Model-Based and Semi-Parametric Estimation of Time Series Components and Mean Square Error of Estimators

Michael Sverchkov
(Bureau of Labor Statistics)


The opinions expressed in this paper are those of the author and do not necessarily represent the policies of the Bureau of Labor Statistics
We define seasonal and trend components under which the X-11 ARIMA estimators of them are almost unbiased at least in the central part of the series.

We define the Variance and Mean Square Error (MSE) of X-11 ARIMA and Basic State-Space Structural Model (BSM) estimators with respect to the newly defined trend and seasonal components and propose estimators for the Bias, Variance and MSE.

We investigate the behavior of the X-11 ARIMA and BSM estimators of the newly defined trend and seasonal components and their MSE estimators in a small simulation study based on real data.
Definitions

\[ y_t = G_t + e_t, \quad t = t_{\text{start}}, \ldots, 0, 1, \ldots, N, N + 1, \ldots, \infty; \]

\( y_t \) - target time series (observed for \( t = 1, \ldots, N \) only)

\( G_t \) - signal (unobserved), \( e_t \) - Error (unobserved)

Assumptions: \( e_t, G_t \) independent, \( E(e_t) = 0, \) \( \text{Var}(e_t) < \infty \)

Signal and Error can be defined in several ways. We consider the following alternative definitions:

I) Signal (\( G_t \)) = Trend (\( T_t \)) + Seasonal component (\( S_t \),

Error (\( e_t \)) = Irregular term (\( I_t \)) + Sampling error (\( \varepsilon_t \)) \,(Pfeffermann, 1994).

II) Signal (\( G_t \)) = Trend (\( T_t \)) + Seasonal component (\( S_t \)) + Irregular term (\( I_t \)),

Error (\( e_t \)) = Sampling error (\( \varepsilon_t \)) \,(Bell and Kramer 1999).

Components of interest

Seasonally Adjusted Series \( A_t = T_t + I_t \), and/or Trend \( T_t \).
X-11 ARIMA Estimators

X-11 ARIMA estimators of trend and seasonal components can be approximated as:

\[ \hat{T}_t = \sum_{k=-(t-1)}^{N-t} w_{kt}^{T} y_{t+k}, \quad \hat{S}_t = \sum_{k=-(t-1)}^{N-t} w_{kt}^{S} y_{t+k} \implies \hat{A}_t = y_t - \hat{S}_t = \sum_k w_{kt}^{A} y_{t-k} \]

Filters \( w_{kt}^{S} \) and \( w_{kt}^{T} \) defined by the X-11 ARIMA program options and length of series, \( N \).

At center part of series, filters time-invariant & symmetric;

\[ w_{kt}^{T} = w_{k}^{T}, \quad w_{-k}^{T} = w_{k}^{T}, \quad a_T \leq t \leq N - a_T; \]
\[ w_{kt}^{S} = w_{k}^{S}, \quad w_{-k}^{S} = w_{k}^{S}, \quad a_S \leq t \leq N - a_S. \]
X-11 ARIMA Estimators (cont.)

Notice: X-11 ARIMA contains “non-linear” operations:
Identification and estimation of ARIMA models,
Identification and gradual replacement of extreme observations.

We assume:
Time series under consideration already corrected for outliers,
Effects of identification and estimation of ARIMA models are negligible (verified in previous studies).
What does X-11 ARIMA estimate?

Model implies: 

\[ E(\hat{S}_t \mid G) = \sum_{k=-(t-1)}^{N-t} w_{kt} G_{t-k} = \tilde{S}_t, \]

where \( G = (G_{t_{\text{start}}}, G_{t_{\text{start}}+1}, \ldots) \).

(Similarly for trend and seasonally adjusted components.)

\[ \Rightarrow \hat{S}_t - \text{unbiased estimator of } \tilde{S}_t \text{ but } \tilde{S}_t \text{ not time-invariant. For example, } \tilde{S}_N \text{ based on } y_1, \ldots, y_N \text{ is not equal to } \tilde{S}_N \text{ based on } y_1, \ldots, y_N, y_{N+1}, \ldots, y_{2N}. \]

\[ \Rightarrow \tilde{S} = \{\tilde{S}_t, \ t = 0, \ldots, \infty\} \text{ not a seasonal component (not defined uniquely).} \]
**X-11 decomposition**

Assuming \( t_{start} << 0 \) define “X-11 trend and seasonal components” as:

\[
S^{x_{11}}_t = \sum_{k=-a_s}^{a_s} w^S_k G_{t-k}, \quad T^{x_{11}}_t = \sum_{k=-a_T}^{a_T} w^T_k G_{t-k}.
\]

(Bell & Kramer, 1999)

\[\Rightarrow \quad \text{X-11 decomposes the observed series into the “X-11-components”}
\]

\[y_t = T^{x_{11}}_t + S^{x_{11}}_t + e^{x_{11}}_t; \quad e^{x_{11}}_t = y_t - T^{x_{11}}_t - S^{x_{11}}_t
\]

\[\quad \text{X-11 estimators almost unbiased at the center part of the series with respect to this decomposition.}\]
X-11 decomposition (cont.)

