Appendix A: Derivation of Variational Bayesian Inference

The evidence lower bound can be derived as

$$\begin{aligned}
J(\tilde{p}) &= \ln p(\mathbf{X}_{Z(1)}, \ldots, \mathbf{X}_{Z(n)}) - KL[\tilde{p}(\boldsymbol{\theta}_a, \mathbf{r}) || p(\boldsymbol{\theta}_a, \mathbf{r} | \mathbf{X}_{Z(1)}, ..., \mathbf{X}_{Z(n)})] \\
&= \ln p(\mathbf{X}_{Z(1)}, \ldots, \mathbf{X}_{Z(n)}) - \mathbb{E}_{\tilde{p}}[\ln \frac{\tilde{p}(\boldsymbol{\theta}_a, \mathbf{r})}{p(\boldsymbol{\theta}_a, \mathbf{r} | \mathbf{X}_{Z(1)}, ..., \mathbf{X}_{Z(n)})}] \\
&= \mathbb{E}_{\tilde{p}}[\ln p(\mathbf{X}_{Z(1)}, ..., \mathbf{X}_{Z(n)}, \boldsymbol{\theta}_a, \mathbf{r}) - \ln \tilde{p}(\boldsymbol{\theta}_a, \mathbf{r})] \\
&= \mathbb{E}_{\tilde{p}}[\sum_{t=1}^{n} \lambda_{nt} \ln p(\mathbf{X}_{Z(t)} | \boldsymbol{\theta}_a) + \ln p(\boldsymbol{\theta}_a | \mathbf{r}) + \ln p(\mathbf{r}) - \ln \tilde{p}(\boldsymbol{\theta}_a, \mathbf{r})].
\end{aligned}$$

The distribution of $\boldsymbol{\theta}_a$ is in spike-slab form. To be specific, with probability $1 - w_j$, $r_j = 0$, $\theta_{a,j} \sim N(0, v\sigma_j^2)$. On the other, with probability w_j, $r_j = 1, \theta_{a,j} \sim N(0, \sigma_j^2)$. So derive the prior distribution of $\boldsymbol{\theta}_a$ as

$$\ln p(\boldsymbol{\theta}_a | \mathbf{r}) = \ln \prod_{j=1}^{k_a} p(\theta_{a,j} | r_j) = \sum_{j=1}^{k_a} \ln p(\theta_{a,j} | r_j) = \sum_{j=1}^{k_a} (-\frac{1}{2} \ln \sigma_j^2 - \frac{\theta_{a,j}^2}{2\sigma_j^2}) r_j + (-\frac{1}{2} \ln(v\sigma_j^2) - \frac{\theta_{a,j}^2}{2v\sigma_j^2})(1 - r_j),$$

and

$$\ln p(\mathbf{r}) = \ln \prod_{j=1}^{k_a} p(r_j) = \sum_{j=1}^{k_a} \ln p(r_j) = \sum_{j=1}^{k_a} r_j \ln w_j + (1 - r_j) \ln(1 - w_j).$$

Then write the joint density of $\mathbf{X}_{Z(n)}, \boldsymbol{\theta}_a$ and $\mathbf{r}$ as

$$\begin{aligned}
&\ln p(\mathbf{X}_{Z(1)}, \ldots, \mathbf{X}_{Z(n)}, \boldsymbol{\theta}_a, \mathbf{r}) = \sum_{t=1}^{n} \lambda_{nt} \ln p(\mathbf{X}_{Z(t)} | \boldsymbol{\theta}_a) + \ln p(\boldsymbol{\theta}_a | \mathbf{r}) + \ln p(\mathbf{r}) \\
&= \sum_{t=1}^{n} \lambda_{nt} \big((\mathbf{X}_{Z(t)} \oslash a(\boldsymbol{\phi}_{Z(t)}))' (\mathbf{B}_{a,Z(t)} \boldsymbol{\theta}_a + \mathbf{B}_{b,Z(t)} \boldsymbol{\theta}_t) - \mathbf{1}_m' b(\mathbf{B}_{a,Z(t)} \boldsymbol{\theta}_a + \mathbf{B}_{b,Z(t)} \boldsymbol{\theta}_t) \oslash a(\boldsymbol{\phi}_{Z(t)}) + \mathbf{1}_m' c(\mathbf{X}_{Z(t)}, \boldsymbol{\phi}_{Z(t)}) \big) \\
&+ \sum_{j=1}^{k_a} (-\frac{1}{2} \ln \sigma_j^2 - \frac{\theta_{a,j}^2}{2\sigma_j^2}) r_j + (-\frac{1}{2} \ln(v\sigma_j^2) - \frac{\theta_{a,j}^2}{2v\sigma_j^2})(1 - r_j) + r_j \ln w_j + (1 - r_j) \ln(1 - w_j).
\end{aligned}$$

To be coincident, model the posterior distribution of $\boldsymbol{\theta}_a$ in the spike-slab form too. With probability $1 - \alpha_j$, $r_j = 0$, and then $\theta_{a,j} \sim N(0, vs_j^2)$. With probability α_j, $r_j = 1$ and then $\theta_{a,j} \sim N(\mu_j, s_j^2)$.

$$\begin{aligned}
\mathbb{E}_{\tilde{p}}[\ln p(\boldsymbol{\theta}_a, \mathbf{r})] &= \mathbb{E}_{\tilde{p}}[\sum_{j=1}^{k_a} (-\frac{1}{2} \ln \sigma_j^2 - \frac{\theta_{a,j}^2}{2\sigma_j^2}) r_j + (-\frac{1}{2} \ln(v\sigma_j^2) - \frac{\theta_{a,j}^2}{2v\sigma_j^2})(1 - r_j) + r_j \ln w_j + (1 - r_j) \ln(1 - w_j)] \\
&= \sum_{j=1}^{k_a} (-\frac{1}{2} \ln \sigma_j^2 - \frac{\mu_j^2 + s_j^2}{2\sigma_j^2}) \alpha_j + (-\frac{1}{2} \ln(v\sigma_j^2) - \frac{vs_j^2}{2v\sigma_j^2})(1 - \alpha_j) + \alpha_j \ln w_j + (1 - \alpha_j) \ln(1 - w_j),
\end{aligned}$$

and

$$\mathbb{E}_{\tilde{p}}[\ln \tilde{p}(\boldsymbol{\theta}_a, \mathbf{r})] = \sum_{j=1}^{k_a} (-\frac{1}{2}\ln s_j^2 - \frac{1}{2})\alpha_j + (-\frac{1}{2}\ln(vs_j^2) - \frac{1}{2})(1-\alpha_j) + \alpha_j \ln \alpha_j + (1-\alpha_j)\ln(1-\alpha_j).$$

We finally derive that

$$\begin{aligned}
J(\tilde{p}) &= \mathbb{E}_{\tilde{p}}[\sum_{t=1}^{n} \lambda_{nt} \ln p(\mathbf{X}_{Z(t)}|\boldsymbol{\theta}_a) + \ln p(\boldsymbol{\theta}_a, \mathbf{r}) - \ln \tilde{p}(\boldsymbol{\theta}_a, \mathbf{r})] \\
&= \sum_{t=1}^{n} \lambda_{nt} \big((\mathbf{X}_{Z(t)} \oslash a(\boldsymbol{\phi}_{Z(t)}))' (\mathbf{B}_{a,Z(t)} \boldsymbol{\mu}_a + \mathbf{B}_{b,Z(t)} \boldsymbol{\theta}_t) \big) \\
&+ \sum_{j=1}^{k_a} \big(\frac{1}{2} - \frac{s_j^2}{2\sigma_j^2} + \frac{1}{2}\ln \frac{s_j^2}{\sigma_j^2} + (\ln w_j - \ln \alpha_j - \frac{\mu_j^2}{2\sigma_j^2})\alpha_j + (\ln(1-w_j) - \ln(1-\alpha_j))(1-\alpha_j) \big) \\
&- \sum_{t=1}^{n} \lambda_{nt} \big(\mathbf{1}_m' \mathbb{E}_{\tilde{p}}[b(\mathbf{B}_{a,Z(t)} \boldsymbol{\theta}_a + \mathbf{B}_{b,Z(t)} \boldsymbol{\theta}_t) \oslash a(\boldsymbol{\phi}_{Z(t)})] + \mathbf{1}_m' c(\mathbf{X}_{Z(t)}, \boldsymbol{\phi}_{Z(t)}) \big),
\end{aligned}$$

where $\boldsymbol{\mu}_a = \boldsymbol{\mu} \circ \boldsymbol{\alpha}$. To simplify the notations, assume $d = \sum_{t=1}^{n} \lambda_{nt} \mathbf{1}_m' \mathbb{E}_{\tilde{p}}[(b(\mathbf{B}_{a,Z(t)} \boldsymbol{\theta}_a + \mathbf{B}_{b,Z(t)} \boldsymbol{\theta}_t) \oslash a(\boldsymbol{\phi}_{Z(t)}))]$. We can derive the derivatives of the parameters as

$$\begin{aligned}
\frac{\partial J(\tilde{p})}{\partial \mu_j} &= \sum_{t=1}^{n} \lambda_{nt} \big((\mathbf{X}_{Z(t)} \oslash a(\boldsymbol{\phi}_{Z(t)}))' \mathbf{B}_{a,Z(t)j} \alpha_j \big) - \frac{\alpha_j \mu_j}{\sigma_j^2} - \frac{\partial d}{\partial \mu_j}, \\
\frac{\partial J(\tilde{p})}{\partial \alpha_j} &= \sum_{t=1}^{n} \lambda_{nt} \big((\mathbf{X}_{Z(t)} \oslash a(\boldsymbol{\phi}_{Z(t)}))' \mathbf{B}_{a,Z(t)j} \mu_j \big) + \ln w_j - \frac{\mu_j^2}{2\sigma_j^2} - \ln \alpha_j - \ln(1-w_j) + \ln(1-\alpha_j) - \frac{\partial d}{\partial \alpha_j}, \\
\frac{\partial J(\tilde{p})}{\partial s_j} &= -\frac{s_j}{\sigma_j^2} + \frac{1}{s_j} - \frac{\partial d}{\partial s_j}.
\end{aligned}$$

Besides $\tilde{p}(\boldsymbol{\theta}_a|\mathbf{r})$ and $\tilde{p}(\mathbf{r})$, $\boldsymbol{\theta}_n$ also needs to be estimated for each time point. Here we adopt the maximum a posteriori estimation method. Since we assume in prior $p_0(\boldsymbol{\theta}_n) \sim N(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0)$, we obtain the posterior of $\boldsymbol{\theta}_n$ as $\tilde{p}(\boldsymbol{\theta}_n) \propto p(\mathbf{X}_{Z(n)}|\boldsymbol{\eta}_{Z(n)}) p_0(\boldsymbol{\theta}_n) = p(\mathbf{X}_{Z(n)}|\boldsymbol{\theta}_n, \boldsymbol{\mu}_a) p_0(\boldsymbol{\theta}_n)$. Then the maximum a posteriori estimation can be achieved via gradient ascent method as well:

$$\frac{\partial \ln \tilde{p}(\boldsymbol{\theta}_n)}{\partial \boldsymbol{\theta}_n} = -\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) + \mathbf{B}_{bZ(n)}' \big((\mathbf{X}_{Z(n)} - b'(\mathbf{B}_{a,Z(n)} \boldsymbol{\mu}_a + \mathbf{B}_{b,Z(n)} \boldsymbol{\theta}_n)) \oslash a(\boldsymbol{\phi}_{Z(n)}) \big).$$

Take $\mathbf{X}$ as Gaussian, Poisson and binomial distributions as examples and derive the specific variational inference for them.

