# Bayesian Forecasting: Exponential Decay Method

Jiacheng Wang*[1] and Soudeep Deb[2]

[1]Department of Statistics, University of Chicago
[2]NBCUniversal
{*jiachengwang@uchicago.edu, Soudeep.Deb@nbcuni.com*}

August 15, 2019

## 1 Introduction

Exponential decay method (EDM) is a new way for time series modelling, which is more flexible compared to classic ARIMA/SARIMA model. The following table is an overview for comparison between ARIMA/SARIMA model and exponential decay method.

Table 1: Comparison between ARIMA/SARIMA and EDM

| What we want? | ARIMA/SARIMA | EDM |
|---|---|---|
| Time-varying covariates | ✓ | ✓ |
| Time-dependent structure | ✓ | ✓ |
| Change points | ✓ | ✓ |
| Missing/multiple data | ✗ | ✓ |
| Weekly-yearly seasonality | ✗ | ✓ |
| Variance flexibility | ✗ | ✓ |
| Fast implementation | ✗ | ✓ |

- Time-varying covariates
  Time series modelling requires to have time-varying covariates, which allows users to have time dependent linear structure for different time periods. Both ARIMA/SARIMA and EDM can achieve this.
- Time-dependent structure
  The most important feature for time series modelling is to capture the time dependent structure in the data. ARIMA/EDM both work for this. Section 3 introduces how basic EDM model incorporates time dependent structure in detail.
- Change points
  Change points are critical for time series modelling, which indicate possible structural changes for observations. ARIMA/EDM allow to add change points. Section 5.1 explains how to add change points and corresponding parameter estimation for EDM.

*Work as data science intern at NBCUniversal summer 2019. We would like to thank Marco Morales for helpful comments. This document is for internal use only, not for external distribution.

- Missing/multiple data
  One critical assumption for ARIMA/SARIMA model is that observations should have some "consecutive" structure. Suppose you have daily data, if there are some missing data in certain week or there are multiple observations on a specific day, the dependent structure breaks down for ARIMA/SARIMA model. In this situation, the EDM still works since covariance structure for EDM only depends on time difference among observations and changes together with data structure.

- Weekly-Yearly seasonality
  ARIMA/SARIMA only allows to add one seasonality into the model, either weekly or yearly while EDM can take both into consideration. See Section 5.2 for details.

- Variance flexibility
  ARIMA/SARIMA and EDM both have a variance parameter, which controls the fluctuate range for residual term. You will see in Section 5.3 that ARIMA/SARIMA model can only treat it as a constant, while EDM can change variance parameter as a function of calendar year. Thus, EDM model has a more flexible structure.

- Fast implementation
  In practice, ARIMA/SARIMA model takes a long time to find the optimal model if the dataset is large. Or even in the worse case, when the regressors are large, time will be longer for ARIMA/SARIMA to get the final result. While EDM takes the advantages of Bayesian framework to estimate the parameter(s), the algorithm will converge in a short time and the MAPE/SMAPE for final models are even better than ARIMA/SARIMA in most of the cases. Thus, EDM is a better choice for time series modelling if there is a time budget for forecasting.

As you will see, basic EDM model is pretty simple and we add some complicated structures into the basic model to achieve better prediction performance. We will try to improve the basic EDM model in the following aspects:

- Add change points
- Add weekly-yearly seasonality
- Add variance flexibility

The whole document's structure is as follows. Section 2 sets up the data structure and notations. Section 3 gives an introduction to basic EDM model. Section 4 provides the framework for parameter estimation in basic EDM model. Section 5 shows three updated versions for basic EDM model and the updated algorithm for parameter estimation. Section 6 implements the EDM model on some real dataset from NBCUniversal with different networks and daypart/show name and compares the performance between ARIMA/SARIMA and EDM. Section 7 summarizes all the things about EDM and future work for this method.

# 2 Data and Notations

Suppose we have time series observations

$$(y_i, \mathbf{x}_i)_{i=1}^N$$

where for observation $i$, $y_i \in \mathbb{R}$ is the response variable and $\mathbf{x_i} \in \mathbb{R}^p$ is the explanatory variable. The number for observations (sample size) is $N$.

The notations used in this document are as follows. $\mathbb{R}^p$ represents set for all p-dimensional real numbers. $\mathbb{1}(\cdot)$ represents the indicator function and *mod* denotes the modulo operator.

# 3 Basic EDM Model

We start from basic EDM model, which includes an extra term into the original linear regression model, residual $u_i$ with EDM covariance. Model structure for basic EDM model is:

$$y_i = \mathbf{x}_i^\top \beta + u_i + \varepsilon_i \tag{1}$$

Or write in matrix form:

$$\mathbf{Y} \in \mathbb{R}^N = \mathbf{X}\beta + \mathbf{U} + \mathcal{E} \tag{2}$$
$$\mathbf{U} \in \mathbb{R}^N \sim \mathcal{N}(0, \sigma^2 R) \tag{3}$$
$$\mathcal{E} \in \mathbb{R}^N \sim \mathcal{N}(0, \sigma^2 \mathbf{I}) \tag{4}$$

where

$$\mathbf{X} \in \mathbb{R}^{N\times p} = [\mathbf{x}_1 ... \mathbf{x}_N]^\top \tag{5}$$
$$\beta \in \mathbb{R}^p = [\beta_1, ... \beta_p]^\top \tag{6}$$

There are total 3 parts in this EDM model.

- Mean structure $\mathbf{X}\beta$
- Residual with EDM covariance structure $\mathbf{U}$
- White noise $\mathcal{E}$.