X-11 ARIMA in common use all over the world. Users trust the method. Important to study what is estimated.

Estimator defines the parameter; not parameter defines the estimator.

Conditioning on signal $\Rightarrow$ viewing trend & seasonal effects as fixed population parameters $\Rightarrow$ conforms with classical sampling theory.

Parameters must be defined by symmetric filters.
Conditional Bias, variance and MSE of X-11 ARIMA estimators

\[ \text{Bias}(\hat{T}_t | G) = E[(\hat{T}_t - T_{t}^{X_{11}}) | G] = \sum_{k=-(t-1)}^{N-t} w_{kt}^{T} G_{t+k} - \sum_{k=-a_{t}}^{a_{t}} w_{k}^{T} G_{t+k}, \]

\[ \text{Var}[\hat{T}_t | G] = E[ \sum_{k=-(t-1)}^{N-t} w_{kt}^{T} (y_{t+k} - G_{t+k})]^{2} = E( \sum_{k=-(t-1)}^{N-t} w_{kt}^{T} e_{t+k})^{2}, \]

\[ \text{MSE}(\hat{T}_t | G) = E[(\hat{T}_t - T_{t}^{X_{11}})^2 | G] = \text{Var}(\hat{T}_t | G) + \text{Bias}^2(\hat{T}_t | G). \]

\[ \text{MSE}[\hat{S}_t] \text{ can be estimated by estimating } \text{Var}[\hat{S}_t | G] \text{ and } \text{Bias}[\hat{S}_t | G] \]

Similar expressions for any estimator \( \tilde{H}_t \approx \sum_{k=-(t-1)}^{N-t} h_{kt} y_{t+k} \) with arbitrary weights \( \{h_{kt}\} \), e.g., Basic Structural Model.
Bell and Kramer (1999)

BK propose same target components. They estimate the components by augmenting the series with sufficient minimum mean squared error forecasts and backcasts under the ARIMA model, such that the symmetric filters can be applied to the augmented series at every time point with observation.

\[
\hat{T}_t^{BK} = \sum_{k=-a_t}^{a_T} w_k^T y_{t+k}; \quad y_{t+k}^* = y_{t+k} \text{ if } y_{t+k} \text{ observed}, \quad y_{t+k}^* = \text{forecast or backcast otherwise.}
\]

\[
E(y_{t+k}^* - y_{t+k}) = 0 \text{ (unconditionally)} \Rightarrow E(\hat{T}_t^{BK} - T_t^{X_{11}}) = E[ \sum_{k=-a_T}^{a_S} w_k^T y_{t+k}^* - \sum_{k=-a_T}^{a_S} w_k^T y_{t+k} ] = 0
\]

(unconditionally).

BK estimate \( \text{Var}(\hat{T}_t^{BK} - T_t^{X_{11}}) \) over distributions of sampling errors and forecast & backcast prediction errors.

Time series irregulars considered as part of the signal.
Where do we differ?

We **condition** on $G = \{G_t, t = t_{\text{start}}, \ldots, \infty\}$, in which case in general $E[\hat{T}_t^{BK} - T_t] \neq 0$.

Bias may also exist even **unconditionally** when extrapolating less than required for use of symmetric filters, depending on distribution of the signal.

Estimation of **conditional MSE** not restricted to full forecasts and backcasts, and can be applied when estimating the target components with only **one** or **two** years of forecasts and backcasts (common case?) or without ARIMA extrapolations, or when estimating the components by fitting different models (**e.g.,** Basic Structural Model).
Where do we differ? (cont.)

We attempt to estimate the **conditional MSE**, given the signal.

Alternatively, when the bias estimator is obtained optimally under a model, it may be viewed as estimating the **unconditional bias** over all possible realizations of the signal under the model, given the observed series (see below).

The proposed approach is applicable also when the signal consists of only the trend and the seasonal effect, and the time series irregular is part of the error.
Estimation of conditional MSE

Conditional variance

When the time series irregular is part of the signal,

\[
\hat{\text{Var}}(\hat{T}_t | G) = \hat{E}( \sum_{k=-(t-1)}^{N-t} w_{kt}^T \varepsilon_{t+k} )^2 = \sum_k \sum_l w_{kt}^T w_{lt}^T \hat{Cov}(\varepsilon_{t+k}, \varepsilon_{t+l}).
\]

Easily computable when estimators of the variances and covariances of the sampling errors are available.

