Bayesian Exponential Decay Method with Spike and Slab prior for time series forecasting

Jiacheng Wang

August 18, 2020

1 Introduction

Bayesian exponential decay method is a well-established method for time series modeling and provides reliable and robust forecasting performance under different scenarios, i.e., different networks (USA, OXYG, UNVSO...) and different level of data (weekly, daily, hourly, half-hourly...). As many popular machine learning techniques, one critical challenge for bayesian exponential decay method is overfitting. Several reasons may account for this issue:

- As more regressors added into the model mean linear regression structure, forecasting performance in the training dataset is improved quite a lot. But the collinearity problems may be exacerbated, which also increases complexity in model interpretability.
- Change Points are part of critical tools to make model better fit the trend in the time series data. Adding 2 or 3 change points to the bayesian exponential decay model can double/triple the number of parameters in the model, which restricted the number of change points added to the model setting.

Bayesian exponential decay method simply employs stepwise AIC criterion to perform variable selection which partially alleviates the problem of overfitting. But there are still some essential problems:

- R^2 values are biased high.
- The F statistics do not have the claimed distribution.
- The standard errors of the parameter estimates are too small.
- Consequently, the confidence intervals around the parameter estimates are too narrow.
- p-values are too low, due to multiple comparisons, and are difficult to correct.
- Parameter estimates are biased away from 0.
- Collinearity problems are exacerbated.
- ...

To improve the basic Bayesian exponential decay method with aim to achieve better prediction performance, we propose a new time series forecasting method, bayesian exponential decay method with spike and slab prior. This provides a powerful way of reducing a large set of correlated variables into a parsimonious model, while also imposing prior beliefs on the model. Furthermore, by using priors on the regressor coefficients, the model incorporates uncertainties of the coefficient estimates when producing the credible interval for the forecasts.

2 Model

2.1 Review of exponential decay method model setting

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. 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:

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

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

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

Based on the current model setting for bayesian exponential decay method, we can derive

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2(R + I)) \tag{9}$$

For simplicity, we denote $\Lambda := R + I$.

2.2 Model framework for bayesian exponential decay method with spike and slab prior

Bayesian exponential decay method with spike and slab prior shares the similar model structure as basic bayesian exponential decay method. The only difference is the prior distribution setting on only one parameter. The spike and slab prior is employed on parameter β and to decrease the number of hyperparameters, we introduce the following hierarchical structure of prior distribution setting:

$$\beta_i \sim \mathbb{1}(\pi_i = 0)\delta_0 + \mathbb{1}(\pi_i = 1)\mathcal{N}(0, \sigma^2\tau^2) \tag{10}$$
$$\pi_i \sim \text{Bernoulli}(\theta) \tag{11}$$
$$\theta \sim \text{Beta}(a, b) \tag{12}$$
$$\tau^2 \sim \text{Inv-Gamma}(\frac{1}{2}, \frac{s^2}{/2}) \tag{13}$$
$$\sigma^2 \sim \text{Inv-Gamma}(\alpha_1, \alpha_2) \tag{14}$$

where π_i is the indicator variable controlling whether the i-th regressor should be included in the fitted model, i.e., $\beta_i = 0$. As the name 'spike and slab' indicates, the prior distribution we put on β consists of two parts: δ_0 which is the dirac delta function (the spike), setting a large probability mass of value for β_i exactly at 0; conjugate normal prior distribution (the slab) which is the original prior distribution we have for β in the basic bayesian exponential decay method. We multiply τ^2 with σ^2 so that the prior naturally scales with the scale of the outcome. If we would not do this, then our results would depend on the measurement units of y. We can visualize the relations between all random variables using Figure 1. Obviously, the user-specified hyperparameters reduce to $s, \alpha_1, \alpha_2, a, b$. This facilitates the tuning parameters procedure a lot and to some extent weakens the effect of prior distribution choice on final forecasting performance.