As you can see, the mean structure and white noise are the same as linear regression model, while the residual $\mathbf{U}$ is the key part for EDM which captures dependent structure in time series data. The covariance matrix for $\mathbf{U}$ is $R \in \mathbb{R}^{N\times N} = [r_{ij}]_{i,j=1,...,N}$, which is an exponential decay covariance matrix with each element $r_{ij}$ as

$$e^{-\alpha|t_i - t_j|} \tag{7}$$

where $t_i$ and $t_j$ are time ID (date or week) for observations $i$ and $j$. Specifically, for consecutive data, $R$ has a special form which is

$$R = \begin{bmatrix} 1 & e^{-\alpha} & e^{-2\alpha} & ... & e^{-(N-1)\alpha} \\ e^{-\alpha} & 1 & e^{-\alpha} & ... & e^{-(N-2)\alpha} \\ ... & ... & ... & ... & ... \\ e^{-(N-1)\alpha} & e^{-(N-2)\alpha} & e^{-(N-3)\alpha} & ... & 1 \end{bmatrix}$$

Thus, you can see that $R$ only depends on the time difference among the observations and the covariance structure will change automatically if there are missing data or multiple observations.

The parameters in basic EDM model are

$$\beta, \sigma^2, \alpha \tag{8}$$

Next we will explain how to use Bayesian framework to estimate these parameters.

# 4 Estimation Method

## 4.1 General Bayesian Method

Before we move to the Bayesian estimation procedure for EDM model. This is a general framework for Bayesian method. Bayesian method is a flexible way to estimate the parameters since it allows you to

evaluate how likely the model is, from the observations, weighted by the prior knowledge. Compared to the classic frequentist maximum likelihood estimation (MLE) methodology, Bayesian framework is more flexible and easy to implement. There are several constraints to use MLE. First, you have to make sure MLE exists for the model used. Second, to get the MLE, the optimal situation is to get a closed form solution for parameters. Even if you cannot get a closed form solution, a convex likelihood function will be required to get a reliable estimate for MLE. While for many complicated model, especially there are some sophisticated dependent structure in residuals, the likelihood function can be non convex, which means that you may get stuck into some sub-optimal points or saddle points when doing the optimization. That will lead to a large gap between parameter estimates and their corresponding true values.

In general, suppose our interested parameter is $\theta$ and we would like to estimate $\theta$ using Bayesian method. Here are some steps about how we develop Bayesian framework.

- Prior distribution
  The prior distribution is the distribution of the parameter(s) before any data is observed. Assume prior distribution for $\theta$ is based on a hyperparameter $\alpha$. Generally speaking, prior distribution might not be easily determined.
  $$p(\theta|\alpha) \tag{9}$$
- Likelihood function
  Once the model for the observation is determined, you can write the likelihood function for the parameters. Likelihood function is a function for the parameters and in our case, it's the same as distribution of the observed data conditional on its parameters,
  $$L(\theta|X) = P(X|\theta) \tag{10}$$
- Posterior distribution
  Posterior is the distribution of the parameter(s) after taking into account the observed data. This is determined by Bayes' rule, which forms the heart of Bayesian inference:
  $$p(\theta|\mathbf{X}, \alpha) = \frac{p(\theta, \mathbf{X}, \alpha)}{p(\mathbf{X}, \alpha)} = \frac{p(\mathbf{X}|\theta, \alpha)p(\theta|\alpha)}{p(\mathbf{X}|\alpha)} \propto p(\mathbf{X}|\theta, \alpha)p(\theta|\alpha) \tag{11}$$
  Since the posterior distribution is proportional to the multiplication of prior distribution and likelihood function, we can only focus on these two parts in most of the cases. Like if we choose the conjugate prior, the posterior distribution is the same as the prior distribution but with different parameters. We can use prior distribution and likelihood function to get the parameters for posterior distribution without calculating the annoying normalizing constant in the denominator for the Bayes formula.

Once the posterior distribution has been derived, you will have an overview about how the parameter distribution will be like. If you would like get "one" value estimate for the parameters, you can pick the one with largest probability density value. Like if the posterior distribution is a normal distribution, you can pick the mean as the estimate for that parameter.

## 4.2 Bayesian Method for basic EDM model

We try to derive the Bayesian framework for EDM model and show the posterior distribution for parameters.

The likelihood and log-likelihood function for basic EDM model (equation 2) are

$$L(\beta, \sigma^2, \alpha|\mathbf{Y}, \mathbf{X}) \propto \det(\sigma^2(R + \mathbf{I}))^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\beta)^\top (R + \mathbf{I})^{-1}(\mathbf{Y} - \mathbf{X}\beta)\right) \tag{12}$$

$$\ell(\beta, \sigma^2, \alpha|\mathbf{Y}, \mathbf{X}) \propto -\frac{1}{2}\ln\det(R + \mathbf{I}) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\beta)^\top (R + \mathbf{I})^{-1}(\mathbf{Y} - \mathbf{X}\beta) \tag{13}$$

Denote $\Lambda \in \mathbb{R}^{N\times N} = R + \mathbf{I} = [\lambda_{ij}]_{i,j=1,\dots,N}$, which is

$$\lambda_{ij} = \begin{cases} 2 & \text{if } i = j \\ e^{-\alpha|t_i - t_j|} & \text{if } i \neq j \end{cases}$$

If observations are daily consecutive data,

$$\Lambda = \begin{bmatrix} 2 & e^{-\alpha} & e^{-2\alpha} & \dots & e^{-(N-1)\alpha} \\ e^{-\alpha} & 2 & e^{-\alpha} & \dots & e^{-(N-2)\alpha} \\ \dots & \dots & \dots & \dots & \dots \\ e^{-(N-1)\alpha} & e^{-(N-2)\alpha} & e^{-(N-3)\alpha} & \dots & 2 \end{bmatrix}$$

We will show how we derive the posterior distribution for $\beta$ and $\sigma^2$ in detail. Since basic EDM model follows multivariate normal distribution, we choose conjugate prior which simplifies derivation for posterior distribution. Besides, we construct cross validation procedure to estimate $\alpha$.