When the irregular is not part of the signal, \( G_t = S_t + T_t, e_t = I_t + \varepsilon ; \)

\[
\hat{\text{Var}}(\hat{T}_t | G) = \sum_k \sum_l w_{kt}^T w_{lt}^T \hat{Cov}(e_{t+k}, e_{t+l}).
\]

Methods for estimating \( \text{Cov}(e_{t+k}, e_{t+l}) \) considered in literature

(Pfeffermann 1994, Pfeffermann & Scott (1997), Chen et al. (2003))
Idea of estimating $\text{Cov}(e_{t+k}, e_{t+l})$:

$$\text{Cov}(\hat{R}_t, \hat{R}_m | G) = \text{Cov}[\sum_{k=-(t-1)}^{N-t} w_{kt}^R e_{t+k}, \sum_{l=-(m-1)}^{N-m} w_{lm}^R e_{m+l}] = \sum_{k} \sum_{l} w_{kt}^R w_{lm}^R \text{Cov}(e_{t+k}, e_{m+l}) \quad (1)$$

$$w_{0t}^R = 1 - w_{0t}^S - w_{0t}^T, \quad w_{kt}^R = 1 - w_{kt}^S - w_{kt}^T, \quad k \neq 0.$$  

$$\hat{R}_t = y_t - \hat{S}_t - \hat{T}_t = \sum_{k=-(t-1)}^{N-t} w_{kt}^R y_{t+k} - \text{observed,}$$

$$\text{Cov}(\hat{R}_t, \hat{R}_m | G) - \text{can be estimated}$$

Assuming $I_t$ - stationary, $\text{Cov}(I_{t+k}, I_{t+l}) = 0, |k-l| > \text{Const}$ and $\text{Cov}(\varepsilon_{t+k}, \varepsilon_{t+l})$ are available

or

assuming $e_t$ - stationary and $\text{Cov}(e_{t+k}, e_{t+l}) = 0, |k-l| > \text{Const}$

(1) can be solved
Estimation of conditional bias

Within X-11 ARIMA framework, the signal can be estimated conveniently (but not efficiently) as:

(a) Use the model chosen by program to forecast–backcast \( m = \max(a_T, a_S) \) observations. Let \( N_{\text{aug}} = N + 2m \).

(b) Estimate \( \hat{G}_{t}^{\text{aug}} = \sum_{k=-(t-1)}^{N_{\text{aug}}-t} w_{kt} S_{t+k} + \sum_{k=-(t-1)}^{N_{\text{aug}}-t} w_{kt} T_{t+k} \); \( t = -m + 1, \ldots, N + m \)

\( y_{t}^{\text{aug}} = y_{t} \) if \( y_{t} \) observed, \( y_{t}^{\text{aug}} = \) forecast (backcast) otherwise.

\[ \text{Biàs} [\hat{T}_t | G] = \hat{E}[(\hat{T}_t - T_t^{X11}) | G] = \sum_{k=-(t-1)}^{N-t} w_{kt} \hat{G}_{t+k}^{\text{aug}} - \sum_{k=-a_T}^{a_T} w_{kt} \hat{G}_{t+k}^{\text{aug}} ; \ t = 1, \ldots, N. \]
Estimation of conditional bias (cont.)

Signal can be estimated more *efficiently* by extracting the models for the trend and seasonal effects using *signal extraction*, and then estimate within the observation period, and forecast and backcast under the extracted models.

Estimators of components are in this case *minimum MSE* (MMSE) under the models.

**MMSE** estimator is the conditional expectation given the observed series. The bias estimator is then the *unconditional expectation* of the bias over all possible realizations of the signal *given the observed series*.

When predicting the signal many time points ahead, estimators may perform *very badly*.

The filter weights $\downarrow 0$ fast when moving away from the time point of interest $\Rightarrow$ possibly large biases of estimators of the signal for distant time points may have little effect on the bias of the bias estimator.
Figure 1. Central and end weights when estimating the trend by default X-11 and under Basic Structural Model.
Estimation of conditional MSE

\[ \text{MSE}(\hat{T}_t \mid G) = \text{Var}(\hat{T}_t \mid G) + \text{Biås}^2 (\hat{T}_t \mid G). \]

Generally, over-estimator because,

\[ E[\text{Biås}^2 (\hat{T}_t \mid G) \mid G] \]
\[ = \{E[\text{Biås}(\hat{T}_t \mid G)] \mid G\}^2 + \text{Var}[\text{Biås}(\hat{T}_t \mid G) \mid G] \geq \{E[\text{Biås}(\hat{T}_t \mid G)] \mid G\}^2. \]

Overestimation corrected by subtracting \( \text{Var}[\text{Biås}(\hat{T}_t \mid G) \mid G] \).