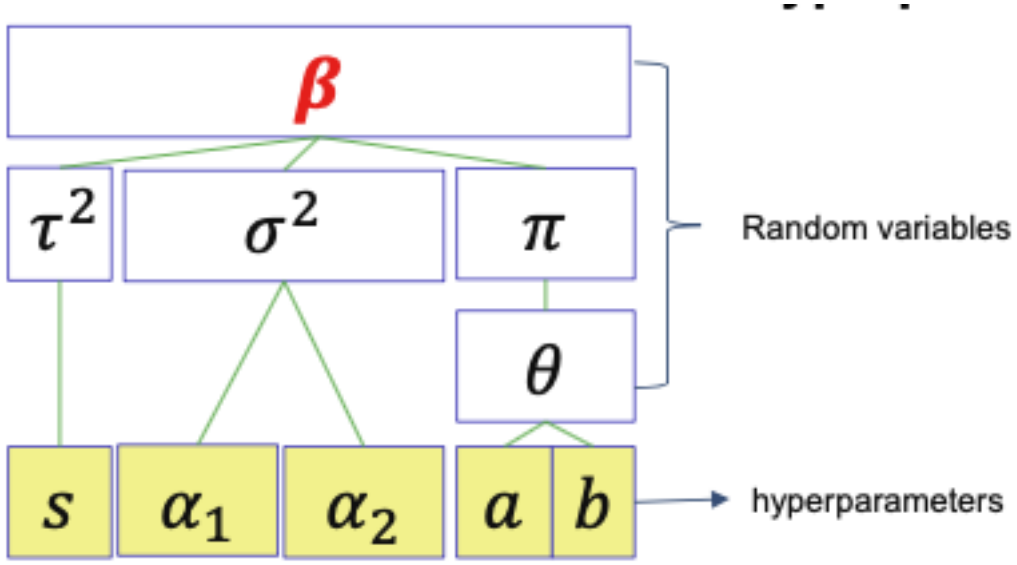

Figure 1: Hierarchical structure of spike and slab prior

3 Estimation

Before we move on to the detailed procedure of posterior distribution derivation, I will illustrate the notations in this subsection.

- $\pi \in \mathbb{R}^p := [\pi_1, ..., \pi_p]^\top$
- $\pi_{(-i)} \in \mathbb{R}^{p-1} := [\pi_1, .., \pi_{(i-1)}, \pi_{(i+1)} ..., \pi_p]^\top$
- $\beta \in \mathbb{R}^p := [\beta_1, ..., \beta_p]^\top$
- $\beta_{(-i)} \in \mathbb{R}^{p-1} := [\beta_1, .., \beta_{(i-1)}, \beta_{(i+1)} ..., \beta_p]^\top$
- $X_i \in \mathbb{R}^n$ which is the ith column in design matrix X.
- $X_{(-i)} \in \mathbb{R}^{n \times (p-1)}$ which remaining matrix in X with ith column removed.

3.1 Posterior distribution for SSB without change points

Posterior distribution for θ

$$p(\theta|\pi) \propto p(\pi|\theta)p(\theta) \tag{15}$$
$$\propto \prod_{i=1}^{p} p(\pi_i|\theta) \frac{1}{Beta(a,b)} \theta^{a-1}(1-\theta)^{b-1} \tag{16}$$
$$\propto \theta^{\sum_{i=1}^p \pi_i}(1-\theta)^{\sum_{i=1}^p (1-\pi_i)} \theta^{a-1}(1-\theta)^{b-1} \tag{17}$$
$$\propto \theta^{a+\sum_i \pi_i - 1}(1-\theta)^{b+p-\sum_i \pi_i - 1} \tag{18}$$

Thus, the posterior distribution for θ is

$$\text{Beta}(a + \sum_{i=1}^{p} \pi_i, b + p - \sum_{i=1}^{p} \pi_i) \tag{19}$$

where $\sum_{i=1}^p \pi_i$ indicates the number of regressors included in forecasting model.

Posterior distribution for τ^2

$$p(\tau^2|\beta, \pi) \propto p(\beta|\tau^2, \pi)p(\pi)p(\tau^2) \tag{20}$$
$$\propto \prod_{i=1}^{p} p(\beta_i|\tau^2, \pi_i)p(\pi_i)p(\tau^2) \tag{21}$$
$$\propto \prod_i \frac{1}{\sqrt{2\pi\sigma^2\tau^2}} \exp\{-\frac{\beta_i^2}{2\sigma^2\tau^2}\} \theta^{\pi_i}(1-\theta)^{1-\pi_i} \frac{(s^2/2)^{1/2}}{\Gamma(1/2)} (\tau^2)^{-1/2-1} \exp\{-\frac{s^2/2}{\tau^2}\} \tag{22}$$
$$\propto (2\pi\sigma^2\tau^2)^{-\frac{\sum_{i=1}^p \pi_i}{2}} \exp\{-\frac{\beta^\top\beta}{2\sigma^2\tau^2}\} \theta^{\sum_i \pi_i}(1-\theta)^{\sum_i (1-\pi_i)} (\tau^2)^{-1/2-1} \exp\{-\frac{s^2/2}{\tau^2}\} \tag{23}$$
$$\propto (\tau^2)^{-\frac{\sum_i \pi_i + 1}{2} - 1} \exp\{-\frac{\beta^\top\beta/2\sigma^2 + s^2/2}{\tau^2}\} \tag{24}$$

The posterior distribution for τ^2 is

$$\text{Inv-Gamma}(\frac{1 + \sum_{i=1}^p \pi_i}{2}, \frac{\beta^\top\beta}{2\sigma^2} + \frac{s^2}{2}) \tag{25}$$

Posterior distribution for σ^2

$$p(\sigma^2|y, \beta) \propto p(y|\beta, \sigma^2)p(\beta|\sigma^2)p(\sigma^2) \tag{26}$$
$$\propto \det(\sigma^2\Lambda)^{-1/2} \exp\{-\frac{1}{2\sigma^2}(y - X\beta)^\top\Lambda^{-1}(y - X\beta)\} (2\pi\sigma^2\tau^2)^{-\frac{\sum_{i=1}^p \pi_i}{2}} \exp\{-\frac{\beta^\top\beta}{2\sigma^2\tau^2}\} \tag{27}$$
$$\times \quad \frac{\alpha_2^{\alpha_1 - 1}}{\Gamma(\alpha_1)} (\sigma^2)^{-\alpha_1 - 1} \exp\{-\frac{\alpha_2}{\sigma^2}\} \tag{28}$$
$$\propto (\sigma^2)^{-n/2 - \sum_{i=1}^p \pi_i/2 - \alpha_1 - 1} \exp\{-\frac{(y - X\beta)^\top\Lambda^{-1}(y - X\beta)/2 + \beta^\top\beta/2\tau^2 + \alpha_2}{\sigma^2}\} \tag{29}$$