**Posterior distribution for $\beta$** Suppose prior distribution for $\beta$ is

$$\beta^{\text{Prior}} \sim \mathcal{N}(\theta, \Omega) \tag{14}$$

The posterior distribution can be derived using Bayes formula (equation 11)

$$\begin{aligned} \ln p(\beta|\sigma^2, \alpha, \mathbf{Y}, \mathbf{X}) &\propto \ell(\beta, \sigma^2, \alpha|\mathbf{Y}, \mathbf{X}) \ln p(\beta) && (15)\\ &\propto (\mathbf{Y} - \mathbf{X}\beta)^\top \Lambda^{-1}(\mathbf{Y} - \mathbf{X}\beta) \cdot (\beta - \theta)^\top \Omega^{-1}(\beta - \theta) && (16)\\ &\propto \beta^\top(\Omega^{-1} + \mathbf{X}^\top\Lambda^{-1}\mathbf{X}/\sigma^2)\beta - 2(\theta^\top\Omega^{-1} + \mathbf{Y}^\top\Lambda^{-1}\mathbf{X}/\sigma^2)\beta && (17) \end{aligned}$$

Thus,

$$\begin{aligned} \beta^{\text{Posterior}}|\sigma^2, \alpha, \mathbf{Y}, \mathbf{X} &\sim \mathcal{N}(\theta', \Omega') && (18)\\ \theta' &= \Omega'(\Omega^{-1}\theta + \mathbf{X}^\top\Lambda^{-1}\mathbf{Y}/\sigma^2) && (19)\\ \Omega' &= (\Omega^{-1} + \mathbf{X}^\top\Lambda^{-1}\mathbf{X}/\sigma^2)^{-1} && (20) \end{aligned}$$

**Posterior distribution for $\sigma^2$** Suppose the prior distribution for $\sigma^2$ is

$$(\sigma^2)^{\text{Prior}} \sim \mathcal{G}^{-1}(\gamma, \tau) \tag{21}$$

where $\mathcal{G}^{-1}$ represents inverse gamma distribution. The posterior distribution is

$$\begin{aligned} p(\sigma^2|\beta, \alpha, \mathbf{Y}, \mathbf{X}) &\propto (\sigma^2)^{-\frac{N}{2}} \exp\Big(-\frac{(\mathbf{Y} - \mathbf{X}\beta)^\top\Lambda^{-1}(\mathbf{Y} - \mathbf{X}\beta)}{2\sigma^2}\Big) \cdot (\sigma^2)^{-\gamma-1} \exp\Big(-\frac{\tau}{\sigma^2}\Big) && (22)\\ &\propto (\sigma^2)^{-(\frac{N}{2}+\gamma+1)} \exp\Big(-\frac{\tau + (\mathbf{Y} - \mathbf{X}\beta)^\top\Lambda^{-1}(\mathbf{Y} - \mathbf{X}\beta)/2}{\sigma^2}\Big) && (23) \end{aligned}$$

Thus,

$$\begin{aligned} (\sigma^2)^{\text{Posterior}}|\beta, \alpha, \mathbf{Y}, \mathbf{X} &\sim \mathcal{G}^{-1}(\gamma', \tau') && (24)\\ \gamma' &= \frac{N}{2} + \gamma && (25)\\ \tau' &= \frac{(\mathbf{Y} - \mathbf{X}\beta)^\top\Lambda^{-1}(\mathbf{Y} - \mathbf{X}\beta)}{2} + \tau && (26) \end{aligned}$$

**How to choose $\alpha$?** We choose $\alpha$ using cross validation (CV) with $CV_\alpha$ range

$$\{0.001, 0.005, 0.01, 0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 0.9, 1, 2, 3, 4, 5, 10\}$$

As you can see, the CV range takes a large range of candidate values. $\alpha$ controls the level of dependency. The large value it takes, the weaker dependency level will be. In the extreme case with $\alpha = 0$, it means that each entry in residual $\mathbf{U}$ are perfectly dependent on the other entries. The CV range is not equally divided between 0 and 10. We choose more candidate values in the interval $[0, 1]$ in that $e^{-0.001} = 0.9990005$ and $e^{-1} = 0.3678794$ have a big difference for the dependency level. However, $e^{-5} = 0.006737947$ and $e^{-10} = 4.539993e-05$ are both pretty small and thus we don't split the interval $[5, 10]$ any further.

### 4.3 Bayesian algorithm for basic EDM model

Algorithm 1 shows how we do Bayesian estimation for EDM model. The key step is to sample $\beta$ and $\sigma^2$ from their posterior distribution respectively. As you can see, the posterior distribution for $\beta$ relies on $\sigma^2$ and similarly, posterior distribution for $\sigma^2$ depends on the value of $\beta$. This is the key idea for Gibbs sampling schema. For calculating the prediction based on the estimated parameters, please refer to (Deb and Tsay, 2017).

**Algorithm 1** Basic EDM model

**Result:** Estimate for $\beta$, $\sigma^2$, $\alpha$
Initialization: $\theta, \Omega, \gamma, \tau, CV, Tol$
Split dataset into training and testing dataset
**for** $\alpha \leftarrow CV_\alpha[1] = 0.001$ *to* $CV_\alpha[end] = 10$ **do**
  **for** $t$ = *1,...* **do**
    **while** $|\beta^t - \beta^{t-1}| <$ *Tol and* $|(\sigma^2)^{t+1} - (\sigma^2)^t| <$ *Tol* **do**
      Sample $\beta^t$ from $\mathcal{N}(\theta'[\alpha, (\sigma^2)^{t-1}], \Omega'[\alpha, (\sigma^2)^{t-1}])$
      Sample $(\sigma^2)^t$ from $\mathcal{G}^{-1}(\gamma'[\alpha, \beta^t], \tau'[\alpha, \beta^t])$
    **end**
  **end**
  Get estimate for $\hat{\beta}[\alpha]$, $\hat{\sigma}^2[\alpha]$
  Calculate prediction on training and testing dataset
  Calculate MAPE/SMAPE on test dataset
**end**
Choose optimal model $\beta$, $\sigma^2$, $\alpha$ with $min$(MAPE) or $min$(SMAPE)

## 5 Model extension

In this section, we will introduce three updated versions of the basic EDM model. As we have mentioned for EDM model structure, EDM model can be split into 3 parts and we will make the corresponding improvement on these three parts.

