce Signum

- We apply Ecce Signum to analyze **starts**.
- Details in R Markdown notebook, and the working paper.
- This example illustrates modeling a multivariate series with 8 latent components, estimated with both MOM and ML.
- Extracted signals are either trend, seasonal, or non-seasonal.
- Two methods (direct matrix and truncated WK) of signal extraction are used.
- Also functions of a signal, such as growth rates, can be obtained; uncertainty is represented by shading.

# Housing Starts

## Signal Extractions Plot

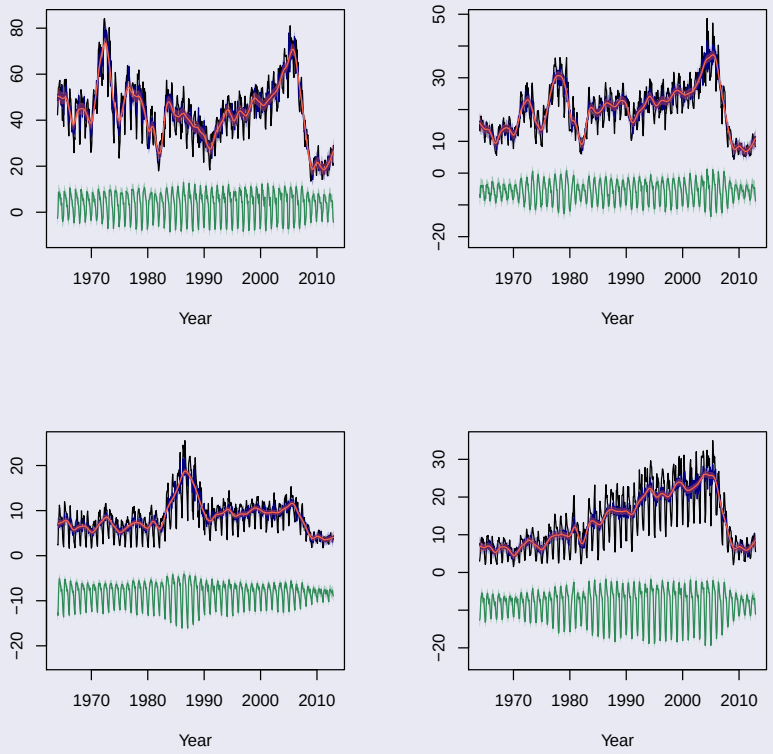


**Figure 2:** Signals of Starts

# Housing Starts

## Trend Growth Rates Plot

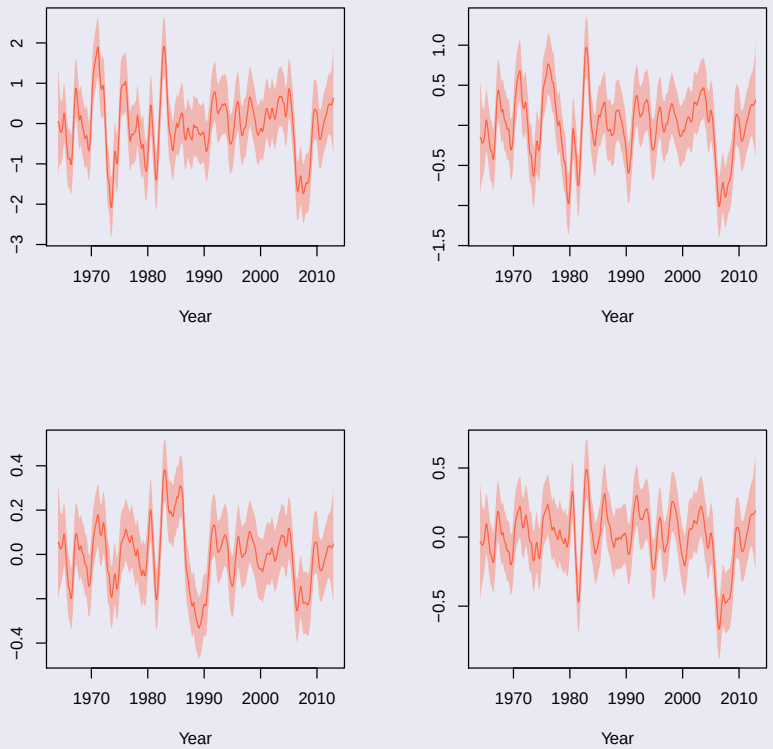


**Figure 3:** Starts Trend Growth Rate

# Conclusion

- *Ecce Signum* is a prototype and work-in-progress, available now as an R package.
- Ragged edge missing values, along with forecasting and signal extraction, can be handled.
- Future work: model identification, more general treatment of outliers (e.g., level shifts), sampling error, ...
- Reincarnation in Python?

# Contact


- email: tucker.s.mcelroy@census.gov, james.a.livsey@census.gov
- Github: https://github.com/tuckermcelroy