The posterior distribution for σ^2 is

$$\text{Inv-Gamma}(\frac{n+\sum_{i=1}^{p}\pi_i}{2}+\alpha_1, \frac{(y-X\beta)^\top\Lambda^{-1}(y-X\beta)}{2}+\frac{\beta^\top\beta}{2\tau^2}+\alpha_2) \tag{30}$$

Posterior distribution for β

$$p(\beta|y,\pi,\tau^2,\sigma^2) \propto p(y|\beta,\sigma^2)p(\beta|\sigma^2,\tau^2) \tag{31}$$

$$\propto \exp\{-\frac{1}{2}(\beta^\top(\frac{X^\top\Lambda^{-1}X}{\sigma^2}+\frac{1}{\sigma^2\tau^2}I)\beta-\frac{2y^\top\Lambda^{-1}X}{\sigma^2}\beta)\} \tag{32}$$

The posterior distribution for β is multivariate normal distribution

$$\mathcal{N}(\Sigma'\frac{X^\top\Lambda^{-1}y}{\sigma^2},\Sigma') \tag{33}$$

where $\Sigma' = (\frac{X^\top\Lambda^{-1}X}{\sigma^2}+\frac{1}{\sigma^2\tau^2}I)^{-1}$.

Posterior distribution for π Let's denote η_j as follows:

$$\eta_i = \frac{p(\pi_i=0|y,\pi_{-i},\beta_{(-i)},\sigma^2,\tau^2,\theta)}{p(\pi_i=1|y,\pi_{-i},\beta_{(-i)},\sigma^2,\tau^2,\theta)+p(\pi_i=0|y,\pi_{-i},\beta_{(-i)},\sigma^2,\tau^2,\theta)} \tag{34}$$

Then we will calculate $p(\pi_i=1|y,\pi_{-i},\beta_{(-i)},\sigma^2,\tau^2,\theta)$ and $p(\pi_i=0|y,\pi_{-i},\beta_{(-i)},\sigma^2,\tau^2,\theta)$ respectively. The latter is easy and we will start from here.

$$p(\pi_i=0|y,\pi_{-i},\beta_{(-i)},\sigma^2,\tau^2,\theta) = \frac{1}{Z}\exp\{-\frac{1}{2\sigma^2}(y-X_{(-i)}\beta_{(-i)})^\top\Lambda(y-X_{(-i)}\beta_{(-i)})\}(1-\theta) \tag{35}$$

$$p(\pi_i=1|y,\pi_{-i},\beta_{(-i)},\sigma^2,\tau^2,\theta) = \frac{1}{Z}\int p(y,\beta_i|\pi_i=1,\pi_{(-i)},\beta_{(-i)},\sigma^2,\tau^2,\theta)p(\pi|\theta)d\beta_i \tag{36}$$

$$= \frac{1}{Z}p(\pi|\theta)\int p(y|\pi_i=1,\pi_{(-i)},\beta_{(-i},\beta_i,\sigma^2,\tau^2,\theta)p(\beta_i|\pi_i,\tau^2)d\beta_i \tag{37}$$

$$= \frac{1}{Z}\theta(2\pi\sigma^2\tau^2)^{-1/2}\int\exp\{-\frac{1}{2\sigma^2}(y-X\beta)^\top\Gamma^{-1}(y-X\beta)-\frac{\beta_i^2}{2\tau^2\sigma^2}\}d\beta_i \tag{38}$$

Denote $z = y - X_{(-i)}\beta_{(-i)}$, then the above equation can be further written as

$$= \frac{1}{Z}\theta(2\pi\sigma^2\tau^2)^{-1/2}\exp\{-\frac{1}{2\sigma^2}z^\top\Lambda^{-1}z\}\int\exp\{-\frac{1}{2}\Big((\frac{X_i^\top\Lambda^{-1}X_i}{\sigma^2}+\frac{1}{\sigma^2\tau^2})\beta_i^2-\frac{2z^\top\Lambda^{-1}X_i}{\sigma^2}\beta_i\Big)\} \tag{39}$$

$$= \frac{1}{Z}\theta(2\pi\sigma^2\tau^2)^{-1/2}\exp\{-\frac{1}{2\sigma^2}z^\top\Lambda^{-1}z\}\exp\{(2\pi/(X_i\Lambda^{(-1)}X_i/\sigma^2+1/\sigma^2\tau^2))^{1/2}\}\exp\{\frac{(z^\top\Lambda^{-1}X_i/\sigma^2)^2}{2(\frac{X_i^\top\Lambda^{-1}X_i}{\sigma^2}+\frac{1}{\sigma^2\tau^2})}\} \tag{40}$$