- Mean structure → Add change points
- Residual with EDM covariance → Add weekly-yearly seasonality + Add variance flexibility
- White noise → Add variance flexibility

Notice that by adding variance flexibility, we also change the structure for residual with EDM covariance since the distribution for residual $\mathbf{U}$ is also determined by the variance parameter $\sigma^2$.

## 5.1 Add change points

Basic EDM model only fits a simple linear regression to the mean structure which might be unreasonable if there are some nonlinear trends in the mean part. See the following simple example (Figure 1), there is an obvious nonlinear trend and one way to improve the model fitting is by adding change

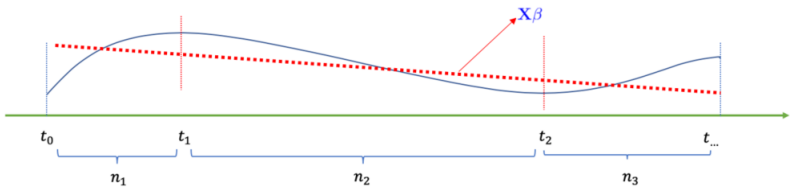


Figure 1: Basic EDM model mean structure fitting

points at specific time points, Now we have different values for $\beta_1, \beta_2, \beta_3 \in \mathbb{R}^p$ for different time periods

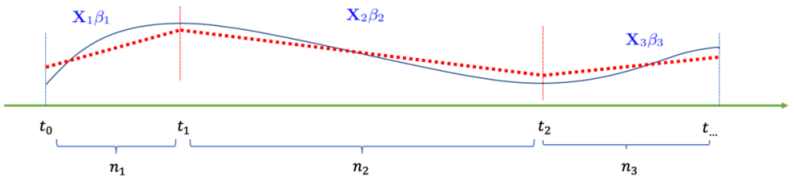


Figure 2: Add change points

and we need to estimate them separately. We stick on to Bayesian method to estimate $\beta_1, \beta_2, \beta_3, \sigma^2$. Let's take example shown in Figure 2 and see how we construct Bayesian framework for $\beta_1, \beta_2, \beta_3$.

In this example, we would like to add two change points at time $t_1$ and $t_2$ and for each time period, we have $n_1, n_2, n_3$ observations. After that, our explanatory variable $\mathbf{X}$ and responses $\mathbf{Y}$ will be split into 3 parts as follows:

$$\mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1n_1} & \dots & x_{1(n_1+n_2)} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_{n_1 1} & \dots & \dots & \dots & \dots & \dots & \dots \\ \hdashline \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_{(n_1+n_2)1} & \dots & \dots & \dots & \dots & \dots & \dots \\ \hdashline \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_{(n_1+n_2+n_3)1} & \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \hdashline \mathbf{X}_2 \\ \hdashline \mathbf{X}_3 \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \dots \\ y_{n_1} \\ \hdashline y_{n_1+1} \\ \dots \\ y_{n_1+n_2} \\ \hdashline y_{n_1+n_2+1} \\ \dots \\ y_{n_1+n_2+n_3} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_1 \\ \hdashline \mathbf{Y}_2 \\ \hdashline \mathbf{Y}_3 \end{bmatrix}$$

The likelihood function changes to

$$\ell(\beta_1, ...\beta_3, \sigma^2, \alpha|\mathbf{Y}, \mathbf{X}) \propto \begin{bmatrix} \mathbf{Y}_1 - \mathbf{X}_1\beta_1 \\ \mathbf{Y}_2 - \mathbf{X}_2\beta_2 \\ \mathbf{Y}_3 - \mathbf{X}_3\beta_3 \end{bmatrix}^\top \Lambda(\alpha)^{-1} \begin{bmatrix} \mathbf{Y}_1 - \mathbf{X}_1\beta_1 \\ \mathbf{Y}_2 - \mathbf{X}_2\beta_2 \\ \mathbf{Y}_3 - \mathbf{X}_3\beta_3 \end{bmatrix} := \mathbf{E}^\top \Lambda(\alpha)^{-1}\mathbf{E}$$

Next we will show that by rearranging $\mathbf{E}$, $\mathbf{X}$, $\mathbf{Y}$ and $\Lambda(\alpha)^{-1}$, the Bayesian posterior distribution for $\beta_1, \beta_2, \beta_3$ can be derived similarly. $\forall\ i = 1, 2, 3$, let $\mathcal{S}_i$ represents the indices in time period $i$ and $\mathcal{S}_{(-i)}$ be indices outside time period $i$. Then we rearrange $\mathbf{E}$, $\mathbf{X}$, $\mathbf{Y}$ and $\Lambda(\alpha)^{-1}$ in the following way,

$$\mathbf{E} = \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix} = \begin{bmatrix} E_i \\ E_{(-i)} \end{bmatrix} \quad \Lambda(\alpha)^{-1} = \begin{bmatrix} \Lambda(\alpha)^{-1}_{ii} & \Lambda(\alpha)^{-1}_{i(-i)} \\ \Lambda(\alpha)^{-1}_{i(-i)}{}^\top & \Lambda(\alpha)^{-1}_{(-i)(-i)} \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_i \\ \mathbf{X}_{(-i)} \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} \mathbf{Y}_i \\ \mathbf{Y}_{(-i)} \end{bmatrix}$$

where $E_i = \mathbf{Y}_i - \mathbf{X}_i\beta_i$, i.e. elements in $\mathbf{E}$ with index belongs to $\mathcal{S}_i$. $E_{(-i)}$ are all elements in $\mathbf{E}$ with index belongs to $\mathcal{S}_{(-i)}$. Similarly for $\mathbf{X}$, $\mathbf{Y}$ and $\Lambda(\alpha)^{-1}$.