If $\mathbf{X}$ follows Gaussian distribution, we formulate the generalized linear model as

$$p(X_{Z(n)_i}|\eta_{Z(n)_i}) = \exp\big(\frac{X_{Z(n)_i}\eta_{Z(n)_i} - \frac{1}{2}(\eta_{Z(n)_i})^2}{\sigma_e^2} - \frac{X_{Z(n)_i}^2}{2\sigma_e^2} - \frac{1}{2}\ln(2\pi\sigma_e^2) \big), \quad i = 1, \ldots, m,$$

$$\boldsymbol{\eta}_{Z(n)} = \mathbf{B}_{a,Z(n)}\boldsymbol{\theta}_a + \mathbf{B}_{b,Z(n)}\boldsymbol{\theta}_n.$$

And derive the inference as

$$\begin{aligned}
d &= \sum_{t=1}^{n} \frac{\lambda_{nt}}{\sigma_e^2} \mathbb{E}_{\tilde{p}}[\frac{1}{2}(\mathbf{B}_{a,Z(t)}\boldsymbol{\theta}_a + \mathbf{B}_{b,Z(t)}\boldsymbol{\theta}_t)'(\mathbf{B}_{a,Z(t)}\boldsymbol{\theta}_a + \mathbf{B}_{b,Z(t)}\boldsymbol{\theta}_t)] \\
&= \sum_{t=1}^{n} \frac{\lambda_{nt}}{2\sigma_e^2} \big(\mathbb{E}_{\tilde{p}}[\boldsymbol{\theta}_a' \mathbf{B}_{a,Z(t)}' \mathbf{B}_{a,Z(t)}\boldsymbol{\theta}_a + 2\boldsymbol{\theta}_t' \mathbf{B}_{b,Z(t)}' \mathbf{B}_{a,Z(t)}\boldsymbol{\theta}_a + \boldsymbol{\theta}_t' \mathbf{B}_{b,Z(t)} \mathbf{B}_{b,Z(t)}\boldsymbol{\theta}_t]\big) \\
&= \sum_{t=1}^{n} \frac{\lambda_{nt}}{2\sigma_e^2} \big(\sum_{j=1}^{k_a} \mathbf{B}_{a,Z(t)j}' \mathbf{B}_{a,Z(t)j} \alpha_j (\mu_j^2 + s_j^2) + 2 \sum_{j=1}^{k_a} \sum_{k \neq j} \mathbf{B}_{a,Z(t)j}' \mathbf{B}_{a,Z(t)k} \mu_k \mu_j \alpha_k \alpha_j \\
&+ 2 \sum_{j=1}^{k_a} \boldsymbol{\theta}_t' \mathbf{B}_{b,Z(t)}' \mathbf{B}_{a,Z(t)j} \alpha_j \mu_j + \boldsymbol{\theta}_t' \mathbf{B}_{b,Z(t)}' \mathbf{B}_{b,Z(t)} \boldsymbol{\theta}_t \big).
\end{aligned}$$

Take the derivatives of $J(\tilde{p})$ with respect to μ_j, α_j, s_j and $\boldsymbol{\theta}_n$ and set the partial derivatives to zero to solve μ_j, α_j, s_j and $\boldsymbol{\theta}_n$:

$$\begin{aligned}
\frac{\partial d}{s_j} &= \sum_{t=1}^{n} \frac{\lambda_{nt}}{\sigma_e^2} \mathbf{B}_{a,Z(t)j}' \mathbf{B}_{a,Z(t)j} s_j \alpha_j, \\
\frac{\partial d}{\partial \mu_j} &= \sum_{t=1}^{n} \frac{\lambda_{nt}}{\sigma_e^2} \big(\mathbf{B}_{a,Z(t)j}' \mathbf{B}_{a,Z(t)j} \mu_j \alpha_j + \sum_{k \neq j} \mathbf{B}_{a,Z(t)j}' \mathbf{B}_{a,Z(t)k} \mu_k \alpha_j \alpha_k + \boldsymbol{\theta}_t' \mathbf{B}_{b,Z(t)}' \mathbf{B}_{a,Z(t)j} \alpha_j \big), \\
\frac{\partial d}{\partial \alpha_j} &= \sum_{t=1}^{n} \frac{\lambda_{nt}}{\sigma_e^2} \big(\frac{1}{2} \mathbf{B}_{a,Z(t)j}' \mathbf{B}_{a,Z(t)j} (\mu_j^2 + s_j^2) + \sum_{k \neq j} \mathbf{B}_{a,Z(t)j}' \mathbf{B}_{a,Z(t)k} \mu_j \mu_k \alpha_k + \boldsymbol{\theta}_t' \mathbf{B}_{b,Z(t)}' \mathbf{B}_{a,Z(t)j} \mu_j \big).
\end{aligned}$$

Update the gradient of $\boldsymbol{\theta}_n$

$$\begin{aligned}
\frac{\partial \ln \tilde{p}(\boldsymbol{\theta}_n)}{\partial \boldsymbol{\theta}_n} &= -\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) + \frac{1}{\sigma_e^2} \mathbf{B}_{b,Z(n)}' (\mathbf{X}_{Z(n)} - \mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_a - \mathbf{B}_{b,Z(n)}\boldsymbol{\theta}_n) = \mathbf{0}, \\
\tilde{\boldsymbol{\theta}}_n &= (\frac{1}{\sigma_e^2} \mathbf{B}_{b,Z(n)}' \mathbf{B}_{b,Z(n)} + \boldsymbol{\Sigma}_0^{-1})^{-1} \big(\frac{1}{\sigma_e^2} \mathbf{B}_{b,Z(n)}' (\mathbf{X}_{Z(n)} - \mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_a) + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\theta}_0 \big).
\end{aligned}$$

If $\mathbf{X}$ follows Poisson distribution, we formulate the generalized linear model as

$$\begin{aligned}
p(X_{Z(n)i}) &= \exp\big(X_{Z(n)_i}\eta_{Z(n)_i} - \exp(\eta_{Z(n)_i}) - \ln\Gamma(X_{Z(n)_i}+1)\big), \quad i = 1, \ldots, m, \\
\boldsymbol{\eta}_{Z(n)} &= \mathbf{B}_{a,Z(n)}\boldsymbol{\theta}_a + \mathbf{B}_{b,Z(n)}\boldsymbol{\theta}_n,
\end{aligned}$$

where Γ denotes the Gamma function. And derive the inference for it:

$$
\begin{aligned}
d &= \sum_{t=1}^{n} \lambda_{nt} \mathbf{1}'_m \big(\mathbb{E}_{\tilde{p}}(\exp(\mathbf{B}_{a,Z(t)}\boldsymbol{\theta}_a + \mathbf{B}_{b,Z(t)}\boldsymbol{\theta}_t))\big) \\
&\approx \sum_{t=1}^{n} \lambda_{nt} \big(\mathbf{1}'_m \exp(\mathbf{B}_{a,Z(t)}\boldsymbol{\mu}_a + \mathbf{B}_{b,Z(t)}\boldsymbol{\theta}_t) + \frac{1}{2}\operatorname{Tr}\big(\bar{\operatorname{diag}}(\exp(\mathbf{B}_{a,Z(t)}\boldsymbol{\mu}_a + \mathbf{B}_{b,Z(t)}\boldsymbol{\theta}_t))\mathbf{B}_{a,Z(t)}\mathbf{K}\mathbf{B}'_{a,Z(t)}\big)\big),
\end{aligned}
$$

where $\bar{\operatorname{diag}}(\cdot)$ represents constructing a diagonal matrix with the vector in the brackets, and $\mathbf{K} = \bar{\operatorname{diag}}(\boldsymbol{\alpha} \circ \mathbf{s} \circ \mathbf{s}) \in \mathcal{R}^{k_a \times k_a}$. The approximated step is based on Taylor expansions for the moments of functions of random variables. For example, since $\mathbb{E}[f(\mathbf{X})] = \mathbb{E}\left[f(\boldsymbol{\mu}_\mathbf{X} + (\mathbf{X} - \boldsymbol{\mu}_\mathbf{X}))\right] \approx \mathbb{E}\left[f(\boldsymbol{\mu}_\mathbf{X}) + f'(\boldsymbol{\mu}_\mathbf{X})(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X}) + \frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})^T H_f(\boldsymbol{\mu}_\mathbf{X})(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})\right]$, where $H_f(\cdot)$ is the Hessian matrix, $\mathbb{E}[f(\mathbf{X})] \approx f(\boldsymbol{\mu}_\mathbf{X}) + \frac{1}{2}\operatorname{Tr}(H_f(\boldsymbol{\mu}_\mathbf{X})\boldsymbol{\Sigma}_\mathbf{X})$ on account of $\mathbb{E}[\mathbf{X} - \boldsymbol{\mu}_\mathbf{X}] = \mathbf{0}$. Take the derivatives of d:

$$
\begin{aligned}
\frac{\partial d}{\partial \boldsymbol{\mu}} &= \sum_{t=1}^{n} \lambda_{nt} \big(\mathbf{B}'_{a,Z(t)} \exp(\mathbf{B}_{a,Z(t)}\boldsymbol{\mu}_a + \mathbf{B}_{b,Z(t)}\boldsymbol{\theta}_t) \circ \boldsymbol{\alpha} + \frac{1}{2}\mathbf{B}'_{a,Z(t)}(\operatorname{diag}(\mathbf{B}_{a,Z(t)}\mathbf{K}\mathbf{B}'_{a,Z(t)}) \circ \exp(\mathbf{B}_{a,Z(t)}\boldsymbol{\mu}_a + \mathbf{B}_{b,Z(t)}\boldsymbol{\theta}_t)) \\
&\quad \circ \boldsymbol{\alpha}\big), \\
\frac{\partial d}{\partial \boldsymbol{\alpha}} &= \sum_{t=1}^{n} \lambda_{nt} \big(\mathbf{B}'_{a,Z(t)} \exp(\mathbf{B}_{a,Z(t)}\boldsymbol{\mu}_a + \mathbf{B}_{b,Z(t)}\boldsymbol{\theta}_t) \circ \boldsymbol{\mu} + \frac{1}{2}\mathbf{B}'_{a,Z(t)}(\operatorname{diag}(\mathbf{B}_{a,Z(t)}\mathbf{K}\mathbf{B}'_{a,Z(t)}) \circ \exp(\mathbf{B}_{a,Z(t)}\boldsymbol{\mu}_a + \mathbf{B}_{b,Z(t)}\boldsymbol{\theta}_t)) \\
&\quad \circ \boldsymbol{\mu} + \frac{1}{2}\operatorname{diag}(\mathbf{B}'_{a,Z(t)}\bar{\operatorname{diag}}(\exp(\mathbf{B}_{a,Z(t)}\boldsymbol{\mu}_a + \mathbf{B}_{b,Z(t)}\boldsymbol{\theta}_t))\mathbf{B}_{a,Z(t)}) \circ \mathbf{s} \circ \mathbf{s}\big), \\
\frac{\partial d}{\partial \mathbf{s}} &= \sum_{t=1}^{n} \lambda_{nt} \operatorname{diag}(\mathbf{B}'_{a,Z(t)}\bar{\operatorname{diag}}(\exp(\mathbf{B}_{a,Z(t)}\boldsymbol{\mu}_a + \mathbf{B}_{b,Z(t)}\boldsymbol{\theta}_t))\mathbf{B}_{a,Z(t)}) \circ \boldsymbol{\alpha} \circ \mathbf{s}.
\end{aligned}
$$

Update the gradient of $\boldsymbol{\theta}_n$

$$
\frac{\partial \ln \tilde{p}(\boldsymbol{\theta}_n)}{\partial \boldsymbol{\theta}_n} = -\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) + \mathbf{B}'_{b,Z(n)}(\mathbf{X}_{Z(n)} - \exp(\mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_a + \mathbf{B}_{b,Z(n)}\boldsymbol{\theta}_n)).
$$

If $\mathbf{X}$ follows binomial distribution with known number of trials of N, we formulated the generalized linear model as

$$
\begin{aligned}
p(X_{Z(n)_i}) &= \exp\left(X_{Z(n)_i}\eta_{Z(n)_i} - N\log(1 + \exp(\eta_{Z(n)_i})) + \binom{N}{X_{Z(n)_i}}\right), \quad i = 1, \ldots, m, \\
\boldsymbol{\eta}_{Z(n)} &= \mathbf{B}_{a,Z(n)}\boldsymbol{\theta}_a + \mathbf{B}_{b,Z(n)}\boldsymbol{\theta}_n.
\end{aligned}
$$

And derive the inference for it:

$$
d = \sum_{t=1}^{n} \lambda_{nt} \mathbf{1}'_m \mathbb{E}_{\tilde{p}}[N \log\left(\mathbf{1}_m + \exp(\mathbf{B}_{a,Z(t)}\boldsymbol{\theta}_a + \mathbf{B}_{b,Z(t)}\boldsymbol{\theta}_t)\right)]
$$

$$\approx \sum_{t=1}^{n} \lambda_{nt} \mathbf{N}'_m \log\big(\mathbf{1}_m + \exp(\mathbf{B}_{a,Z(t)}\boldsymbol{\mu}_a + \mathbf{B}_{b,Z(t)}\boldsymbol{\theta}_t)\big)$$
$$+ \frac{1}{2}\operatorname{Tr}(N\bar{\operatorname{diag}}\big(\exp(\mathbf{B}_{a,Z(t)}\boldsymbol{\mu}_a + \mathbf{B}_{b,Z(t)}\boldsymbol{\theta}_t) \oslash (\mathbf{1}_m + \exp(\mathbf{B}_{a,Z(t)}\boldsymbol{\mu}_a + \mathbf{B}_{b,Z(t)}\boldsymbol{\theta}_t))^2\big)\mathbf{B}_{a,Z(t)}\mathbf{K}\mathbf{B}'_{a,Z(t)}).$$