Thus,

$$1-\eta_i = \frac{1-\theta}{1-\theta+\theta(\sigma^2\tau^2)^{-1/2}(X_j\Lambda^{(-1)}X_j/\sigma^2+1/\sigma^2\tau^2)^{-1/2}\exp\{\frac{(z^\top\Lambda^{-1}X_i/\sigma^2)^2}{2(\frac{X_i^\top\Lambda^{-1}X_i}{\sigma^2}+\frac{1}{\sigma^2\tau^2})}\}} \tag{41}$$

The posterior distribution for π_i is

$$\text{Bernoulli}(1-\eta_i) \tag{42}$$

3.2 'Optimal' choice for α

We choose α using cross validation (CV) with CV_α range

$$\{0.001, 0.005, 0.01, 0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 0.9, 1, 2, 3, 4, 5, 10, 11, 12, 20\}$$

As you can see, the CV range takes a large range of candidate values. α 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.

3.3 Bayesian Algorithm

Algorithm 1 shows framework for how we provide time series forecasting under spike and slab bayesian exponential decay method. The key step is to sample parameters from their posterior distribution respectively. For calculating the prediction based on the estimated parameters, please refer to (**?**).

Algorithm 1 Spike and slab bayesian exponential decay method

Result: Estimate for β, σ^2, α, τ^2, θ, π, α
Initialization: $a, b, s, \alpha_1, \alpha_2, CV, No.MCMC, Burnin$
Split dataset into training and testing dataset
for $\alpha \leftarrow CV_\alpha[1]$ *to* $CV_\alpha[end]$ **do**
  for *t = 1,..., No.MCMC* **do**
    Sample θ^t from Beta$(a + \sum_{i=1}^{p} \pi_i, b + p - \sum_{i=1}^{p} \pi_i)$
    Sample $(\tau^2)^t$ from Inv-Gamma$(\frac{1+\sum_{i=1}^{p} \pi_i}{2}, \frac{\beta^\top \beta}{2\sigma^2} + \frac{s^2}{2})$
    Sample $(\sigma^2)^t$ from Inv-Gamma$(\frac{n+\sum_{i=1}^{p} \pi_i}{2} + \alpha_1, \frac{(y-X\beta)^\top \Lambda^{-1} (y-X\beta)}{2} + \frac{\beta^\top \beta}{2\tau^2} + \alpha_2)$
    Sample β^t from $\mathcal{N}(\Sigma' \frac{X^\top \Lambda^{-1} y}{\sigma^2}, \Sigma')$
    Sample π_i^t from Bernoulli$(1 - \eta_i)$, $\quad \forall i \in \{1, 2, ..., p\}$
    Do pairwise multiplication on β^t and π^t
    Calculate predictions on training and testing dataset
  end
  Discard the burn-in period
  Calculate mean(median, weighted mean) of prediction on training and testing based on MCMC samples
  Calculate MAPE/SMAPE on test dataset
end
Choose optimal model α with min(MAPE) or min(SMAPE)

4 Extension to model with change points

4.1 Introduction to change points model structure

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 2), there is an obvious nonlinear trend and one way to improve the model fitting is by adding change 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

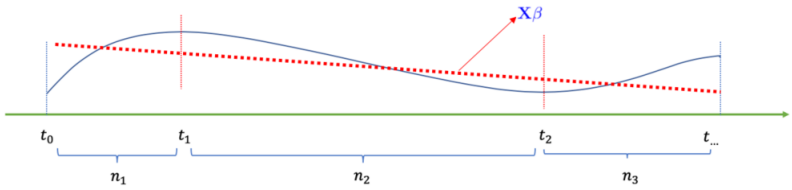

Figure 2: Basic EDM model mean structure fitting

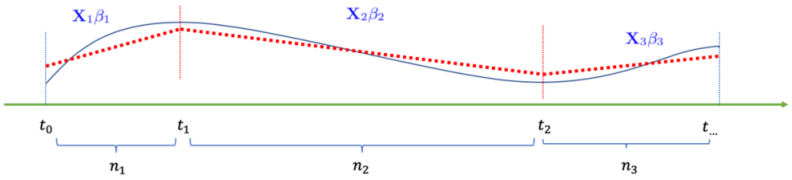

Figure 3: Add change points

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 3 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 X and responses y will be split into 3 parts as follows:

$$X = \begin{bmatrix} x_{11} & \dots & x_{1n_1} & \dots & x_{1(n_1+n_2)} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_{n_1 1} & \dots & \dots & \dots & \dots & \dots & \dots \\ \hdashline \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_{(n_1+n_2)1} & \dots & \dots & \dots & \dots & \dots & \dots \\ \hdashline \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_{(n_1+n_2+n_3)1} & \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix} = \begin{bmatrix} X_1 \\ \hdashline X_2 \\ \hdashline X_3 \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ \dots \\ y_{n_1} \\ \hdashline \\ y_{n_1+1} \\ \dots \\ y_{n_1+n_2} \\ \hdashline \\ y_{n_1+n_2+1} \\ \dots \\ y_{n_1+n_2+n_3} \end{bmatrix} = \begin{bmatrix} y_1 \\ \hdashline \\ y_2 \quad f \\ \hdashline \\ y_3 \end{bmatrix}$$