Taking $i = 2$ as example, $\mathcal{S}_2 = \{n_1+1, ..., n_1+n_2\}$ and $\mathcal{S}_{(-2)} = \{1, 2, ..., n_1, n_1+n_2+1, ..., n_1+n_2+n_3\}$. $\mathbf{E}$ and $\Lambda(\alpha)^{-1}$ can be derived as follows (Figure 3, 4, 5, 6,7) :

$$\mathbf{E} = \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix} = \begin{bmatrix} E_2 \\ E_{(-2)} \end{bmatrix}$$

$$\Lambda(\alpha)^{-1} = \begin{bmatrix} \Lambda(\alpha)^{-1}_{22} & \Lambda(\alpha)^{-1}_{2(-2)} \\ \Lambda(\alpha)^{-1}_{2(-2)}{}^\top & \Lambda(\alpha)^{-1}_{(-2)(-2)} \end{bmatrix}$$

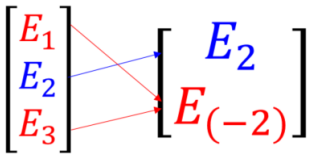


Figure 3: Rearrange $\mathbf{E}$

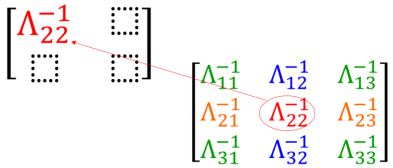


Figure 4: Rearrange $\Lambda(\alpha)^{-1}$: element (1,1)

The posterior distribution for $\beta_i$ can be derived similarly to basic EDM model.

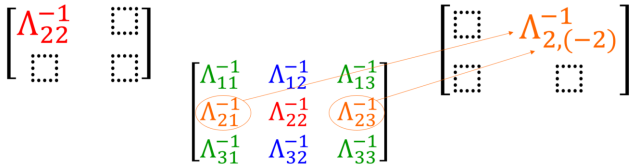

Figure 5: Rearrange $\Lambda(\alpha)^{-1}$: element (1,2)

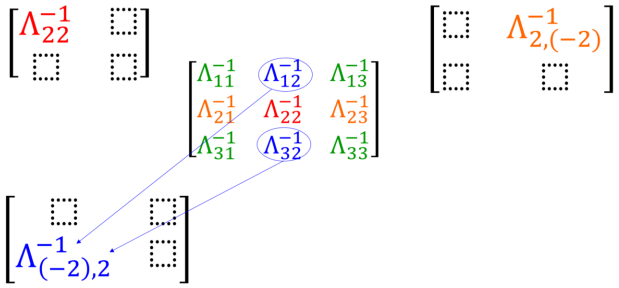

Figure 6: Rearrange $\Lambda(\alpha)^{-1}$: element (2,1)

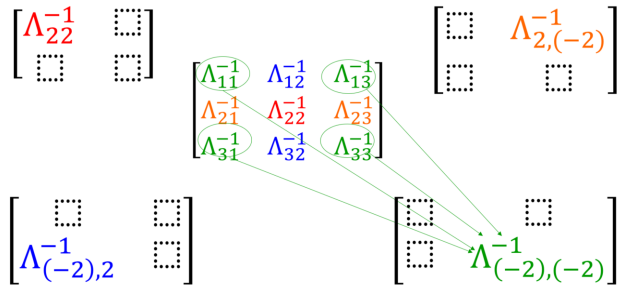

Figure 7: Rearrange $\Lambda(\alpha)^{-1}$: element (2,2)

**Posterior distribution** In general case, suppose the number for change point is *cpt* and prior distribution for $\beta_i$, $\forall\ i = 1, ..., cpt$ is

$$\beta_i^{\text{Prior}} \sim \mathcal{N}(\theta_{1i}, \Omega_{1i})$$
$$\beta_{(-i)}^{\text{Prior}} \sim \mathcal{N}(\theta_{2i}, \Omega_{2i})$$

The posterior distribution is

$$\begin{aligned}\ln p(\beta_i|\beta_{(-i)}, \sigma^2, \alpha, \mathbf{Y}, \mathbf{X}) &\propto \ell(\beta_i, \beta_{(-i)}, \sigma^2, \alpha|\mathbf{Y}, \mathbf{X}) \ln p(\beta_i|\theta_{1i}, \Omega_{1i}) \\ &\propto \beta_i^\top \mathbf{X}_i^\top \Lambda(\alpha)_{ii}^{-1} \mathbf{X}_i \beta_i - 2(\mathbf{Y}_i^\top \Lambda(\alpha)_{ii}^{-1} \mathbf{X}_i + E_{(-i)} \Lambda(\alpha)_{i(-i)}^{-1} \mathbf{X}_i)\beta_i) \\ &\cdot (\beta_i - \theta_{1i})^\top \Omega_{1i}^{-1} (\beta_i - \theta_{1i})\end{aligned}$$

$$\begin{aligned}\beta_i^{\text{Posterior}}|\sigma^2, \alpha, \mathbf{Y}, \mathbf{X} &\sim \mathcal{N}(\theta'_{1i}, \Omega'_{1i}) \\ \theta'_{1i} &= \Omega'_{1i}(\Omega_{1i}^{-1}\theta_{1i} + (\mathbf{X}_i^\top \Lambda_{ii}^{-1} \mathbf{Y}_i + \mathbf{X}_i^\top \Lambda_{i(-i)}^{-1} E_{(-i)})/\sigma^2) \\ \Omega'_{1i} &= (\Omega_{1i}^{-1} + \mathbf{X}_i^\top \Lambda_{ii}^{-1} \mathbf{X}_i/\sigma^2)^{-1}\end{aligned}$$

Similar case for $\beta_{(-i)}^{\text{Posterior}}$

$$\begin{aligned}\beta_{(-i)}^{\text{Posterior}}|\sigma^2,\alpha,\mathbf{Y},\mathbf{X} &\sim \mathcal{N}(\theta'_{2i},\Omega'_{2i})\\ \theta'_{2i} &= \Omega'_{2i}(\Omega_{2i}^{-1}\theta_{2i}+(\mathbf{X}_{(-i)}^{\top}\Lambda_{(-i)(-i)}^{-1}\mathbf{Y}_{(-i)}+\mathbf{X}_{(-i)}^{\top}\Lambda_{i(-i)}^{-1}{}^{\top}E_i)/\sigma^2)\\ \Omega'_{2i} &= (\Omega_{2i}^{-1}+\mathbf{X}_{(-i)}^{\top}\Lambda_{(-i)(-i)}^{-1}\mathbf{X}_{(-i)}/\sigma^2)^{-1}\end{aligned}$$