Take the derivatives of d:

$$\boldsymbol{\eta}_t = \mathbf{B}_{a,Z(t)}\boldsymbol{\mu}_a + \mathbf{B}_{b,Z(t)}\boldsymbol{\theta}_t,$$
$$\frac{\partial d}{\partial \boldsymbol{\mu}} = \sum_{t=1}^{n} \lambda_{nt}\big(N\mathbf{B}'_{a,Z(t)}(\exp(\boldsymbol{\eta}_t) \oslash (\mathbf{1}_m + \exp(\boldsymbol{\eta}_t))) \circ \boldsymbol{\alpha} + \frac{1}{2}N\mathbf{B}'_{a,Z(t)}(\operatorname{diag}(\mathbf{B}_{a,Z(t)}\mathbf{K}\mathbf{B}'_{a,Z(t)}) \circ \exp(\boldsymbol{\eta}_t)$$
$$\oslash (\mathbf{1}_m + \exp(\boldsymbol{\eta}_t))^2) \circ \boldsymbol{\alpha} - N\mathbf{B}'_{a,Z(t)}(\operatorname{diag}(\mathbf{B}_{a,Z(t)}\mathbf{K}\mathbf{B}'_{a,Z(t)}) \circ \exp(\boldsymbol{\eta}_t)^2 \oslash (\mathbf{1}_m + \exp(\boldsymbol{\eta}_t))^3) \circ \boldsymbol{\alpha}\big),$$
$$\frac{\partial d}{\partial \boldsymbol{\alpha}} = \sum_{t=1}^{n} \lambda_{nt}\big(N\mathbf{B}'_{a,Z(t)}(\exp(\boldsymbol{\eta}_t) \oslash (\mathbf{1}_m + \exp(\boldsymbol{\eta}_t))) \circ \boldsymbol{\mu} + \frac{1}{2}N\mathbf{B}'_{a,Z(t)}(\operatorname{diag}(\mathbf{B}_{a,Z(t)}\mathbf{K}\mathbf{B}'_{a,Z(t)}) \circ \exp(\boldsymbol{\eta}_t)$$
$$\oslash (\mathbf{1}_m + \exp(\boldsymbol{\eta}_t))^2) \circ \boldsymbol{\mu} - N\mathbf{B}'_{a,Z(t)}(\operatorname{diag}(\mathbf{B}_{a,Z(t)}\mathbf{K}\mathbf{B}'_{a,Z(t)}) \circ \exp(\boldsymbol{\eta}_t)^2 \oslash (\mathbf{1}_m + \exp(\boldsymbol{\eta}_t))^3) \circ \boldsymbol{\mu}$$
$$+ \frac{1}{2}N\operatorname{diag}(\mathbf{B}'_{a,Z(t)}\bar{\operatorname{diag}}(\exp(\boldsymbol{\eta}_t) \oslash (\mathbf{1}_m + \exp(\boldsymbol{\eta}_t))^2)\mathbf{B}_{a,Z(t)}) \circ \mathbf{s}^2\big),$$
$$\frac{\partial d}{\partial \mathbf{s}} = \sum_{t=1}^{n} \lambda_{nt}\big(N\operatorname{diag}(\mathbf{B}'_{a,Z(t)}\bar{\operatorname{diag}}(\exp(\boldsymbol{\eta}_t) \oslash (\mathbf{1}_m + \exp(\boldsymbol{\eta}_t))^2)\mathbf{B}_{a,Z(t)}) \circ \boldsymbol{\alpha} \circ \mathbf{s}\big).$$

Update the gradient of $\boldsymbol{\theta}_n$

$$\frac{\partial \ln \tilde{p}(\boldsymbol{\theta}_n)}{\partial \boldsymbol{\theta}_n} = -\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) + \mathbf{B}'_{b,Z(n)}\big((\mathbf{X}_{Z(n)} - N\exp(\mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_a + \mathbf{B}_{b,Z(n)}\boldsymbol{\theta}_n) \oslash (\mathbf{1}_m + \exp(\mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_a$$
$$+ \mathbf{B}_{b,Z(n)}\boldsymbol{\theta}_n))\big)).$$

Appendix B: Deviation of Test Statistics

Simple Likelihood Ratio Test (SLRT)

For simple likelihood ratio test,

$$\Lambda^{SLRT}_{\boldsymbol{\theta}_a}(\mathbf{X}_{Z(n)}) = 2\ln\frac{p(\mathbf{X}_{Z(n)}|H_1)}{p(\mathbf{X}_{Z(n)}|H_0)}$$
$$= (\mathbf{X}_{Z(n)} \oslash a(\boldsymbol{\phi}_{Z(n)}))'\mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_a - \mathbf{1}'_m(b(\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n + \mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_a) \oslash a(\boldsymbol{\phi}_{Z(n)})) + \mathbf{1}'_m(b(\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n) \oslash a(\boldsymbol{\phi}_{Z(n)})).$$

If $\mathbf{X}$ follows Gaussian distribution, it can be defined specifically as

$$\Lambda^{SLRT}_{\boldsymbol{\theta}_a}(\mathbf{X}_{Z(n)}) = \frac{1}{\sigma_e^2}\big(\mathbf{X}'_{Z(n)}\mathbf{B}_{a,Z(t)}\boldsymbol{\mu}_a - \frac{1}{2}(\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n + \mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_a)'(\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n + \mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_a) + \frac{1}{2}\tilde{\boldsymbol{\theta}}'_n\mathbf{B}'_{b,Z(n)}\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n\big),$$

$$= 2(\mathbf{X}_{Z(n)} - \mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)'\mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_a - \boldsymbol{\mu}_a'\mathbf{B}_{a,Z(n)}'\mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_a.$$

If $\mathbf{X}$ follows Poisson distribution, it can be defined specifically as

$$\Lambda_{\boldsymbol{\theta}_a}^{SLRT}(\mathbf{X}_{Z(n)}) = \mathbf{X}_{Z(n)}'\mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_a + \mathbf{1}_m'\exp(\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n) - \mathbf{1}_m'\exp(\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n + \mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_a).$$

If $\mathbf{X}$ follows binomial distribution, it can be defined specifically as

$$\Lambda_{\boldsymbol{\theta}_a}^{SLRT}(\mathbf{X}_{Z(n)}) = \mathbf{X}_{Z(n)}'\mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_a + \mathbf{N}_m'\ln(\mathbf{1}_m + \exp(\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)) - \mathbf{N}_m'\ln(\mathbf{1}_m + \exp(\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n + \mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_a)).$$

General Likelihood Ratio Test (GLRT)

For general likelihood ratio test,

$$\Lambda_{\boldsymbol{\theta}_a}^{GLRT}(\mathbf{X}_{Z(n)}) = 2\max_{\boldsymbol{\theta}_a}\sum_{t=1}^{n}\lambda_{nt}\ln\frac{p(\mathbf{X}_{Z(t)}|H_1)}{p(\mathbf{X}_{Z(t)}|H_0)}$$
$$= \sum_{t=1}^{n}\lambda_{nt}\big((\mathbf{X}_{Z(t)} \oslash a(\boldsymbol{\phi}_{Z(t)}))'\mathbf{B}_{a,Z(t)}\hat{\boldsymbol{\theta}}_a - \mathbf{1}_m'(b(\mathbf{B}_{b,Z(t)}\tilde{\boldsymbol{\theta}}_t + \mathbf{B}_{a,Z(t)}\hat{\boldsymbol{\theta}}_a) \oslash a(\boldsymbol{\phi}_{Z(t)})) + \mathbf{1}_m'(b(\mathbf{B}_{b,Z(t)}\tilde{\boldsymbol{\theta}}_t) \oslash a(\boldsymbol{\phi}_{Z(t)}))\big).$$

$\hat{\boldsymbol{\theta}}_a$ maximizes the weighted log-likelihood ratio, i.e., $\sum_{t=1}^{n}\lambda_{nt}\ln\frac{p(\mathbf{X}_{Z(t)}|H_1)}{p(\mathbf{X}_{Z(t)}|H_0)}$ and we use Newton's method to compute it.

If $\mathbf{X}$ follows Gaussian distribution, it can be defined specifically as

$$\Lambda_{\boldsymbol{\theta}_a}^{GLRT}(\mathbf{X}_{Z(n)}) = \sum_{t=1}^{n}\lambda_{nt}\big(2(\mathbf{X}_{Z(t)} - \mathbf{B}_{b,Z(t)}\tilde{\boldsymbol{\theta}}_t)'\mathbf{B}_{a,Z(t)}\hat{\boldsymbol{\theta}}_a - \hat{\boldsymbol{\theta}}_a'\mathbf{B}_{a,Z(t)}'\mathbf{B}_{a,Z(t)}\hat{\boldsymbol{\theta}}_a\big).$$

If $\mathbf{X}$ follows Poisson distribution, it can be defined specifically as

$$\Lambda_{\boldsymbol{\theta}_a}^{GLRT}(\mathbf{X}_{Z(n)}) = \sum_{t=1}^{n}\lambda_{nt}\big(\mathbf{X}_{Z(t)}'\mathbf{B}_{a,Z(t)}\hat{\boldsymbol{\theta}}_a + \mathbf{1}_m'\exp(\mathbf{B}_{b,Z(t)}\tilde{\boldsymbol{\theta}}_t) - \mathbf{1}_m'\exp(\mathbf{B}_{b,Z(t)}\tilde{\boldsymbol{\theta}}_t + \mathbf{B}_{a,Z(t)}\hat{\boldsymbol{\theta}}_a)\big).$$

If $\mathbf{X}$ follows binomial distribution, it can be defined specifically as

$$\Lambda_{\boldsymbol{\theta}_a}^{GLRT}(\mathbf{X}_{Z(n)}) = \sum_{t=1}^{n}\lambda_{nt}\big(\mathbf{X}_{Z(t)}'\mathbf{B}_{a,Z(t)}\hat{\boldsymbol{\theta}}_a + \mathbf{N}_m'\ln(\mathbf{1}_m + \exp(\mathbf{B}_{b,Z(t)}\tilde{\boldsymbol{\theta}}_t)) - \mathbf{N}_m'\ln(\mathbf{1}_m + \exp(\mathbf{B}_{b,Z(t)}\tilde{\boldsymbol{\theta}}_t + \mathbf{B}_{a,Z(t)}\hat{\boldsymbol{\theta}}_a))\big).$$

Posterior Bayes Factor (PBF)

For posterior Bayes factor, reformulate the hypothesis test as

$$H_0: \boldsymbol{\theta}_a = \mathbf{0}, \quad H_1: \boldsymbol{\theta}_a \sim N(\boldsymbol{\mu}_{\mathbf{r}}, \mathbf{K}_{\mathbf{r}}), \mathbf{r} \in \mathbb{R},$$

where $\mathbb{R}$ is set of all possible $\mathbf{r}$. $\boldsymbol{\mu}_{\mathbf{r}} = \boldsymbol{\mu} \circ \mathbf{r}$. $\mathbf{K}_{\mathbf{r}} = \bar{\text{diag}}(((1-r_1)v + r_1)s_1^2, \ldots, ((1-r_{k_a})v + r_{k_a})s_{k_a}^2)$. Compute $\Lambda_{\boldsymbol{\theta}_a}^{PBF}(\mathbf{X}_{Z(n)})$ as

$$\Lambda_{\boldsymbol{\theta}_a}^{PBF}(\mathbf{X}_{Z(n)}) = \frac{L_1}{L_0} = \frac{\sum_{\mathbf{r}\in\mathbb{R}} \tilde{p}(\mathbf{r}) \int \tilde{p}(\boldsymbol{\theta}_a|\mathbf{r}) p(\mathbf{X}_{Z(n)}|\tilde{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_a) d\boldsymbol{\theta}_a}{p(\mathbf{X}_{Z(n)}|\tilde{\boldsymbol{\theta}}_n)},$$

with

$$L_0 = p(\mathbf{X}_{Z(n)}|\tilde{\boldsymbol{\theta}}_n) = \exp\big((\mathbf{X}_{Z(n)} \oslash a(\boldsymbol{\phi}_{Z(t)}))' \mathbf{B}_{b,Z(n)} \tilde{\boldsymbol{\theta}}_n - \mathbf{1}_m' b(\mathbf{B}_{b,Z(n)} \tilde{\boldsymbol{\theta}}_n) \oslash a(\boldsymbol{\phi}_{Z(t)}) - \mathbf{1}_m' c(\mathbf{X}_{Z(n)}, \boldsymbol{\phi}_{Z(t)})\big).$$

To derive L_1, we need to simplify the following integral,

$$\int \tilde{p}(\boldsymbol{\theta}_a|\mathbf{r}) p(\mathbf{X}_{Z(n)}|\tilde{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_a) d\boldsymbol{\theta}_a = (2\pi)^{-k_a/2} |\mathbf{K}_{\mathbf{r}}|^{-1/2} \int \exp(-\frac{1}{2}(\boldsymbol{\theta}_a - \boldsymbol{\mu}_{\mathbf{r}})' \mathbf{K}_{\mathbf{r}}^{-1} (\boldsymbol{\theta}_a - \boldsymbol{\mu}_{\mathbf{r}}) + l) d\boldsymbol{\theta}_a,$$

where $l = (\mathbf{X}_{Z(n)} \oslash a(\boldsymbol{\phi}_{Z(t)}))' \boldsymbol{\eta}_{Z(n)} - \mathbf{1}_m' (b(\boldsymbol{\eta}_{Z(n)}) \oslash a(\boldsymbol{\phi}_{Z(t)})) + \mathbf{1}_m' c(\mathbf{X}_{Z(n)}, \boldsymbol{\phi}_{Z(t)})$.

To apply Laplace's method, denote $h(\boldsymbol{\theta}_a) = -\frac{1}{2}(\boldsymbol{\theta}_a - \boldsymbol{\mu}_{\mathbf{r}})' \mathbf{K}_{\mathbf{r}}^{-1} (\boldsymbol{\theta}_a - \boldsymbol{\mu}_{\mathbf{r}}) + l$. Assume $\tilde{\boldsymbol{\theta}}_a$ is the maximizer of $h(\boldsymbol{\theta}_a)$. The following letters with a tilde identification are all evaluated at $\tilde{\boldsymbol{\theta}}_a$. We find derivatives up to the sixth order:

1) $h(\tilde{\boldsymbol{\theta}}_a) = \tilde{l} - \frac{1}{2}(\tilde{\boldsymbol{\theta}}_a - \boldsymbol{\mu}_{\mathbf{r}})' \mathbf{K}_{\mathbf{r}}^{-1} (\tilde{\boldsymbol{\theta}}_a - \boldsymbol{\mu}_{\mathbf{r}})$.