The likelihood function changes to

$$\ell(\beta_1, \dots \beta_3, \sigma^2, \alpha | y, X) \propto \begin{bmatrix} y_1 - X_1\beta_1 \\ y_2 - X_2\beta_2 \\ y_3 - X_3\beta_3 \end{bmatrix}^\top \Lambda(\alpha)^{-1} \begin{bmatrix} y_1 - X_1\beta_1 \\ y_2 - X_2\beta_2 \\ y_3 - X_3\beta_3 \end{bmatrix} := E^\top \Lambda(\alpha)^{-1} E$$

Next we will show that by rearranging E, X, 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 E, X, y and $\Lambda(\alpha)^{-1}$ in the following way,

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

$$X = \begin{bmatrix} X_i \\ X_{(-i)} \end{bmatrix} \quad y = \begin{bmatrix} y_i \\ y_{(-i)} \end{bmatrix}$$

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

Taking $i = 2$ as example, $\mathcal{S}_2 = \{n_1+1, ..., n_1+n_2\}$ and $\mathcal{S}_{(-2)} = \{1, 2, ..., n_1, n_1+n_2+1, ..., n_1+n_2+n_3\}$. E and $\Lambda(\alpha)^{-1}$ can be derived as follows (Figure 4, 5, 6, 7,8) :

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

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

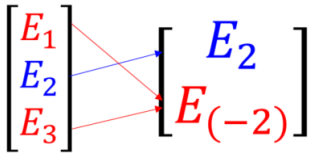

Figure 4: Rearrange E

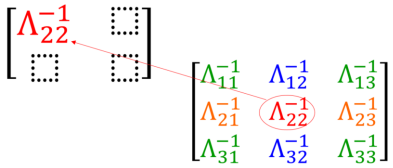

Figure 5: Rearrange $\Lambda(\alpha)^{-1}$: element (1,1)

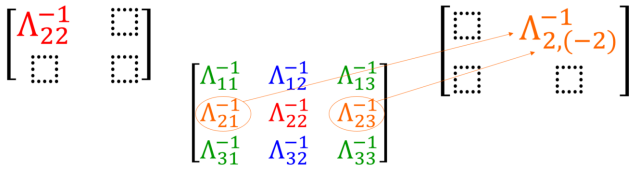

Figure 6: Rearrange $\Lambda(\alpha)^{-1}$: element (1,2)

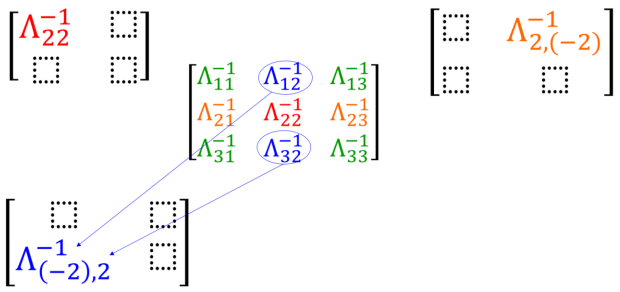

Figure 7: Rearrange $\Lambda(\alpha)^{-1}$: element (2,1)

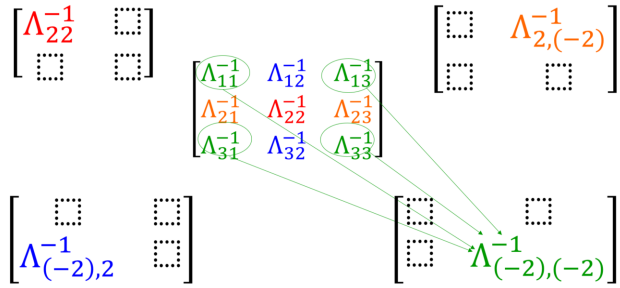

Figure 8: Rearrange $\Lambda(\alpha)^{-1}$: element (2,2)

4.2 Prior distribution

The prior distribution settings are similar to the scenarios without change points.

$$\pi_{i,j} \sim \text{Bernoulli}(\theta), \quad \forall i \in \{1, 2, ..., cpt + 1\}, j \in \{1, 2, ..., p\} \tag{43}$$

$$\theta \sim \text{Beta}(a, b) \tag{44}$$

$$\tau^2 \sim \text{Inv-Gamma}(\frac{1}{2}, \frac{s^2}{2}) \tag{45}$$

$$\sigma^2 \sim \text{Inv-Gamma}(\alpha_1, \alpha_2) \tag{46}$$

$$\beta_i \sim \mathbb{1}(\pi_i = \mathbf{0})\delta_0 + \mathbb{1}(\pi_i = \mathbf{1})\mathcal{N}(0, \sigma^2\tau^2 I), \quad \forall i \in \{1, 2, ..., cpt + 1\} \tag{47}$$

Similarly, $a, b, s, \alpha_1, \alpha_2$ are the pre-specified hyperparameters.