Also,

$$(\sigma^2)^{\text{Posterior}}|\beta,\alpha,\mathbf{Y},\mathbf{X}\sim\mathcal{G}^{-1}(\frac{N}{2}+\gamma,\frac{\begin{bmatrix}\mathbf{Y}_1-\mathbf{X}_1\beta_1\\ \dots\\ \mathbf{Y}_{cpt}-\mathbf{X}_{cpt}\beta_{cpt}\end{bmatrix}^{\top}\Lambda^{-1}\begin{bmatrix}\mathbf{Y}_1-\mathbf{X}_1\beta_1\\ \dots\\ \mathbf{Y}_{cpt}-\mathbf{X}_{cpt}\beta_{cpt}\end{bmatrix}}{2}+\tau) \quad (27)$$

## 5.2 Add weekly-yearly seasonality

Basic EDM model only captures one time unit dependency structure for the observations. For example, if observations are daily level, basic EDM model only considers daily time dependency. We will show the drawback of basic EDM model by the following simple experiment. We fit the basic EDM model on SalesPrime dataset from network USA. This is a daily level dataset with 7 observations per week. Then we do PACF diagnostic on the residual (Figure 8) from that model to check if there is still some dependent structure in the residuals. Obviously, there are large PACF values at time lag $7, 14, 21, ...$

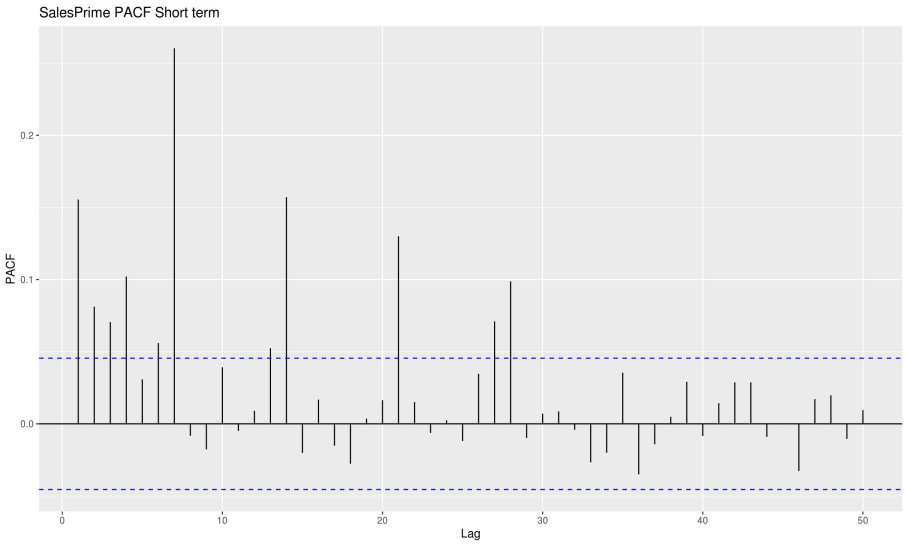


Figure 8: Residual PACF diagnostic for basic EDM model

indicating that basic EDM model doesn't capture the weekly seasonality perfectly. We solve this by adding weekly seasonality to residual covariance matrix.

**Weekly seasonality** Remember element for residual covariance matrix $R$ is

$$r_{ij}=e^{-\alpha|t_i-t_j|}$$

We try to add weekly seasonality dependency into pairs of observation with time lag exactly equal to 1,2,3... weeks. Elements in $R$ will change to

$$r_{ij}=e^{-\alpha_1|t_i-t_j|-\mathbb{1}(mod(|t_i-t_j|/7)=0)\alpha_2(|t_i-t_j|/7)} \quad (28)$$

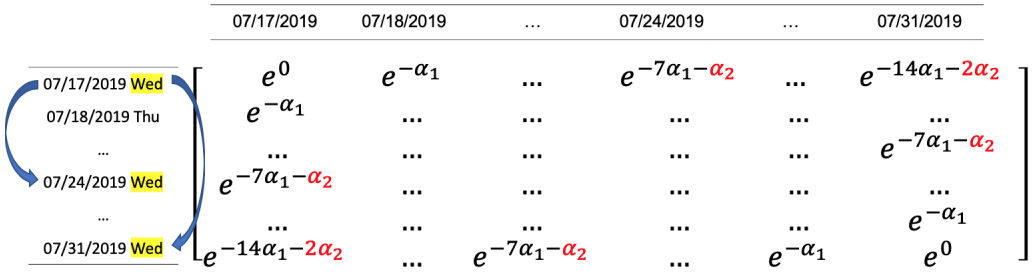


Figure 9: Add weekly seasonality

Consider a simple example, assume we have observations from July 17, 2019 to July 31, 2019, the residual EDM covariance matrix will look like Figure 9. We still use CV to estimate $\alpha_2$ and the CV range for $\alpha_2$ is the same as $\alpha_1$. This time we still use the SalesPrime dataset to test residual dependency for EDM model with weekly seasonality. See Figure 10, we get a better model fitting by adding weekly seasonality!