2) $h^{(1)}(\tilde{\boldsymbol{\theta}}_a) = \tilde{l}^{(1)} - \mathbf{K}_{\mathbf{r}}^{-1} (\tilde{\boldsymbol{\theta}}_a - \boldsymbol{\mu}_{\mathbf{r}})$, where

$$\tilde{l}^{(1)} = \frac{\partial l}{\partial \boldsymbol{\theta}_a}\bigg|_{\boldsymbol{\theta}_a = \tilde{\boldsymbol{\theta}}_a} = \mathbf{B}_{a,Z(n)}' (\mathbf{X}_{Z(n)} - \tilde{\boldsymbol{\pi}}_{Z(n)}) \oslash a(\boldsymbol{\phi}_{Z(t)}) = \mathbf{B}_{a,Z(n)}' \tilde{\mathbf{W}}_{Z(n)} (\mathbf{X}_{Z(n)}^{\star} - \tilde{\boldsymbol{\eta}}_{Z(n)}).$$

Here $\boldsymbol{\pi}_{Z(n)} = \mathbb{E}[\mathbf{X}_{Z(n)}]$. $\mathbf{X}_{Z(n)}^{\star} = \tilde{\mathbf{W}}_{Z(n)}^{-1} (\mathbf{X}_{Z(n)} - \tilde{\boldsymbol{\pi}}_{Z(n)}) \oslash a(\boldsymbol{\phi}_{Z(t)}) + \tilde{\boldsymbol{\eta}}_{Z(n)}$,
$\tilde{\mathbf{W}}_{Z(n)} = \text{diag}(\tilde{w}_{Z(n)1}, \tilde{w}_{Z(n)2}, \ldots \tilde{w}_{Z(n)m})$, and $\tilde{w}_{Z(n)i} = \frac{d\tilde{\pi}_{Z(n)i}}{d\tilde{\eta}_{Z(n)i}}$.

3) $h^{(2)}(\tilde{\boldsymbol{\theta}}_a) = \tilde{l}^{(2)} - \mathbf{K}_{\mathbf{r}}^{-1} = -\mathbf{B}_{a,Z(n)}' \tilde{\mathbf{W}}_{Z(n)} \mathbf{B}_{a,Z(n)} - \mathbf{K}_{\mathbf{r}}^{-1}$.

4) For $k \geq 3$, $h^{(k)}(\tilde{\boldsymbol{\theta}}_a) = \tilde{l}^{(k)} = -\sum_{i=1}^{m} \tilde{w}_{Z(n)_i}^{(k)} \overset{k}{\otimes} \mathbf{B}_{a,Z(n)_i}'$, where $\tilde{w}_{Z(n)_i}^{(k)}$ is the $(k-1)^{th}$ derivative of $\tilde{\pi}_{Z(n)i}$ with respect to $\tilde{\eta}_{Z(n)_i}$, which is $\frac{d^{k-1}\tilde{\pi}_{Z(n)_i}}{d\tilde{\eta}_{Z(n)_i}^{k-1}}$.

According to Taylor series, $h(\boldsymbol{\theta}_a)$ can be expanded as

$$h(\boldsymbol{\theta}_a) = h(\tilde{\boldsymbol{\theta}}_a) + h^{(1)}(\tilde{\boldsymbol{\theta}}_a)(\boldsymbol{\theta}_a - \tilde{\boldsymbol{\theta}}_a) + \sum_{k=2}^{\infty} \frac{1}{k!} \overset{k-1}{\otimes} (\boldsymbol{\theta}_a - \tilde{\boldsymbol{\theta}}_a)' h^{(k)}(\tilde{\boldsymbol{\theta}}_a)(\boldsymbol{\theta}_a - \tilde{\boldsymbol{\theta}}_a).$$

Then we can express the following integral as

$$\begin{aligned}
&\int \tilde{p}(\boldsymbol{\theta}_a|\mathbf{r}) p(\mathbf{X}_{Z(n)}|\tilde{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_a) d\boldsymbol{\theta}_a \\
=&(2\pi)^{-\frac{k_a}{2}} |\mathbf{K}_\mathbf{r}|^{-\frac{1}{2}} \int \exp(-\frac{1}{2}(\boldsymbol{\theta}_a - \boldsymbol{\mu}_\mathbf{r})' \mathbf{K}'_\mathbf{r} (\boldsymbol{\theta}_a - \boldsymbol{\mu}_\mathbf{r}) + l) d\boldsymbol{\theta}_a \\
=&(2\pi)^{-\frac{k_a}{2}} |\mathbf{K}_\mathbf{r}|^{-\frac{1}{2}} \exp(\tilde{l} - \frac{1}{2}(\tilde{\boldsymbol{\theta}}_a - \boldsymbol{\mu}_\mathbf{r})' \mathbf{K}_\mathbf{r}^{-1} (\tilde{\boldsymbol{\theta}}_a - \boldsymbol{\mu}_\mathbf{r})) \int \exp(-\frac{1}{2}(\boldsymbol{\theta}_a - \tilde{\boldsymbol{\theta}}_a)' h^{(2)}(\tilde{\boldsymbol{\theta}}_a)(\boldsymbol{\theta}_a - \tilde{\boldsymbol{\theta}}_a)) \exp(R) d\boldsymbol{\theta}_a \\
=&|\mathbf{K}_\mathbf{r}|^{-\frac{1}{2}} |h^{(2)}(\tilde{\boldsymbol{\theta}}_a)|^{-\frac{1}{2}} \exp(\tilde{l} - \frac{1}{2}(\tilde{\boldsymbol{\theta}}_a - \boldsymbol{\mu}_\mathbf{r})' \mathbf{K}_\mathbf{r}^{-1} (\tilde{\boldsymbol{\theta}}_a - \boldsymbol{\mu}_\mathbf{r})) \mathbb{E}[\exp(R)],
\end{aligned}$$

where $R = \sum_{k=3}^{\infty} T_k$ and $T_k = \frac{1}{k!}(\overset{k-1}{\otimes}(\boldsymbol{\theta}_a - \tilde{\boldsymbol{\theta}}_a)' h^{(k)}(\tilde{\boldsymbol{\theta}}_a)(\boldsymbol{\theta}_a - \tilde{\boldsymbol{\theta}}_a))$. This follows the Laplace approximation in Raudenbush et al. (2000). Since $\exp(R) = 1 + R + \frac{1}{2}R^2 + \cdots + \frac{1}{k!}R^k + \cdots$, using the same approximation as Raudenbush et al. (2000), $\mathbb{E}[\exp(R)] = 1 + \mathbb{E}[T_4] + \mathbb{E}[T_6] + \frac{1}{2}\mathbb{E}[T_3^2]$. $\mathbb{E}[T_k] = 0$, for odd $k, k \geq 3$ and $\mathbb{E}[T_k] = \frac{(k-1)(k-3)\cdots 3}{k!} \text{vec}'(\overset{\frac{k}{2}}{\otimes}(-h^{(2)}(\tilde{\boldsymbol{\theta}}_a))^{-1}) \text{vec}(h^{(k)}(\tilde{\boldsymbol{\theta}}_a))$ for even $k, k \geq 3$. Also $\mathbb{E}[T_k T_l] = 0$, for odd $(k+l), k \geq 3, l \geq 3$ and $\mathbb{E}[T_k T_l] = \frac{(k+l-1)(k+l-3)\cdots 3}{k!l!} \text{vec}'(\overset{\frac{k+l}{2}}{\otimes}(-h^{(2)}(\tilde{\boldsymbol{\theta}}_a))^{-1}) \text{vec}(h^{(k)}(\tilde{b}) \otimes h^{(l)}(\tilde{b}))$ for even $(k+l), k \geq 3, l \geq 3$. Such that we can derive the approximation form of L_1,

$$\begin{aligned}
L_1 =& \sum_{\mathbf{r} \in \mathbb{R}} \tilde{p}(\mathbf{r}) \int \tilde{p}(\boldsymbol{\theta}_a|\mathbf{r}) p(\mathbf{X}_{Z(n)}|\tilde{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_a) d\boldsymbol{\theta}_a \\
\approx& \sum_{\mathbf{r} \in \mathbb{R}} \tilde{p}(\mathbf{r}) \big(|\mathbf{K}_\mathbf{r}|^{-\frac{1}{2}} |\mathbf{B}'_{a,Z(n)} \tilde{\mathbf{W}}_{Z(n)} \mathbf{B}_{a,Z(n)} + \mathbf{K}_\mathbf{r}^{-1}|^{-\frac{1}{2}} \exp\{\tilde{l} - \frac{1}{2}(\tilde{\boldsymbol{\theta}}_a - \boldsymbol{\mu}_\mathbf{r})' \mathbf{K}_\mathbf{r}^{-1} (\tilde{\boldsymbol{\theta}}_a - \boldsymbol{\mu}_\mathbf{r})\} \\
&\times \big(1 + \frac{1}{8} \text{vec}'(\overset{2}{\otimes}(\mathbf{B}'_{a,Z(n)} \tilde{\mathbf{W}}_{Z(n)} \mathbf{B}_{a,Z(n)} + \mathbf{K}_\mathbf{r}^{-1})^{-1}) \text{vec}(-\sum_{i=1}^{m} \tilde{w}^{(4)}_{Z(n)_i} (\overset{4}{\otimes} \mathbf{B}'_{a,Z(n)_i})) \\
&+ \frac{1}{48} \text{vec}'(\overset{3}{\otimes}(\mathbf{B}'_{a,Z(n)} \tilde{\mathbf{W}}_{Z(n)} \mathbf{B}_{a,Z(n)} + \mathbf{K}_\mathbf{r}^{-1})^{-1}) \text{vec}(-\sum_{i=1}^{m} \tilde{w}^{(6)}_{Z(n)_i} (\overset{6}{\otimes} \mathbf{B}'_{a,Z(n)_i})) \big) \\
&+ \frac{15}{72} \text{vec}'(\overset{3}{\otimes}(\mathbf{B}'_{a,Z(n)} \tilde{\mathbf{W}}_{Z(n)} \mathbf{B}_{a,Z(n)} + \mathbf{K}_\mathbf{r}^{-1})^{-1}) \text{vec}\big((\sum_{i=1}^{m} \tilde{w}^{(3)}_{Z(n)_i} (\overset{3}{\otimes} \mathbf{B}'_{a,Z(n)_i})) \otimes (\sum_{i=1}^{m} \tilde{w}^{(3)}_{Z(n)_i} (\overset{3}{\otimes} \mathbf{B}'_{a,Z(n)_i})) \big) \big).
\end{aligned}$$

Here $\tilde{\boldsymbol{\theta}}_a$ can be obtained using gradient ascent method. Having got L_1, we can compute PBF. That is the whole procedure suitable for general distributions in exponential family.

However, for Gaussian distribution, we can derive a close-form $\tilde{\boldsymbol{\theta}}_a$, so there exists a simpler form for PBF. $\alpha(\phi_i) = \text{Var}(\mathbf{X}_i|\boldsymbol{\theta}_a, \boldsymbol{\theta}_n) = \sigma_e^2, i = 1, \ldots, p$, $\tilde{w}_{Z(n)_i} = 1$ and $\tilde{w}^{(k)}_{Z(n)_i} = 0$, for $k \geq 3$. Let

$h^{(1)}(\tilde{\boldsymbol{\theta}}_a)=\mathbf{0}$, we can obtain that

$$\frac{1}{\sigma_e^2}\mathbf{B}'_{a,Z(n)}(\mathbf{X}_{Z(n)}-\mathbf{B}_{a,Z(n)}\tilde{\boldsymbol{\theta}}_a-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)-\mathbf{K}_{\mathbf{r}}^{-1}(\tilde{\boldsymbol{\theta}}_a-\boldsymbol{\mu}_{\mathbf{r}})=\mathbf{0}.$$

Undo $\tilde{\boldsymbol{\theta}}_a$ as

$$\begin{aligned}\tilde{\boldsymbol{\theta}}_a&=(\frac{\mathbf{B}'_{a,Z(n)}\mathbf{B}_{a,Z(t)}}{\sigma_e^2}+\mathbf{K}_{\mathbf{r}}^{-1})^{-1}(\frac{1}{\sigma_e^2}\mathbf{B}'_{a,Z(n)}(\mathbf{X}_{Z(n)}-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)+\mathbf{K}_{\mathbf{r}}^{-1}\boldsymbol{\mu}_{\mathbf{r}})\\&\approx\frac{1}{\sigma_e^2}\mathbf{K}_{\mathbf{r}}\mathbf{B}'_{a,Z(n)}(\mathbf{X}_{Z(n)}-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)+\boldsymbol{\mu}_{\mathbf{r}}(\textit{The terms in }\mathbf{K}_{\mathbf{r}}\textit{ are always quite small}).\end{aligned}$$