4.3 Posterior distribution

Same as the previous subsection, Let's get familiar with the notations. With a little bit of abuse of notation, we have

- $\pi \in \mathbb{R}^{(cpt+1)\times p} := [\pi_{i,j}]_{i=1,j=1}^{cpt+1,p} = \begin{bmatrix} \pi_1^\top \\ ... \\ \pi_p^\top \end{bmatrix}$
- $\pi_{i,(-j)} \in \mathbb{R}^{p-1}$ is the ith row in the π with jth element removed.
- $\pi_{(-i)} \in \mathbb{R}^{cpt\times p}$ is the remaining matrix in π with the ith row removed.
- $\beta \in \mathbb{R}^{(cpt+1)\times p} := [\beta_{i,j}]_{i=1,j=1}^{cpt+1,p} = \begin{bmatrix} \beta_1^\top \\ ... \\ \beta_p^\top \end{bmatrix}$
- $\beta_{i,(-j)} \in \mathbb{R}^{p-1}$ is the ith row in the β with jth element removed.

- $\beta_{(-i)} \in \mathbb{R}^{cpt \times p}$ is the remaining matrix in β with the ith row removed.
- $X_i \in \mathbb{R}^{n_i \times p}$ which is the design matrix for ith time period after adding change points.
- $X_{i,j} \in \mathbb{R}^{n_i}$ which is the jth column in X_i.
- $z_i \in \mathbb{R}^{n_i} := y_i - X_{i,(-j)}\beta_{i,(-j)}$ which can be explained as partial residual.

Posterior distribution for θ Same as the case with no change points.

Posterior distribution for τ^2

$$p(\tau^2|\beta,\pi) \propto \prod_{i=1}^{cpt+1} p(\beta_i|\tau^2,\pi)p(\pi)p(\tau^2) \tag{48}$$

$$\propto \prod_{i=1}^{cpt+1}\prod_{j=1}^{p} p(\beta_{ij}|\tau^2,\pi_{ij})p(\pi_{ij})p(\tau^2) \tag{49}$$

$$\propto \prod_{i,j} \frac{1}{\sqrt{2\pi\sigma^2\tau^2}} \exp\{-\frac{\beta_{ij}^2}{2\sigma^2\tau^2}\}\theta^{\pi_{ij}}(1-\theta)^{1-\pi_{ij}} \frac{(s^2/2)^{1/2}}{\Gamma(1/2)}(\tau^2)^{-1/2-1}\exp\{-\frac{s^2/2}{\tau^2}\} \tag{50}$$

$$\propto (2\pi\sigma^2\tau^2)^{-\frac{\sum_{i,j}\pi_{ij}}{2}} \exp\{-\frac{\sum_i \beta_i^\top\beta_i}{2\sigma^2\tau^2}\}\theta^{\sum_{i,j}\pi_{i,j}}(1-\theta)^{\sum_{i,j}(1-\pi_{i.j})}(\tau^2)^{-1/2-1}\exp\{-\frac{s^2/2}{\tau^2}\} \tag{51}$$

$$\propto (\tau^2)^{-\frac{\sum_{i,j}\pi_{i,j}+1}{2}-1}\exp\{-\frac{\sum_i \beta_i^\top\beta_i/2\sigma^2 + s^2/2}{\tau^2}\} \tag{52}$$

The posterior distribution for τ^2 is

$$\text{Inv-Gamma}(\frac{1+\sum_{i=1}^{cpt+1}\sum_{j=1}^{p}\pi_{i,j}}{2}, \frac{\sum_{i=1}^{cpt+1}\beta_i^\top\beta_i}{2\sigma^2} + \frac{s^2}{2}) \tag{53}$$

Posterior distribution for σ^2

$$p(\sigma^2|y,\beta) \propto p(y|\beta,\sigma^2)p(\beta|\sigma^2)p(\sigma^2) \tag{54}$$

$$\propto \det(\sigma^2\Lambda)^{-1/2}\exp\{-\frac{1}{2\sigma^2}E^\top\Lambda^{-1}E\}(2\pi\sigma^2\tau^2)^{-\frac{\sum_{i=1}^{cpt+1}\sum_{j=1}^{p}\pi_i}{2}}\exp\{-\frac{\sum_{i=1}^{cpt+1}\beta_i^\top\beta_i}{2\sigma^2\tau^2}\} \tag{55}$$

$$\times \quad \frac{\alpha_2^{\alpha_1-1}}{\Gamma(\alpha_1)}(\sigma^2)^{-\alpha_1-1}\exp\{-\frac{\alpha_2}{\sigma^2}\} \tag{56}$$

$$\propto (\sigma^2)^{-n/2-\sum_{i=1}^{cpt+1}\sum_{j=1}^{p}\pi_{i,j}/2-\alpha_1-1}\exp\{-\frac{E^\top\Lambda^{-1}E/2 + \sum_{i=1}^{cpt+1}\beta_i^\top\beta_i/2\tau^2 + \alpha_2}{\sigma^2}\} \tag{57}$$

The posterior distribution for σ^2 is

$$\text{Inv-Gamma}(\frac{n+\sum_{i=1}^{cpt+1}\sum_{j=1}^{p}\pi_{i,j}}{2}+\alpha_1, \frac{E^\top\Lambda^{-1}E}{2} + \frac{\sum_{i=1}^{cpt+1}\beta_i^\top\beta_i}{2\tau^2}+\alpha_2) \tag{58}$$