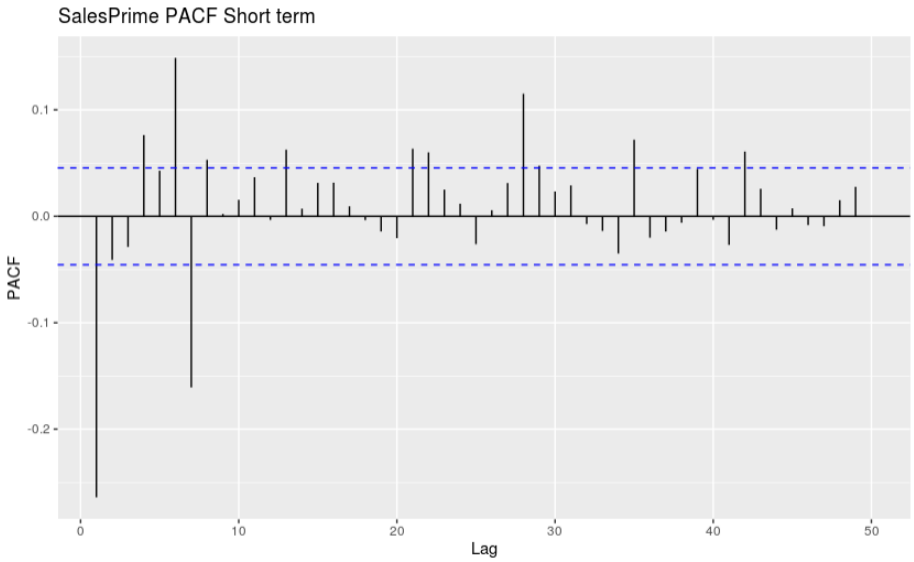


Figure 10: Residual PACF diagnostic for EDM model with weekly seasonality

**Yearly seasonality** To add yearly seasonality, all the framework will be the same except for adding an extra parameter $\alpha_3$ to covariance matrix $R$ representing the yearly seasonality. The element in $R$ will be

$$r_{ij} = e^{-\alpha_1|t_i-t_j|-\mathbb{1}(mod(|t_i-t_j|/7)=0)\alpha_2(|t_i-t_j|/7)-\mathbb{1}(mod(|t_i-t_j|/365)=0)\alpha_3(|t_i-t_j|/365)} \tag{29}$$

Similarly, we use CV to estimate $\alpha_3$.

## 5.3 Add variance flexibility

As we have mentioned, ARIMA/SARIMA model and basic EDM model treat variance parameter $\sigma^2$ as a constant. Let's check whether this assumption is reasonable. We pick a sample dataset Daytime

from network USA and calculate the variance for 'SC3_Impressions' from year 2012 to 2019 and it's clear that there is an exponential decay trend (Figure 11).

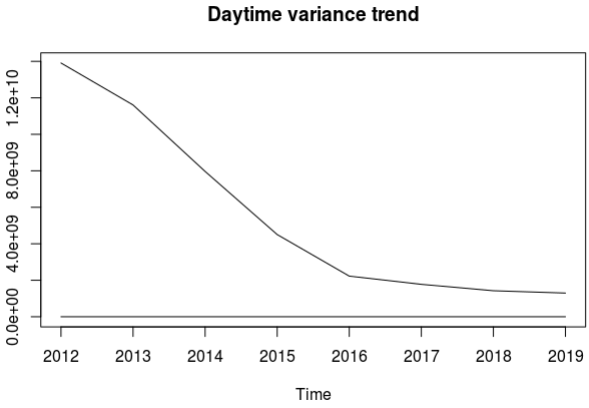


Figure 11: Variance trend for Daytime

We try to write the variance parameter in the matrix form as follows (which is exactly the same as previous basic EDM model (equation 2)):

$$\mathbf{Y} \in \mathbb{R}^N = \mathbf{X}\beta + \mathbf{U} + \mathcal{E} \tag{30}$$

$$\mathbf{U} \in \mathbb{R}^N \sim \mathcal{N}(0, MR) \tag{31}$$

$$\mathcal{E} \in \mathbb{R}^N \sim \mathcal{N}(0, M\mathbf{I}) \tag{32}$$

where

$$M = \begin{bmatrix} \sigma^2 & 0 & 0 & ... & 0 \\ 0 & \sigma^2 & 0 & ... & 0 \\ ... & ... & ... & ... & ... \\ ... & ... & \sigma^2 & ... & ... \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}$$

We would like to add variance flexibility to variance matrix but also avoid the problem of changing variance parameter frequently. Thus, we choose to let variance parameter decay exponentially every two years. Suppose we have observations from year 2014 to year 2019, the new variance matrix will be $M^{new} = \sigma^2\mathbf{C}$ (Figure 12), where $\mathbf{C} \in \mathbb{R}^{N\times N}$ is a diagonal matrix with diagonal elements as a exponential decay function of $c$.

Now, we have an extra parameter $c$ and we will use the CV to estimate. The CV range for $c$ is

$$CV_c = \{0.1, 0.2, ..., 0.9, 1\}$$

Specifically, $c = 1$ represents that the variance parameter $\sigma^2$ is constant over the years.

All the estimation procedure will be the same except for changing the structure for $\Lambda$. Remember that $\Lambda$ in posterior distribution for $\beta_1, ..., \beta_{cpt}$ and $\sigma^2$ will be

$$\Lambda = \mathbf{C}(R + \mathbf{I}) \tag{33}$$

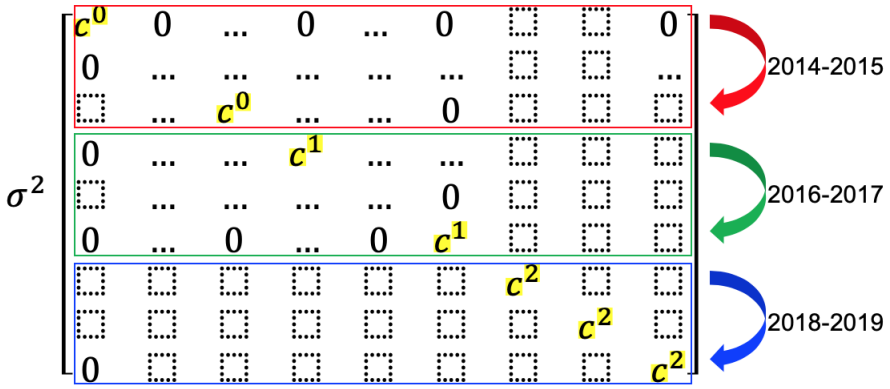


Figure 12: Variance matrix with exponential decay trend

## 5.4 Updated Algorithm

Taking change points, weekly-yearly seasonality and variance flexibility into consideration, the algorithm will be updated to the following version.