So the item in integral L_1 can be derived as

$$\begin{aligned}&h(\tilde{\boldsymbol{\theta}}_a)=\tilde{l}-\frac{1}{2}(\tilde{\boldsymbol{\theta}}_a-\boldsymbol{\mu}_{\mathbf{r}})'\mathbf{K}_{\mathbf{r}}^{-1}(\tilde{\boldsymbol{\theta}}_a-\boldsymbol{\mu}_{\mathbf{r}})\\
=&-\frac{1}{\sigma_e^2}(\mathbf{X}_{Z(n)}-\mathbf{B}_{a,Z(n)}\tilde{\boldsymbol{\theta}}_a-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)'(\mathbf{X}_{Z(n)}-\mathbf{B}_{a,Z(n)}\tilde{\boldsymbol{\theta}}_a-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)-\frac{1}{2}(\tilde{\boldsymbol{\theta}}_a-\boldsymbol{\mu}_{\mathbf{r}})'\mathbf{K}_{\mathbf{r}}^{-1}(\tilde{\boldsymbol{\theta}}_a-\boldsymbol{\mu}_{\mathbf{r}})\\
=&-\frac{1}{\sigma_e^2}(\mathbf{X}_{Z(n)}-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)'(\mathbf{X}_{Z(n)}-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)+\frac{2}{\sigma_e^2}(\mathbf{X}_{Z(n)}-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)'\mathbf{B}_{a,Z(n)}\tilde{\boldsymbol{\theta}}_a-\frac{1}{\sigma_e^2}\tilde{\boldsymbol{\theta}}'_a\mathbf{B}'_{a,Z(n)}\mathbf{B}_{a,Z(n)}\tilde{\boldsymbol{\theta}}_a\\
&-\frac{1}{2}\tilde{\boldsymbol{\theta}}'_a\mathbf{K}_{\mathbf{r}}^{-1}\tilde{\boldsymbol{\theta}}_a+\boldsymbol{\mu}'_{\mathbf{r}}\mathbf{K}_{\mathbf{r}}^{-1}\tilde{\boldsymbol{\theta}}_a-\frac{1}{2}\boldsymbol{\mu}'_{\mathbf{r}}\mathbf{K}_{\mathbf{r}}^{-1}\boldsymbol{\mu}_{\mathbf{r}}\\
\approx&-\frac{1}{\sigma_e^2}(\mathbf{X}_{Z(n)}-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)'(\mathbf{X}_{Z(n)}-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)+\frac{2}{\sigma_e^2}(\mathbf{X}_{Z(n)}-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)'\mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_{\mathbf{r}}-\frac{1}{\sigma_e^2}\boldsymbol{\mu}'_{\mathbf{r}}\mathbf{B}'_{a,Z(n)}\mathbf{B}_{a,Z(t)}\boldsymbol{\mu}_{\mathbf{r}}\\
&-\frac{1}{2}\boldsymbol{\mu}'_{\mathbf{r}}\mathbf{K}_{\mathbf{r}}^{-1}\boldsymbol{\mu}_{\mathbf{r}}-\frac{1}{\sigma_e^2}(\mathbf{X}_{Z(n)}-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)'\mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_{\mathbf{r}}+\frac{1}{\sigma_e^2}\boldsymbol{\mu}'_{\mathbf{r}}\mathbf{B}'_{a,Z(n)}(\mathbf{X}_{Z(n)}-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)+\boldsymbol{\mu}'_{\mathbf{r}}\mathbf{K}_{\mathbf{r}}^{-1}\boldsymbol{\mu}_{\mathbf{r}}\\
&-\frac{1}{2}\boldsymbol{\mu}'_{\mathbf{r}}\mathbf{K}_{\mathbf{r}}^{-1}\boldsymbol{\mu}_{\mathbf{r}}(\textit{Delete the terms containing two or higher orders of }\mathbf{K}_{\mathbf{r}})\\
=&-\frac{1}{\sigma_e^2}(\mathbf{X}_{Z(n)}-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)'(\mathbf{X}_{Z(n)}-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)+\frac{2}{\sigma_e^2}(\mathbf{X}_{Z(n)}-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)'\mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_{\mathbf{r}}-\frac{1}{\sigma_e^2}\boldsymbol{\mu}'_{\mathbf{r}}\mathbf{B}'_{a,Z(n)}\mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_{\mathbf{r}}.\end{aligned}$$

Further simplify Λ_n^{PBF} as

$$\begin{aligned}\Lambda_n^{PBF}&\propto\frac{\mathbb{E}_{\mathbf{r}}\left[|\mathbf{K}_{\mathbf{r}}|^{-\frac{1}{2}}|\mathbf{B}'_{a,Z(n)}\tilde{\mathbf{W}}_{Z(n)}\mathbf{B}_{a,Z(n)}+\mathbf{K}_{\mathbf{r}}^{-1}|^{-\frac{1}{2}}\exp\{\tilde{l}-\frac{1}{2}(\tilde{\boldsymbol{\theta}}_a-\boldsymbol{\mu}_{\mathbf{r}})'\mathbf{K}_{\mathbf{r}}^{-1}(\tilde{\boldsymbol{\theta}}_a-\boldsymbol{\mu}_{\mathbf{r}})\}\right]}{\exp(-\frac{1}{\sigma_e^2}(\mathbf{X}_{Z(n)}-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)'(\mathbf{X}_{Z(n)}-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n))}\\
&=\mathbb{E}_{\mathbf{r}}[|\mathbf{K}_{\mathbf{r}}|^{-\frac{1}{2}}|\mathbf{B}'_{a,Z(n)}\tilde{\mathbf{W}}_{Z(n)}\mathbf{B}_{a,Z(n)}+\mathbf{K}_{\mathbf{r}}^{-1}|^{-\frac{1}{2}}\exp(\frac{1}{\sigma_e^2}(2(\mathbf{X}_{Z(n)}-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)'\mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_{\mathbf{r}}-\boldsymbol{\mu}'_{\mathbf{r}}\mathbf{B}'_{a,Z(n)}\mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_{\mathbf{r}}))]\\
&\approx\mathbb{E}_{\mathbf{r}}\left[2(\mathbf{X}_{Z(n)}-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)'\mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_{\mathbf{r}}-\boldsymbol{\mu}'_{\mathbf{r}}\mathbf{B}'_{a,Z(n)}\mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_{\mathbf{r}}\right]\ (\textit{Since the terms in }\exp(\cdot)\textit{ close to }0,\\
&\textit{we use the first order expansion of Taylor series as an approximation})\\
&=2(\mathbf{X}_{Z(n)}-\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n)'\mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_a-\boldsymbol{\mu}'(\mathbf{B}'_{a,Z(n)}\mathbf{B}_{a,Z(n)}\circ\bar{\mathbf{A}})\boldsymbol{\mu}.\end{aligned}$$

Here $\bar{\mathbf{A}}$ has diagonal items $\bar{A}_{ii}=\alpha_i$, $i=1,\ldots,k_a$, and other items $\bar{A}_{ij}=\alpha_i\alpha_j$, $\forall i,j=1,\ldots,k_a$, $i\neq j$.

Appendix C: Proof of Theorem 1 and Theorem 2

As suggested by Algorithm 2, Thompson sampling procedure based on SLRT is to sample Z by ranking $\Lambda_i^{SLRT} = \frac{1}{a(\phi_i)}(X_i^\star \mathbf{B}_{a,i:}\boldsymbol{\mu}_a - b(\mathbf{B}_{b,i:}\boldsymbol{\theta}_0 + \mathbf{B}_{a,i:}\boldsymbol{\mu}_a) + b(\mathbf{B}_{b,i:}\boldsymbol{\theta}_0))$, for $i = 1, \ldots, p$, from the largest to the smallest and select the top m ones.

To guarantee the algorithm can detect the anomaly efficiently and select the most anomalous variables, we need $\boldsymbol{\theta}_a$ is accurately estimated. However, the estimation of $\boldsymbol{\theta}_n$ and $\boldsymbol{\theta}_a$ are coupled, such that $\boldsymbol{\theta}_n$ should also be estimated accurately. This is guaranteed by Bayesian theory(Gelman et al. 2013). That is, when $p \to \infty$, $m \to \infty$, and $0 < m/p \leq 1$, according to Bayesian theory, the maximum a posteriori estimation $\tilde{\boldsymbol{\theta}}_n$, will converge to its true value, i.e., $\boldsymbol{\theta}_n$. To guarantee $\boldsymbol{\theta}_a$ is estimated accurately, according to Theorem 5 of Wang and Blei (2019), the variational Bayesian posterior converges to the point mass of the true parameter value in distribution. Note that the posterior in Equation (7) is not exactly the true posterior, but the exponentially weighted posterior, which may render Theorem 5 of Wang and Blei (2019) not valid. Thus we need $\lambda \to 0$, i.e., representing that all samples receive almost equivalent weights, to guarantee as $n \to \infty$, all the samples could be used equivalently for estimating $\boldsymbol{\theta}_a$. This assumption of $\lambda \to 0$ is common in theoretical analysis of other online process monitoring schemes with exponential weights (Zhou et al. 2012; Zou et al. 2012).

In our case, under normal condition, the true value of $\theta_{a,j}$ equals 0, $\forall j = 1, \ldots, k_a$. The posterior distribution that we obtain through variational Bayesian method is of spike-slab form. For example, $q(\theta_{a,j}) \sim N(\mu_j, s_j^2)$ with probability α_j and $q(\theta_{a,j}) \sim N(0, vs_j^2)$ with probability $1 - \alpha_j$. Then suggested by Theorem 5 of Wang and Blei (2019), as $n \to \infty$,

$$q(\theta_{a,j}) \xrightarrow{d} \delta_0, \forall j, \tag{1}$$

where δ_0 is a point mass at 0. That suggests $\mu_j \to 0$ and $s_j^2 \to 0, \forall j = 1, \ldots, k_a$. So in normal condition, $\Lambda_i^{SLRT} = \frac{1}{a(\phi_i)}(X_i^\star \mathbf{B}_{a,i:}\boldsymbol{\mu}_a - b(\mathbf{B}_{b,i:}\boldsymbol{\theta}_0 + \mathbf{B}_{a,i:}\boldsymbol{\mu}_a) + b(\mathbf{B}_{b,i:}\boldsymbol{\theta}_0)) \to 0$ in the same rate of $\sqrt{n}$-convergence (Wang and Blei 2019), i.e., $|\Lambda_i^{SLRT} - 0| = O(1/\sqrt{n})$, for $i = 1, \ldots, p$, so we can not say any Λ_i^{SLRT} is larger than others, which means under the limit conditions, we sample the variables $Z(n+1)$ randomly.

Following a similar way, under abnormal condition, assume the anomaly relates to a certain basis set $\mathcal{A} \subset \{1, \ldots, p\}$. For $l \in \mathcal{A}$, assume the anomaly relates to the l^{th} base has change magnitude ξ_l. Then suggested by Theorem 5 of Wang and Blei (2019), as $n \to \infty$,

$$q_l(\theta_{a,l}) \xrightarrow{d} \delta_{\xi_l}, \forall l \in \mathcal{A}, \tag{2}$$

$$q_j(\theta_{a,j}) \xrightarrow{d} \delta_0, \forall j \in \{1, \ldots, p\} - \mathcal{A}, \tag{3}$$

where δ_{ξ_l} is a point mass at ξ_l. That suggests $\mu_l \to \xi_l$, $\alpha_l \to 1$ and $s_l^2 \to 0, \forall l \in \mathcal{A}$. The same as normal condition, $\mu_j \to 0$ and $s_j^2 \to 0$, $\forall j \notin \mathcal{A}$.

So in abnormal condition, $\mathbb{E}[\Lambda_i^{SLRT}] \to \frac{1}{a(\phi_i)}(b'(\mathbf{B}_{a,i:}\boldsymbol{\xi} + \mathbf{B}_{b,i:}\boldsymbol{\theta}_0)\mathbf{B}_{a,i:}\boldsymbol{\xi} - b(\mathbf{B}_{a,i:}\boldsymbol{\xi} + \mathbf{B}_{b,i:}\boldsymbol{\theta}_0) + b(\mathbf{B}_{b,i:}\boldsymbol{\theta}_0))$, where $\boldsymbol{\xi} = (0, \ldots, \xi_l, \ldots, 0)'$, $\forall i = 1, \ldots, p$. Compute the derivative with respect to $\mathbf{B}_{a,i:}\boldsymbol{\xi}$ as $\lim_{n\to\infty} \frac{\partial \mathbb{E}[\Lambda_i^{SLRT}]}{\partial(\mathbf{B}_{a,i:}\boldsymbol{\xi})} = \frac{1}{a(\phi_i)} b''(\mathbf{B}_{a,i:}\boldsymbol{\xi} + \mathbf{B}_{b,i:}\boldsymbol{\theta}_0)\mathbf{B}_{a,i:}\boldsymbol{\xi}$. At the same time, $\frac{1}{a(\phi_i)} b''(\mathbf{B}_{a,i:}\boldsymbol{\xi} + \mathbf{B}_{b,i:}\boldsymbol{\theta}_0) = \frac{1}{a(\phi_i)^2}\mathrm{Var}(X_i) \geq 0$. That means that if $\mathbf{B}_{a,i:}\boldsymbol{\xi} \geq 0$, $\lim_{n\to\infty} \frac{\partial \mathbb{E}[\Lambda_i^{SLRT}]}{\partial(\mathbf{B}_{a,i:}\boldsymbol{\xi})} \geq 0$. On the other hand, if $\mathbf{B}_{a,i:}\boldsymbol{\xi} \leq 0$, $\lim_{n\to\infty} \frac{\partial \mathbb{E}[\Lambda_i^{SLRT}]}{\partial(\mathbf{B}_{a,i:}\boldsymbol{\xi})} \leq 0$. In other words, in limiting condition, $\mathbb{E}[\Lambda_i^{SLRT}]$ monotonically increases with $|\mathbf{B}_{a,i:}\boldsymbol{\xi}|$. In a word, under abnormal condition, whichever the change direction, we prefer to choose those variables with large absolute true changed value $|\xi_l|$ and large relation to the l^{th} base(i.e., $\mathbf{B}_{a,il}$), for $\forall l \in \mathcal{A}$.