Posterior distribution for $\beta_i, \forall \quad i \in \{1,2,...,cpt+1\}$

$$p(\beta_i|y,\pi,\tau^2,\sigma^2,\beta_{(-i)}) \propto \exp\{-\beta_i^\top X_i^\top\Lambda_{ii}^{-1}X_i\beta_i - 2(y_i^\top\Lambda_{ii}^{-1}X_i + E_{(-i)}^\top\Lambda_{i,(-i)}^{-1}X_i)\beta_i\}\exp\{-\frac{\beta_i^\top\beta_i}{2\sigma^2\tau^2}\} \tag{59}$$

$$\propto \exp\{-\frac{1}{2}(\beta_i^\top(\frac{X_i^\top\Lambda_{ii}^{-1}X_i}{\sigma^2} + \frac{1}{\sigma^2\tau^2}I)\beta_i - \frac{2(y_i^\top\Lambda_{ii}^{-1}X_i + E_{(-i)}^\top\Lambda_{i,(-i)}^{-1}X_i)}{\sigma^2}\beta_i)\} \tag{60}$$

The posterior distribution for β_i is multivariate normal distribution

$$\mathcal{N}(\Sigma' \frac{X_i^\top \Lambda_{ii}^{-1} y + X_i^\top \Lambda_{i,(-i)}^{-1} E_{(-i)}}{\sigma^2}, \Sigma'), \quad \text{where} \quad \Sigma' = (\frac{X_i^\top \Lambda_{ii}^{-1} X_i}{\sigma^2} + \frac{1}{\sigma^2\tau^2} I)^{-1} \tag{61}$$

where $\Sigma' = (\frac{X_i^\top \Lambda_{ii}^{-1} X_i}{\sigma^2} + \frac{1}{\sigma^2\tau^2} I)^{-1}$.

Posterior distribution for $\pi_i, \forall \quad i \in \{1, 2, ..., cpt + 1\}$ Let $\eta_{i,j}$ represents the probability of the jth regressor is omited in the ith time period after adding change points to the mean structure, i.e.,

$$\eta_{i,j} = \frac{p(\pi_{i,j} = 0 | y, \pi_{i,(-j)}, \pi_{(-i)}, \beta_{i,(-j)}, \beta_{(-i)}, \sigma^2, \tau^2, \theta)}{p(\pi_i = 1 | y, \pi_{i,(-j)}, \pi_{(-i)}, \beta_{i,(-j)}, \beta_{(-i)}, \sigma^2, \tau^2, \theta) + p(\pi_i = 0 | y, \pi_{i,(-j)}, \pi_{(-i)}, \beta_{i,(-j)}, \beta_{(-i)}, \sigma^2, \tau^2, \theta)} \tag{62}$$

Then we show how to derive the numerator and denominator in the above equation respectively. For simplicity, we denote $\eta_{i,j} = \frac{p(\pi_{i,j}=0|\Theta)}{p(\pi_{i,j}=0|\Theta) p(\pi_{i,j}=|\Theta)}$.

$$p(\pi_{i,j} = 0 | \Theta) = \frac{1}{Z} p(y | \pi_{i,j} = 0, \pi_{i,(-j)}, \pi_{(-i)}, \beta_{i,(-j)}, \beta_{(-i)}, \sigma^2, \tau^2, \theta) p(\pi_{i,(-j)}, \pi_{(-i)} | \theta) p(\beta_{i,(-j)}, \beta_{(-i)} | \sigma^2, \tau^2, \theta) \tag{63}$$

$$\times \, p(\sigma^2) p(\tau^2) p(\theta) \tag{64}$$

$$= \frac{1}{Z} \exp\{-\frac{1}{2\sigma^2}(z_i^\top \Lambda_{i,i}^{-1} z_i + 2E_{(-i)}^\top \Lambda_{(-i),i}^{-1} z + E_{(-i)}^\top \Lambda_{(-i),(-i)}^{(-1)} E_{(-i)})\}(1 - \theta) \tag{65}$$

$$p(\pi_{i,j} = 1 | \Theta) = \frac{1}{Z} \int p(y, \beta_{ij} | \pi_{i,j} = 1, \pi_{i,(-j)}, \pi_{(-i)}, \beta_{i,(-j)}, \beta_{(-i)}, \sigma^2, \tau^2, \theta) p(\beta_{i,j} | \pi_{i,j} = 1) p(\pi_{i,j} | \theta) d\beta_{i,j} \tag{66}$$

$$= \theta (2\pi\sigma^2\tau^2)^{-1/2} \exp\{-\frac{1}{2\sigma^2}(z_i^\top \Lambda_{i,i}^{-1} z_i + 2E_{(-i)}^\top \Lambda_{(-i),i}^{-1} z_i + E_{(-i)}^\top \Lambda_{(-i),(-i)}^{(-1)} E_{(-i)})\} \tag{67}$$