**Algorithm 2** EDM + Change points + Weekly-yearly seasonality + Variance Flexibility

**Result:** Estimate for $\beta_1$,...$\beta_{cpt+1}$, $\sigma^2$, $\alpha$
initialization: $\theta_{1i}, \Omega_{1i}, \theta_{2i}, \Omega_{2i}, \gamma, \tau, CV_{\alpha_1}, CV_{\alpha_2}, CV_{\alpha_3}, CV_c, Tol$
Split dataset into train and test
**for** $\alpha_1 \leftarrow CV_{\alpha_1}[1] = 0.001$ *to* $CV_{\alpha_1}[end] = 10$ **do**
  **for** $\alpha_2 \leftarrow CV_{\alpha_2}[1] = 0.001$ *to* $CV_{\alpha_2}[end] = 10$ **do**
    **for** $\alpha_3 \leftarrow CV_{\alpha_3}[1] = 0.001$ *to* $CV_{\alpha_3}[end] = 10$ **do**
      **for** $c \leftarrow CV_c[1] = 0.1$ *to* $CV_c[end] = 1$ **do**
        **for** $t$ = *1,...* **do**
          **while** $|\beta_i^t - \beta_i^{t-1}| <$ *Tol*, $\forall i = 1, ..., cpt$ *and* $|(\sigma^2)^{t+1} - (\sigma^2)^t| <$ *Tol* **do**
            **for** $i$ = *1,..., cpt+1* **do**
              Sample $\beta_i^t$ from $\mathcal{N}(\theta'_{1i}[\alpha_1, \alpha_2, \alpha_3, c, (\sigma^2)^{t-1}], \Omega'_{1i}[\alpha_1, \alpha_2, \alpha_3, c, (\sigma^2)^{t-1}])$
              Sample $\beta_{(-i)}^t$ from $\mathcal{N}(\theta'_{2i}[\alpha_1, \alpha_2, \alpha_3, c, (\sigma^2)^{t-1}], \Omega'_{2i}[\alpha_1, \alpha_2, \alpha_3, c, (\sigma^2)^{t-1}])$
            **end**
            Sample $(\sigma^2)^t$ from $\mathcal{G}^{-1}(\gamma'[\alpha_1, \alpha_2, \alpha_3, c, \beta_1^t, ..., \beta_{cpt+1}^t], \tau'[\alpha_1, \alpha_2, \alpha_3, c, \beta_1^t, ..., \beta_{cpt+1}^t])$
          **end**
        **end**
        Get estimate for $\hat{\beta}_1[\alpha_1, \alpha_2, \alpha_3, c], ..., \hat{\beta}_{cpt+1}[\alpha_1, \alpha_2, \alpha_3, c]$, $\hat{\sigma}^2[\alpha_1, \alpha_2, \alpha_3, c]$
        Calculate prediction on training and testing dataset
        Calculate MAPE/SMAPE on test dataset
      **end**
    **end**
  **end**
**end**
Choose optimal model $\beta_1$, ..., $\beta_{cpt}$, $\sigma^2$, $\alpha_1$, $\alpha_2$, $\alpha_3$, $c$ with $min$(MAPE) or $min$(SMAPE)

# 6 Real dataset performance

We compare EDM model with ARIMA/SARIMA model on 8 different experiments. Each experiment uses one dataset from some networks with different daypart or show name. We choose 4 networks, which are UNVSO, OXYG, NBC_NEWS, USA. The quarterly average MAPE/SMAPE are summarized in the Table 2.

Table 2: ARIMA/SARIMA and BF+EDM quarterly average MAPE/SMAPE comparison: from 2012-08-27 to 2019-06-30, OOS = 6 quarters (2018-01-01)

| ARIMA/SARIMA v.s. BF + EDM | | | | | | |
|---|---|---|---|---|---|---|
| Network | Daypart/Show name | Time ID | SMAPE_test (%) | | MAPE_test (%) | |
| | | | ARIMA | BF | ARIMA | BF |
| UNVSO | Daytime | weekly | 22.0 | 5.01↓ | 52.3 | 9.56↓ |
| UNVSO | Weekday | daily | 12.9 | 6.32↓ | 28.4 | 12.9↓ |
| OXYG | Access | weekly | 11.9 | 2.63↓ | 21.1 | 5.07↓ |
| OXYG | SalesPrime | weekly | 5.60 | 2.80↓ | 10.5 | 5.49↓ |
| NBC_NEWS | Today I | weekly | 9.40 | 2.87↓ | 21.1 | 5.97↓ |
| NBC_NEWS | Today III | weekly | 16.0 | 3.25↓ | 38.8 | 6.51↓ |
| USA | SalesPrime | daily | 5.70 | 2.56↓ | 12.1 | 5.27↓ |
| USA | Daytime | daily | 1.90 | 1.68↓ | 3.60 | 3.41↓ |

In all these 8 experiments, EDM achieves better prediction performance than ARIMA/SARIMA!

# 7 Summary

Exponential decay method is a new, more flexible way to model time series data. By introducing change points, weekly-yearly seasonality and variance flexibility, EDM method achieves surprising prediction performance on real dataset.

Currently we treat the parameters $\beta_i$ for each time period to be different and that easily lead to overfitting problem! For further work, how to add change points with specific structure, like broken stick linear regression or smoothing splines will be an interesting topic.

Another possible area to improve the model performance is choice for prior distribution. We use simple conjugate prior for multivariate normal distribution now and other prior distribution may have better prediction and quick convergence rate.

# 8 Acknowledgements

A very special gratitude goes out to all down at NBCUniversal for helping and providing opportunity and dataset for the work.

Special thanks to Soudeep for his patient guidance through the whole project. Also, I'm grateful to Marco who provided lots of helpful feedbacks.

With a special mention to Jiabin, Tong, Sebastien, Anup, Juan, Amanda and Adam. It was fantastic to have the opportunity to work as a team with you.

Thanks to my intern cohort, Xiao, Derek and Steven!

# References

Deb, S. and Tsay, R. S. (2017). Spatio-temporal models with space-time interaction and their applications to air pollution data. *arXiv preprint arXiv:1801.00211*.