Appendix D: Results for Numerical Studies

Choice of Parameters of EF-BSSCD

Choice of bases: For $\mathbf{B}_b$, there are two methods. First, set it from notable complete basis space, such as B-spline basis, Fourier basis, kernel functions, wavelet basis, etc., which are often used in high-dimensional profile data modeling(Chang and Yadama 2010; Paynabar and Jin 2011). Second, learn it from historical reference data by feature extraction methods, such as PCA for exponential family (Collins et al. 2001). As for the number of bases, k_b, it should be set large enough to capture various features of the background. For example, using PCA, the number of bases would better be large enough to guarantee the explainable variance of data exceeds a threshold, say 95%. For $\mathbf{B}_a$, it can be designed according to our prior interested abnormal patterns (Yan et al. 2017). Or it can also be estimated from historical abnormal data through PCA methods. $\mathbf{B}_a = \mathbf{I}$ can also be a choice when there is no correlation between the changed variables. As for k_a, it can be set similarly as k_b.

Choice of w_j, σ_j and v: For w_j, a prior such as a degenerate point value set according to the chance of change on every basis or a $Beta(a, b)$ distribution (Ročková and George 2014) are both compatible (Ishwaran et al. 2005). For sparse change, we recommend small a and large b. For σ_j, a prior such as Cauchy distribution, or double exponential distribution (Ročková and George 2014) or a fixed value are both suitable (George and McCulloch 1997). In the numerical studies, assume we know the maximum change magnitude, i.e., ξ, in advance, we can set $\sigma_j \geq |\xi|$ (i.e., ξ is in the 1-level sigma of $N(0, \sigma_j)$) to guarantee that there is a high probability to accommodate all plausible $\theta_{a,j} \neq 0$ in the slab distribution. However, too large σ_j is not recommended, since it may induce unreasonably large $\theta_{a,j}$. Thus a suitably large σ_j is enough. Generally, v is recommended to be a suitably small value (George and McCulloch 1993). In particular, it depends on how small

the exclusion threshold of unimportant nonzero values is. For example, in our case, we assume the noise is in the scale of 10^{-3}, and $\theta_{a,j}$ smaller than this scale can be regarded as zero. So we set $v = (10^{-3})^2 = 10^{-6}$.

Choice of λ: According to Montgomery (2007), λ is set as a small value in $[0.05, 0.25]$. The smaller λ is, the better detection for small changes is.

Choice of m**:** If there is a change in the system, the larger m is, the smaller the ADD will be. Consequently, in our case, we can set m as the maximum number of sensors that can be allocated at each time point.

Choice of h**:** The detection threshold h is determined by controlling a pre-specific false alarm rate, or equivalently setting the in-control ARL to a specific large number, e.g., $ARL = 200$. An empirical method can be used to find the threshold based on the empirical distribution of $\Lambda_{\boldsymbol{\theta}_a}(\mathbf{X}_{Z(n)})$. The most commonly used method is based on Monte Carlo simulation, and the detailed algorithm can be referred to Zhang et al. (2020).

Parameter Settings

In simulations, for Gaussian, for TRAS, we set its parameters $r = 3$, $\mu_{min} = 0.5$ and $\Delta = 10^{-3}$. For SASAM, we set $\boldsymbol{\theta}_1 = 0.1$, $\boldsymbol{\theta}_2 = 0.7$, $\mu_{min} = 0.5$, $h = 1$. As to CMAB(s), we set $\lambda = 0.1$ for exponentially weighted posterior. For NAS, we set $g_{p+1} = 0.1$, $\Delta = 0.1$, $k = 0.5$. For R-SADA, we set $\mu_{min} = 0.5$, $k = 0.5$. For TSSRP, we set $r = 2$, $\mu_{1,true} = 0.5$.

For Poisson, for TRAS, we set its parameters $r = 3$, $\mu_{min} = 0.5$ and $\Delta = 10^{-3}$. For SASAM, we set $\boldsymbol{\theta}_1 = 0.1$, $\boldsymbol{\theta}_2 = 0.7$, $\mu_{min} = 0.5$, $h = 1$. As to CMAB(s), we set $\lambda = 0.1$ for exponentially weighted posterior. For NAS, we set $g_{p+1} = 0.1$, $\Delta = 0.1$, $k = 0.5$. For R-SADA, we set $\mu_{min} = 0.5$, $k = 0.5$. For TSSRP, we set $r = 2$, $\mu_{1,true} = 0.5$.

For binomial, for TRAS, we set its parameters $r = 3$, $\mu_{min} = 0.5$ and $\Delta = 10^{-3}$. For SASAM, we set $\boldsymbol{\theta}_1 = 0.1$, $\boldsymbol{\theta}_2 = 0.7$, $\mu_{min} = 0.5$, $h = 1$. As to CMAB(s), we set $\lambda = 0.1$ for exponentially weighted posterior. For NAS, we set $g_{p+1} = 0.1$, $\Delta = 0.1$, $k = 0.5$. For R-SADA, we set $\mu_{min} = 0.5$, $k = 0.5$. For TSSRP, we set $r = 2$, $\mu_{1,true} = 10$.

For solar flare detection in the three case studies, the parameters of TRAS are selected to be $r = 40$, $\mu_{min} = 2.1$ and $\Delta = 5 * 10^{-6}$. The parameters of SASAM are set to be $\theta_1 = 0.1$, $\theta_2 = 0.4$, $h = 5$ and $\mu_{min} = 2$. The parameters of NAS are selected as $g_{p+1} = 10^{-3}$, $\Delta = 1.47 * 10^{-5}$, $k = 0.1$, $\lambda_1 = 1.6 * 10^{-3}$ and $\lambda_0 = 0.0103$. The parameters of R-SADA are set as $k = 2.5 * 10^4$ and $\mu_{min} = 1$. The parameters of TSSRP are set $r = 40$, and $\mu_{1,true} = 0.1$.

For COVID-19 infectious rate change detection, for TRAS, we set its parameters $r = 1$, $\mu_{min} = 2$ and $\Delta = 10^{-3}$ For SASAM, we set $\theta_1 = 0.3$, $\theta_2 = 0.4$, $\mu_{min} = 1$, $h = 1$. For CMAB(s), we set $\lambda = 0.1$

for exponentially weighted posterior. For NAS, we set $g_{p+1} = 0.01$, $\Delta = 10^{-1}$, $k = 0.2$. For R-SADA, we set $\mu_{min} = 1$, $k = 1.5$. For TSSRP, we set $r = 1$, $\mu_{1,true} = 0.5$.

For manufacturing defect change detection, for TRAS, we set its parameters $r = 1$, $\mu_{min} = 1$ and $\Delta = 10^{-3}$. For SASAM, we set $\boldsymbol{\theta}_1 = 0.1$, $\boldsymbol{\theta}_2 = 0.7$, $\mu_{min} = 1$, $h = 1$. As to CMAB(s), we set $a = 0.1$ for exponentially weighted posterior. For NAS, we set $g_{p+1} = 0.1$, $\Delta = 5 * 10^{-2}$, $k = 7$. For R-SADA, we set $\mu_{min} = 2$, $k = 1$. For TSSRP, we set $r = 1$, $\mu_{1,true} = 0.3$.

Simulation Results

First, we show the specific Average Detection Delays(ADDs) and the corresponding Standard Deviation of Detection Delays(STDs) in Table 1, Table 2 and Table 3 for the experiments in Section 5.1, 5.2 and 5.3 respectively.

Table 1 Average Detection Delays/ADDs(Standard Deviation of Detection Delays/STDs) for Gaussian data with $p = 10$, $m = 3$

ξ	NAS	R-SADA	TRAS	CMAB(s)	SASAM	TSSRP	EF-BSSCD(S)	EF-BSSCD(G)	EF-BSSCD(P)	ORACLE
0.0	200(201)	200(197)	200(180)	200(170)	200(141)	200(127)	200(244)	200(243)	200(235)	200(235)
0.2	127(131)	95.5(100)	89.0(71.3)	19.1(11.6)	111(78.7)	44.2(11.5)	**7.30**(8.29)	8.97(8.80)	8.40(9.78)	1.01(0.10)
0.4	55.0(66.0)	34.9(36.0)	53.8(50.2)	7.88(4.97)	55.8(43.2)	29.5(8.86)	**3.16**(2.85)	4.91(5.57)	3.41(2.98)	1.00(0.00)
0.6	29.5(30.4)	16.2(15.7)	41.0(42.9)	4.60(3.20)	35.3(26.8)	20.7(5.93)	**2.49**(2.02)	3.97(4.41)	2.67(2.26)	1.00(0.00)
0.8	18.6(16.3)	10.4(10.7)	35.9(41.8)	3.21(2.53)	24.0(20.1)	15.2(3.83)	**2.23**(1.75)	3.52(3.48)	2.31(1.71)	1.00(0.00)
1.0	14.7(12.2)	7.87(8.43)	30.1(37.8)	2.98(2.69)	18.8(16.9)	12.5(3.03)	**2.03**(1.61)	3.41(3.14)	2.11(1.71)	1.00(0.00)
1.2	11.7(7.71)	7.39(8.59)	28.2(35.7)	2.67(2.45)	14.9(7.31)	10.7(2.40)	**1.85**(1.40)	3.17(4.29)	1.94(1.53)	1.00(0.00)
1.4	10.7(7.23)	6.84(7.57)	24.9(33.0)	2.35(2.05)	12.8(8.99)	9.60(2.21)	**1.74**(1.19)	2.93(2.86)	1.89(1.40)	1.00(0.00)
1.6	9.62(5.57)	6.40(8.00)	22.7(33.2)	2.35(2.24)	10.7(4.90)	8.71(2.11)	**1.76**(1.31)	2.89(2.58)	1.84(1.39)	1.00(0.00)
1.8	9.05(4.81)	5.93(7.78)	21.9(32.5)	2.10(1.86)	9.42(4.51)	8.06(1.98)	**1.70**(1.23)	2.80(3.57)	1.83(1.38)	1.00(0.00)
2.0	8.78(4.64)	6.24(7.44)	21.9(34.3)	2.10(1.94)	8.45(4.11)	7.37(1.97)	**1.66**(1.17)	2.78(2.80)	**1.66**(1.13)	1.00(0.00)

Table 2 Average Detection Delays/ADDs(Standard Deviation of Detection Delays/STDs) for Poisson data with $p = 10$, $m = 3$

ξ	NAS	R-SADA	TRAS	CMAB(s)	SASAM	TSSRP	EF-BSSCD(S)	EF-BSSCD(G)	EF-BSSCD(P)	ORACLE
0	200(239)	200(205)	200(202)	200(189)	200(161)	200(169)	200(247)	200(261)	200(199)	200(162)
0.2	188(204)	168(168)	155(135)	162(145)	144(126)	206(169)	151(180)	**125**(160)	143(127)	2.23(1.56)
0.4	150(170)	121(129)	103(92.8)	109(96.5)	100(78.1)	204(166)	**80.9**(104)	122(164)	136(133)	2.49(1.98)
0.6	120(140)	75.9(86.0)	81.9(77.7)	64.8(61.1)	66.5(47.3)	199(170)	**61.9**(88.7)	66.3(95.7)	106(112)	2.07(1.28)
0.8	81.8(94.7)	46.5(48.4)	61.9(61.2)	39.4(38.7)	46.1(32.6)	203(163)	**33.5**(45.6)	50.0(71.7)	87.7(99.5)	1.86(1.12)
1.0	65.3(87.1)	32.2(33.6)	55.5(58.8)	24.5(24.8)	35.5(24.3)	217(176)	**22.0**(34.8)	25.3(31.8)	58.6(54.6)	1.63(0.81)
1.2	42.2(50.2)	23.2(22.7)	42.8(48.8)	13.4(13.2)	26.0(17.6)	209(174)	**12.3**(16.1)	20.2(26.7)	48.7(66.3)	1.81(0.92)
1.4	32.6(35.8)	18.5(17.6)	38.7(45.3)	11.5(12.7)	21.0(13.5)	206(179)	**11.6**(14.3)	12.2(14.5)	34.0(42.5)	1.75(0.86)
1.6	22.8(24.2)	15.2(14.5)	36.4(50.4)	**7.03**(7.20)	17.1(11.3)	212(169)	7.09(8.74)	11.2(13.3)	24.0(35.6)	1.69(0.81)
1.8	19.1(18.5)	12.8(12.2)	30.7(43.3)	**5.26**(4.41)	13.8(8.79)	203(158)	6.67(9.29)	8.99(10.4)	17.0(20.8)	1.50(0.61)
2.0	14.8(12.9)	10.4(8.70)	28.7(42.4)	4.21(4.07)	11.4(7.69)	213(169)	**3.87**(4.22)	7.71(8.44)	12.4(16.4)	1.49(0.59)

Second, we also run experiments of $p = 100$ with $m = 30$. We set $\mathbf{B}_b \in \mathcal{R}^{100\times 3}$ are three lowest frequency Fourier bases and $\mathbf{B}_a \in \mathcal{R}^{100\times 26}$ are 26 three-order B-spline bases with 29 equally spaced knots.