$$\times \int \exp\{-\frac{1}{2\sigma^2}(X_{ij}^\top \Lambda_{i,i}^{-1} X_{ij} \beta_{i,j}^2 - 2(z_i^\top \Lambda_{i,i}^{-1} X_{i,j} + E_{(-i)}^\top \Lambda_{(-i),i}^{-1} X_{i,j}) \beta_{i,j})\} \exp\{-\frac{1}{2\sigma^2\tau^2} \beta_{i,j}^2\} d\beta_{i,j} \tag{68}$$

$$= \theta (2\pi\sigma^2\tau^2)^{-1/2} \exp\Big\{ -\frac{1}{2\sigma^2}(z_i^\top \Lambda_{i,i}^{-1} z_i + 2E_{(-i)}^\top \Lambda_{(-i),i}^{-1} z_i + E_{(-i)}^\top \Lambda_{(-i),(-i)}^{(-1)} E_{(-i)}) \Big\} \tag{69}$$

$$\times \Big(\frac{2\pi}{\frac{X_{ij}^\top \Lambda_{i,i}^{-1} X_{ij}}{\sigma^2} + \frac{1}{\sigma^2\tau^2}} \Big)^{1/2} \exp\Big\{ \frac{(z_i^\top \Lambda_{i,i}^{-1} X_{i,j} + E_{(-i)}^\top \Lambda_{(-i),i}^{-1} X_{i,j})^2}{2(\frac{X_{ij}^\top \Lambda_{i,i}^{-1} X_{ij}}{\sigma^2} + \frac{1}{\sigma^2\tau^2})} \Big\} \tag{70}$$

Thus,

$$\eta_{i,j} = \frac{1 - \theta}{1 - \theta + \theta(\sigma^2\tau^2)^{-1/2}(\frac{X_{ij}^\top \Lambda_{i,i}^{-1} X_{ij}}{\sigma^2} + \frac{1}{\sigma^2\tau^2})^{-1/2} \exp\Big\{ \frac{(z_i^\top \Lambda_{i,i}^{-1} X_{i,j} + E_{(-i)}^\top \Lambda_{(-i),i}^{-1} X_{i,j})^2}{2(\frac{X_{ij}^\top \Lambda_{i,i}^{-1} X_{ij}}{\sigma^2} + \frac{1}{\sigma^2\tau^2})} \Big\}} \tag{71}$$

The posterior distribution for $\pi_{i,j}$ is

$$\text{Bernoulli}(1 - \eta_{i,j}) \tag{72}$$

5 Real Dataset Performance

We compare bayesian exponential decay method with spike and slab prior with basic bayesian exponential decay method on 16 different experiments. Each experiment uses one dataset from some

networks with different daypart and time ID. We choose 3 networks, which are UNVSO, OXYG, USA. The quarterly average MAPE/SMAPE are summarized in the Table 1.

Table 1: Basic EDM (B-EDM) and Spike and slab EDM (SS-EDM) quarterly average MAPE/SMAPE comparison: from 2012-08-27 to 2019-06-30, OOS = 6 quarters (2018-01-01)

Basic EDM v.s. Spike and slab EDM						
Network	Day-part/Time	Time ID	SMAPE_test (%)		MAPE_test (%)	
			BEDM	SSEDM	BEDM	SSEDM
USA	SalesPrime	Weekly	4.65	1.82↓	8.82	3.51↓
USA	Daytime	Weekly	3.84	3.5↓	8.1	7.37↓
USA	EarlyFringe	Weekly	3.66	0.72↓	7.63	1.46↓
USA	Overnight	Weekly	8.85	5.01↓	19.8	10.54↓
USA	LateNight	Weekly	3.62	3.28↓	7.01	6.23↓
USA	Weekend	Weekly	2.26	1.4↓	4.54	2.75↓
USA	WWE	Weekly	2.07	4.89↑	4.12	10.3↑
USA	SalesPrime	Daily	2.66	2.23↓	5.50	4.39↓
USA	Daytime	Daily	5.98	4.13↓	11.13	7.71↓
OXYG	SalesPrime 00:00	Half Hour	7.04	6.95↓	12.99	11.92↓
OXYG	SalesPrime 00:30	Half Hour	7.62	7.02↓	14.11	13.77↓
OXYG	SalesPrime 01:00	Half Hour	18.23	11.77↓	10.09	6.27↓
OXYG	SalesPrime 01:30	Half Hour	17.32	14.65↓	9.55	7.90↓
OXYG	Access 16:00	Half Hour	19.60	14.78↓	11.10	8.16↓
OXYG	Access 16:30	Half Hour	18.1	16.29↓	10.2	9.05↓
UNVSO	Daytime	Daily	34.0	35.12↑	13.6	21.4↑

In 14 experiments, bayesian exponential decay method with spike and slab prior achieves better prediction performance than basic bayesian exponential decay method! We notice that bayesian exponential decay method seems perform not so well on UNVSO network, even after adding spike and slab prior to the basic model setting. That might due to the fact that UNVSO network is noisy in the data collection procedure from Nielsen.

6 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 Marco for his patient guidance and valuable suggestion through the whole project. Also, I'm grateful to Juan who provided lots of helpful insights and ideas.

With a special mention to Jiabin, Sebastien and Xiao. It was fantastic to have the opportunity to work as a team with you.

Thanks to my intern cohort, Lisa!