\( \text{Biås}(\hat{T}_t \mid G) \) is again a linear combination of the observed values so the variance is estimated similarly to the estimation of the variance of the estimators of the components.
**Simulation study**

Simulate series from model fitted to the series “Employment to Population Ratio in DC”.

**Erratic series**: X-11 ARIMA residual component explains 55% of month to month changes and 32% of yearly changes. Major portion of the residual is **sampling error**.
Figure 2. Employment to Population Ratio in DC, 1999-2000. Original series and trends estimated by X-11 ARIMA with 12 forecasts and under BSM.

Original series; Trend X11 ARIMA; Trend BSM
Basic Structural Model (BSM) used for simulation

\[ y_t = Y_t + \varepsilon_t = T_t + S_t + I_t + \varepsilon_t \]

\[ T_t = T_{t-1} + R_t; \quad R_t = R_{t-1} + \eta_{Rt}; \quad S_t = \sum_{j=1}^{6} S_{j,t}; \]

\[ S_{jt} = \cos \omega_j S_{j,t-1} + \sin \omega_j S^*_{j,t-1} + \eta_{Sjt} \]

\[ S^*_t = -\sin \omega_j S_{j,t-1} + \cos \omega_j S^*_{j,t-1} + \eta^*_S_{jt} \]

\[ \omega_j = 2\pi j/12, \quad j = 1,\ldots, 6 \]

\[ I_t, \eta_{Rt}, \eta_{Sj,t}, \eta^*_{Sj,t} \] are mutually independent normal disturbances.

\[ \varepsilon_t (\text{sam.err}) \sim \text{AR}(15) \] accounts for CPS sampling design.
**Simulation plan**

**S1**- Generate 1,000 series \( y_t^b, b = 1, \ldots, 1,000 \) of length 300 from BSM; 
\[
y_t^b = T_t^b + S_t^b + I_t^b + \varepsilon_t^b, \quad t = 1, \ldots, 300.
\]
For present study, \( G_t = T_t + S_t + I_t \).

Generate additional 1,000 series as: fix the signal of 2\(^{nd} \) series and add the sampling errors from the first 1,000 series to the fixed signal.

- **First set** of series illustrates *unconditional* properties.
- 2\(^{nd} \) set illustrates *conditional* properties, given signal.

**S2**- Compute default X-11 estimator of trend and seasonal component for each simulated signal to obtain the target X-11 components for central 180 months.

**S3**- Remove first and last 60 observations from simulated series and apply X-11 ARIMA with 12 and 60 forecasts, using default filters and ARIMA \((0,1,1),(0,1,1)\) model. (Model chosen by method for first 10 series in each of two data sets.)

**S4**- Fit the BSM for each of the series of length 180.
Figure 3. Empirical unconditional bias of X-11 ARIMA trend estimates with 60 forecasts and mean of bias estimates, signals estimated by X-11 ARIMA.
Figure 4. Empirical **conditional Bias** of X-11 ARIMA trend estimates with 60 forecasts and **mean of bias estimates**, signal estimated by X-11 ARIMA.
Figure 5. Empirical **unconditional bias** of BSM trend estimates and **mean of bias estimates**, signals estimated by X-11 ARIMA.
Figure 6. Empirical conditional bias of BSM trend estimates and mean of bias estimates, signal estimated by X-11 ARIMA.
Figure 7. Empirical unconditional RMSE of X-11 ARIMA trend estimates with 60 forecasts and mean of RMSE estimates, signals estimated by X-11 ARIMA.
Figure 8. Empirical conditional RMSE of X-11 ARIMA trend estimates with 60 forecasts and mean of RMSE estimates, signal estimated by X-11 ARIMA.
Figure 9. Empirical unconditional RMSE of BSM trend estimates and mean of RMSE estimates, signals generated from BSM and estimated by X-11 ARIMA.
Figure 10. Empirical conditional RMSE of BSM trend estimates and mean of RMSE estimates, signal estimated by X-11 ARIMA.
Figure 11. Empirical unconditional RMSE of X-11 ARIMA trend estimates with 12 forecasts (blue), 60 forecasts (red) and BSM (black), signals estimated by X-11 ARIMA.
Figure 12. Empirical conditional RMSE of X-11 ARIMA trend estimates with 12 forecasts (blue), 60 forecasts (red) and BSM (black), signal estimated by X-11 ARIMA.
Conclusions

- Method seems to work but need to experiment with many more simulated and real series,

- Investigate the robustness of the method to possible model misspecification

- Study efficient estimation of signal under appropriate models.

Thanks!!!  (Sverchkov.Michael@bls.gov)