For Gaussian data, for EF-BSSCD methods, we set $\lambda = 0.1$, $w_j = 0.1$, $\sigma_j = 2$, $v = 10^{-6}$. As to other baseline methods, we set all the parameters according to the recommendations in their

Table 3 Average Detection Delays/ADDs(Standard Deviation of Detection Delays/STDs) for binomial data with $p=10$, $m=3$, $N=10$

ϕ	NAS	R-SADA	TRAS	CMAB(s)	SASAM	TSSRP	EF-BSSCD(S)	EF-BSSCD(G)	EF-BSSCD(P)	ORACLE
0.0	200(200)	200(201)	200(180)	200(180)	200(157)	200(207)	200(223)	200(193)	200(200)	200(206)
0.2	191(186)	173(171)	148(128)	200(203)	170(132)	212(209)	**111**(160)	134(149)	148(161)	2.02(1.17)
0.4	160(163)	129(128)	91.4(75.5)	117(117)	122(97.9)	214(209)	**65.7**(97.6)	91.9(91.6)	140(165)	1.85(1.10)
0.6	125(138)	84.7(92.0)	71.0(60.8)	67.0(66.5)	89.5(80.1)	207(208)	**47.0**(69.8)	61.2(71.3)	91.5(114)	1.74(1.04)
0.8	94.5(108)	53.8(58.1)	57.5(54.8)	39.8(37.9)	65.7(57.9)	223(221)	**30.6**(50.0)	32.2(34.2)	69.7(84.3)	1.65(0.84)
1.0	69.2(75.2)	34.3(34.5)	53.3(55.3)	26.1(22.7)	49.5(41.7)	211(215)	**19.7**(31.6)	26.0(27.6)	58.4(70.2)	1.65(0.75)
1.2	53.9(60.0)	25.8(25.0)	45.0(50.6)	17.9(14.7)	38.5(30.9)	221(215)	**11.2**(17.4)	19.0(20.6)	47.5(73.4)	1.70(0.84)
1.4	42.4(50.7)	19.1(18.8)	41.5(46.3)	12.4(8.50)	34.2(30.6)	204(194)	**6.86**(9.63)	16.1(15.9)	23.9(29.7)	1.55(0.72)
1.6	36.6(40.0)	15.0(13.2)	36.6(43.8)	10.5(6.52)	28.8(19.4)	208(209)	**5.57**(7.37)	10.2(11.0)	21.6(24.8)	1.52(0.59)
1.8	27.3(29.9)	12.2(10.0)	39.3(47.4)	8.77(5.08)	24.1(12.5)	208(197)	**4.88**(5.90)	9.47(8.39)	12.7(16.5)	1.46(0.57)
2.0	22.7(22.2)	10.4(9.30)	32.4(40.6)	7.61(3.99)	23.3(15.2)	213(203)	**4.30**(4.93)	8.40(7.58)	9.55(12.5)	1.47(0.74)

papers. For TRAS, we set its parameters $r=7$, $\mu_{min}=0.5$ and $\Delta=10^{-4}$. For SASAM, we set $\boldsymbol{\theta}_1=0.1$, $\boldsymbol{\theta}_2=0.7$, $\mu_{min}=0.5$, $h=4$. As to CMAB(s), we set $\lambda=0.1$ for exponentially weighted posterior. For NAS, we set $g_{p+1}=0.1$, $\Delta=10^{-2}$, $k=1$. For R-SADA, we set $\mu_{min}=0.5$, $k=1$. For TSSRP, we set $r=7$, $\mu_{1,true}=1$.

For Poisson data, for EF-BSSCD(S) and EF-BSSCD(G), we set $\lambda=0.1$, $w_j=0.5$, $\sigma_j=2$, $v=10^{-6}$. EF-BSSCD(P) is not included due to its high computational complexity. As to other baseline methods, we set their parameters according to the recommendations in their papers. For TRAS, we set its parameters $r=10$, $\mu_{min}=0.5$ and $\Delta=10^{-3}$. For SASAM, we set $\boldsymbol{\theta}_1=0.1$, $\boldsymbol{\theta}_2=0.7$, $\mu_{min}=0.5$, $h=3$. As to CMAB(s), we set $\lambda=0.1$ for exponentially weighted posterior. For NAS, we set $g_{p+1}=0.1$, $\Delta=10^{-2}$, $k=1$. For R-SADA, we set $\mu_{min}=0.5$, $k=2$. For TSSRP, we set $r=7$, $\mu_{1,true}=0.5$.

For binomial data, for EF-BSSCD(S) and EF-BSSCD(G), we set $\lambda=0.1$, $w_j=0.3$, $\sigma_j=2$, $v=10^{-6}$. EF-BSSCD(P) is not included due to its high computational complexity. As to other baseline methods, we set their parameters according to the recommendations in their papers. For TRAS, we set its parameters $r=10$, $\mu_{min}=0.5$ and $\Delta=10^{-3}$. For SASAM, we set $\boldsymbol{\theta}_1=0.1$, $\boldsymbol{\theta}_2=0.7$, $\mu_{min}=0.5$, $h=3$. As to CMAB(s), we set $\lambda=0.1$ for exponentially weighted posterior. For NAS, we set $g_{p+1}=0.1$, $\Delta=10^{-2}$, $k=1$. For R-SADA, we set $\mu_{min}=0.5$, $k=1$. For TSSRP, we set $r=7$, $\mu_{1,true}=10$.

We set the change magnitude $\xi=0.5,0.8,1.0,1.5,2.0$. The results are shown in Table 4, 5 and 6. For Gaussian data, EF-BSSCD(S) has the smallest ADD generally, followed by EF-BSSCD(P), since they both use the spike-slab estimation of $\boldsymbol{\theta}_a$ in their test statistics, which contains sparse change information. EF-BSSCD(G) performs a little inferior to them, since it uses the maximum likelihood estimation (MLE) of $\boldsymbol{\theta}_a$ instead, and consequently cannot track the sparse change very accurately. CMAB(s) achieves good performance over the whole range of change magnitudes, but a little inferior to EF-BSSCD methods. SASAM and TSSRP perform better than TRAS since SASAM considers the spatial correlation information between variables and TSSRP balances exploration and exploitation through Thompson sampling. But R-SADA does not perform as well as

NAS under the high-dimensional cases, although it augments the correlation information between variables. In general, EF-BSSCD-based methods have much superior performance than other baselines.

For Poisson data, EF-BSSCD(S) performs best, except at one small change magnitude, $\xi = 0.5$, inferior to SASAM. EF-BSSCD(G) is a little inferior to EF-BSSCD(S) since EF-BSSCD(G) only uses the MLE of $\boldsymbol{\theta}_a$, which can not estimate the parameter accurately especially when it is a sparse vector. CMAB(s), R-SADA and SASAM perform closely, better than TRAS and NAS at most change magnitudes, since they both use the correlation information between variables. As to TSSRP, it does not detect the change due to the form of Shiryaev–Roberts–Pollak statistic is sensitive to data distributions.

For binomial data, EF-BSSCD(S) performs best, except at one small change magnitude, $\xi = 0.5$, inferior to SASAM. EF-BSSCD(G) are a little inferior to EF-BSSCD(S), due to the same reason as Poisson data. CMAB(s) and SASAM perform better than TRAS since they both considers the correlation information between variables. But R-SADA does not detect the change when the change magnitude is smaller than 1 under high-dimensional cases, such that it can not beat NAS although with the correlation information between variables. TSSRP also does not detect the change due to the form of Shiryaev–Roberts–Pollak statistic is sensitive to data distributions.

Table 4 Average Detection Delays/ADDs(Standard Deviation of Detection Delays/STDs) for Gaussian data with $p = 100$, $m = 30$

ξ	NAS	R-SADA	TRAS	CMAB(s)	SASAM	TSSRP	EF-BSSCD(S)	EF-BSSCD(G)	EF-BSSCD(P)
0.0	200(363)	200(200)	200(186)	200(184)	200(194)	200(185)	200(283)	200(296)	200(277)
0.5	9.61(10.7)	146(158)	48.5(40.4)	8.88(6.22)	47.2(43.1)	59.1(47.3)	**1.25**(0.61)	1.28(5.19)	1.26(0.60)
0.8	6.19(4.88)	33.4(39.4)	30.3(31.8)	7.23(6.22)	16.1(12.4)	17.1(8.14)	1.16(0.46)	1.22(3.17)	**1.15**(0.40)
1.0	5.23(3.37)	17.5(16.0)	24.6(25.8)	6.42(5.86)	10.3(6.56)	10.7(4.05)	**1.10**(0.34)	1.06(0.73)	1.12(0.41)
1.5	4.22(2.28)	9.88(11.6)	17.1(21.3)	5.55(5.51)	5.63(2.84)	5.86(1.45)	**1.08**(0.31)	1.09(1.22)	1.09(0.33)
2.0	3.95(1.90)	9.29(11.6)	13.1(15.8)	5.59(5.65)	4.02(1.70)	4.10(1.04)	**1.07**(0.32)	1.13(2.20)	1.08(0.35)

Table 5 Average Detection Delays/ADDs(Standard Deviation of Detection Delays/STDs) for Poisson data with $p = 100$, $m = 30$

ξ	NAS	R-SADA	TRAS	CMAB(s)	SASAM	TSSRP	EF-BSSCD(S)	EF-BSSCD(G)
0.0	200(244)	200(282)	200(155)	200(250)	200(187)	200(140)	200(216)	200(213)
0.5	109(158)	75.3(122)	76.1(46.8)	77.5(91.7)	**58.9**(47.4)	186(135)	67.3(94.5)	92.4(112)
0.8	36.8(52.5)	25.7(35.7)	35.1(33.2)	46.2(58.7)	25.6(18.3)	185(136)	**15.5**(22.3)	32.7(34.5)
1.0	23.0(28.0)	15.6(20.6)	35.3(27.0)	18.8(13.5)	17.5(10.6)	188(137)	**8.89**(13.9)	21.2(26.9)
1.5	9.31(9.69)	6.44(7.14)	20.4(16.8)	8.36(5.39)	9.10(4.77)	180(131)	**4.27**(5.29)	7.66(6.57)
2.0	4.99(3.27)	3.79(3.66)	14.5(13.3)	5.03(3.31)	5.87(2.43)	184(135)	**2.81**(1.98)	4.13(3.23)

Table 6 Average Detection Delays/ADDs(Standard Deviation of Detection Delays/STDs) for binomial data with $p=100$, $m=30$, $N=10$

ξ	NAS	R-SADA	TRAS	CMAB(s)	SASAM	TSSRP	EF-BSSCD(S)	EF-BSSCD(G)
0.0	200(282)	200(228)	200(175)	200(222)	200(180)	200(199)	200(215)	200(262)
0.5	109(158)	224(248)	69.9(43.7)	86.3(102)	**52.0**(39.1)	188(199)	57.7(69.1)	68.8(61.7)
0.8	36.8(52.5)	310(312)	44.9(30.6)	33.3(27.8)	25.4(14.0)	182(195)	**15.7**(16.8)	21.2(19.7)
1.0	23.0(28.0)	267(304)	36.0(24.7)	19.4(13.5)	19.1(10.5)	169(176)	**13.7**(12.6)	13.9(9.23)
1.5	9.31(9.69)	54.5(92.1)	25.7(19.0)	9.67(4.83)	11.7(5.28)	170(174)	**5.76**(4.09)	7.98(5.02)
2.0	4.99(3.27)	16.1(19.1)	20.8(16.0)	6.63(3.39)	8.55(3.11)	183(189)	**3.24**(2.44)	5.73(2.68)

Appendix E: Estimation of $\boldsymbol{\theta}_0$ and $\boldsymbol{\Sigma}_0$ from Historical Before-Change Data

In practice, we can estimate $\boldsymbol{\theta}_0$ and $\boldsymbol{\Sigma}_0$ from historical reference samples in Phase I. Assume we have N_0 offline in-control samples, $\mathbf{X}_t$, $t=1,\ldots,N_0$.

For Gaussian distribution, the fitted coefficients can be estimated using MLE, i.e., $\hat{\boldsymbol{\theta}}_t = (\mathbf{B}_b'\mathbf{B}_b)^{-1}\mathbf{B}_b'\mathbf{X}_t$ and the fitted residuals are $\hat{\mathbf{E}}_t = (\mathbf{I}-\mathbf{B}_b(\mathbf{B}_b'\mathbf{B}_b)^{-1}\mathbf{B}_b')\mathbf{X}_t$, $t=1,\ldots,N_0$. Therefore, $\hat{\boldsymbol{\theta}}_0 = \frac{1}{N_0}\sum_{t=1}^{N_0}\hat{\boldsymbol{\theta}}_t$ and $\hat{\boldsymbol{\Sigma}}_0 = \mathrm{Cov}(\hat{\boldsymbol{\theta}}_1,\ldots,\hat{\boldsymbol{\theta}}_{N_0})$. If $\boldsymbol{\Sigma}_0$ is assumed as diagonal matrix,i.e., $\boldsymbol{\Sigma}_0=\sigma_0^2\mathbf{I}$, we estimate σ_0 as $\hat{\sigma}_0 = \sqrt{\frac{1}{k_b}\sum_{j=1}^{k_b}\hat{\boldsymbol{\Sigma}}_{0jj}}$. Another parameter special to Gaussian distribution is $\boldsymbol{\Sigma}_e=\sigma_e^2\mathbf{I}$. It is estimated as $\hat{\sigma}_e = \sqrt{\frac{1}{N_0 p}\sum_{t=1}^{N_0}||\hat{\mathbf{E}}_t-\bar{\mathbf{E}}||^2}$, where $\bar{\mathbf{E}}$ is the mean of $\hat{\mathbf{E}}_t$.

For Poisson distribution and binomial distributions, to obtain MLE for $\boldsymbol{\theta}_t$, $t=1,\ldots,N_0$, we can derive the first derivative and the second derivative and use Newton's method to obtain the MLEs iteratively(Hardin et al. 2012). The log-likelihood is $\mathcal{L}(\mathbf{X}_t) = (\mathbf{X}_t \oslash a(\boldsymbol{\phi}_t))'\mathbf{B}_b\boldsymbol{\theta}_t - \mathbf{1}_m'(b(\mathbf{B}_b\tilde{\boldsymbol{\theta}}_t)\oslash a(\boldsymbol{\phi}_t))$. Specifically, the derivatives of $\boldsymbol{\theta}_t$ are

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}_t} = \mathbf{B}_b'(\mathbf{X}_t\oslash a(\boldsymbol{\phi}_t)) - \mathbf{B}_b'(b'(\mathbf{B}_b\boldsymbol{\theta}_t)\oslash a(\boldsymbol{\phi}_t)), \tag{4}$$

$$\frac{\partial^2\mathcal{L}}{\partial\boldsymbol{\theta}_t^2} = -\mathbf{B}_b'\big(b''(\mathbf{B}_b\boldsymbol{\theta}_t)\oslash a(\boldsymbol{\phi}_t)\big)\mathbf{B}_b. \tag{5}$$

Then update $\boldsymbol{\theta}_t$ according to Newton's method until convergence:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_t - (\frac{\partial^2\mathcal{L}}{\partial\boldsymbol{\theta}_t^2})^{-1}\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}_t}. \tag{6}$$

After we get the MLE $\hat{\boldsymbol{\theta}}_t$, $t=1,\ldots,N_0$, $\boldsymbol{\theta}_0$ and $\boldsymbol{\Sigma}_0$ are estimated the same as Gaussian distribution.

Appendix F: Computational Complexity of EF-BSSCD

Here, we analyze the computational complexity of our proposed method. For the computational complexity of Algorithm 1, since the gradients are of different forms for different distributions, it is case by case, e.g., $O\big((mk_ak_b+mk_a^2+k_b^2+mk_b)n_{iter2}\big)$ under Gaussian, $O\Big(\big((mk_a^2+m^2k_a+$

$mk_b)n_{iter1} + k_b^2 + mk_b)n_{iter2}\Big)$ under Poisson and binomial, where n_{iter1} is the number of iterations to guarantee convergence for $\boldsymbol{\mu}, \boldsymbol{\alpha}, \mathbf{s}$ using gradient ascent method to maximize the evidence lower bound, and n_{iter2} is the number of iterations to guarantee convergence between $\boldsymbol{\theta}_n$ and $\boldsymbol{\theta}_a$.

For Thompson sampling procedure, it is $O(pk_b + pk_a + p\ln p)$ based on SLRT and GLRT. Based on PBF, it is $O(pk_a^2 + pk_b + p\log p)$ for Gaussian, $O(p^m(mk_a^6 + mk_b)n_{iter3})$ for Poisson and binomial, where n_{iter3} is the number of iterations to compute $\tilde{\boldsymbol{\theta}}_a$, which is explained in Appendix B.

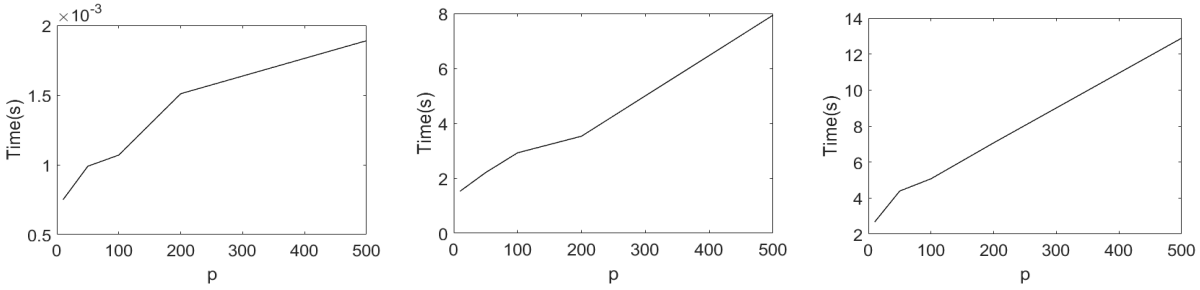
For computing the three test statistics, the complexity of SLRT is $O(mk_a + mk_b)$, and GLRT is $O((mk_a^2 + mk_b + k_a^3)n_{iter4})$, where n_{iter4} is the number of iterations to compute $\hat{\boldsymbol{\theta}}_a$ and PBF is $O(mk_a^2 + mk_b)$ for Gaussian, $O(2^{k_a}(mk_a^6 + mk_b)n_{iter3})$ for Poisson and binomial.

Take a specific case for example. For a process with $p = 100$, $k_a = 26$, $k_b = 3$, for an experiment setting with $n_{iter1} = 1000$, $n_{iter2} = 30$, $n_{iter3} = 40$, $n_{iter4} = 40$ and $m = 30$ using our computer with Intel(R) Core(TM) i9-9900KF CPU @ 3.60GHz, we summarize the processing time for one time point observation in Table 7. To show how the computation complexity changes with the

Table 7 Processing time

	SLRT	GLRT	PBF
Gaussian	0.001694s	0.001814s	0.001806s
Poisson	4.564s	4.756s	None
binomial	11.41s	12.99s	None

dimension of variables, we take EF-BSSCD(S) as an example. We set six levels for p, i.e., $p = 10, 50, 100, 200, 500$, $m = 0.1p$, $k_b = 2, 2, 3, 4, 5$, $k_a = 6, 11, 14, 21, 26$ respectively and compute the average processing time for EF-BSSCD(S) for one time point using 100 observations under Gaussian, Poisson and binomial distributions, and plot them in Figure 1. As we can see, the processing time changes almost linearly with p, which is acceptable for practical use.



(a) Gaussian (b) Poisson (c) Binomial

Figure 1 Processing time varies with p under different distributions

Appendix G: Sensitivity Analysis

We conduct sensitivity analysis to see the performance of our method under different numbers of bases, types of bases, prior parameters, mis-specified distribution or mis-specified bases and different extents of colinearity in bases. We only apply EF-BSSCD(S) for change detection, and the regular patterns are almost the same for EF-BSSCD(G) and EF-BSSCD(P).

For different numbers of bases, we consider the following three cases: generating Poisson data by setting 1) $\mathbf{B}_a \in \mathcal{R}^{10\times 7}$ as 7 four-order B-spline bases with 11 equally spaced knots, 2) $\mathbf{B}_a \in \mathcal{R}^{10\times 5}$ as 5 four-order B-spline bases with 9 equally spaced knots and 3) $\mathbf{B}_a \in \mathcal{R}^{10\times 3}$ as 3 four-order B-spline bases with 7 equally spaced knots. The detection results are shown in Table 8. Generally, we see that for the same type of bases, the detection delay is increasing with the number of bases. The more bases we have, the more parameters we need to estimate and the larger set of observations we need to obtain. So under the same size of observation set, i.e., $m=3$, we can get a better estimation for $\boldsymbol{\theta}_a$ with $k_a=3$ than $k_a=5$ and $k_a=7$, especially when the change magnitude is small so that the change is difficult to detect.

For different types of $\mathbf{B}_a$, we consider the following three cases: generating binomial data by setting 1) $\mathbf{B}_a \in \mathcal{R}^{32\times 12}$ as 12 three-order B-spline bases with 15 equally spaced knots, 2) $\mathbf{B}_a \in \mathcal{R}^{32\times 12}$ as 12 lowest frequency Fourier bases and 3) $\mathbf{B}_a \in \mathcal{R}^{32\times 12}$ as the first 12 Vaidyanathan wavelet bases with a grid of 32 equally spaced sensing points in $(0,1)$. Also, to guarantee their detection performance can be compared fairly, we also normalize these bases to make their Euclidean norm equal to 1. The detection results are shown in Table 9. We can see that the ADDs for B-spline bases are less than those for Fourier bases and Wavelet bases. It is because Fourier bases and Wavelet bases are both orthogonal bases, but B-spline bases have colinearities between bases. When colinearities exist, the elements of $\boldsymbol{\theta}_a$ will be correlated as well. Then when a change happens on one of these colinear bases, their coefficients would be nonzero values at the same time, which can be used jointly to detect the change. So the test statistics will be larger and trigger alarm more quickly. However, too strong colinearity should be avoided when we design $\mathbf{B}_a$ or $\mathbf{B}_b$ since the coefficients of those strongly correlated bases will be strongly correlated as well, which will degrade the diagnosis of the change, i.e., determining which variables are anomalous according to the estimated values on the bases.

For different settings of prior parameters, we set $w=0.01, 0.1, 0.9$ for Gaussian data and set $p=100$, $m=5$, with $\mathbf{B}_b \in \mathcal{R}^{100\times 3}$ are three lowest frequency Fourier bases and $\mathbf{B}_a \in \mathcal{R}^{100\times 52}$ are 52 four-order B-spline bases with 56 equally spaced knots. The detection results are shown in Table 10. As we can see, our results are not much influenced by w. But if we set it as an unrealistic value, such as 0.9, meaning there is a very large probability to change for every basis, the detection results will degenerate a little. Therefore, we'd better set w as the chance of change for every

basis according to our prior knowledge and recommend $w < 0.5$. Then we also set $\sigma_j = 0.1, 3, 10$ for Gaussian data with the same setting as sensitivity analysis for w and the detection results are in Table 11. We find the results are not much influenced by σ_j when it is in a realistic range. But too small σ_j may weaken the detection of large magnitude shift since it prefers to generate small $\boldsymbol{\mu}$, although it benefits the detection of small magnitude shift. As it is suggested in Appendix D, with prior information about the maximum change magnitude, we would better set $\sigma_j \geq |\xi|$ (within 1-sigma of $N(0, \sigma_j)$) to guarantee that there is a high probability to accommodate all plausible $\theta_{aj} \neq 0$ in the slab distribution. In a word, our results are not much influenced by the setting of w and σ_j in spike-slab prior, since the impact of the prior is diluted in the weighted posterior $p\left(\boldsymbol{\theta}_a, \mathbf{r} | \mathbf{X}_{Z(1)}, \ldots, \mathbf{X}_{Z(n)}\right) \propto p_0(\boldsymbol{\theta}_a, \mathbf{r}) \prod_{t=1}^{n} p(\mathbf{X}_{Z(t)} | \boldsymbol{\theta}_a, \mathbf{r})^{\lambda_{nt}}$ by a large number of sequential observations from $t = 1$ to $t = n$.

Table 8 Average Detection Delays/ADDs(Standard Deviation of Detection Delays/STDs) for Poisson data with $p = 10$, $m = 3$ for sensitivity analysis for number of bases

ξ	7 four-order B-spline $\mathbf{B}_a$	5 four-order B-spline $\mathbf{B}_a$	3 four-order B-spline $\mathbf{B}_a$
0.0	200(212)	200(265)	200(291)
0.2	151(180)	131(199)	103(147)
0.4	80.9(104)	81.8(118)	45.9(72.3)
0.6	61.9(88.7)	33.8(56.0)	26.4(47.3)
0.8	33.5(45.6)	17.4(24.2)	13.9(19.1)
1.0	22.0(34.8)	9.85(15.6)	10.1(15.5)
1.2	12.3(16.1)	6.47(10.5)	6.65(10.3)
1.4	11.0(14.3)	5.26(6.04)	3.26(4.25)
1.6	7.09(8.74)	3.59(4.25)	3.34(5.04)
1.8	6.67(9.29)	2.92(3.35)	2.32(2.48)
2.0	3.87(4.22)	2.59(2.70)	2.40(3.04)

Table 9 Average Detection Delays/ADDs(Standard Deviation of Detection Delays/STDs) for binomial data with $p = 32$, $m = 8$ for sensitivity analysis for type of $\mathbf{B}_a$

ξ	B-spline $\mathbf{B}_a$	Fourier $\mathbf{B}_a$	Wavelet $\mathbf{B}_a$
0.0	200(297)	200(303)	200(306)
0.2	31.9(60.0)	100(192)	128(247)
0.4	16.4(36.3)	75.1(162	98.3(150)
0.6	9.38(22.1)	30.8(85.6)	25.2(51.4)
0.8	7.00(14.2)	31.0(105)	18.9(33.7)
1.0	3.64(6.33)	15.0(32.1)	10.5(17.9)
1.2	2.36(2.56)	9.59(24.3)	5.39(7.48)
1.4	2.22(2.29)	5.33(15.4)	2.80(3.30)
1.6	1.91(2.00)	3.55(7.73)	2.71(2.59)
1.8	1.76(1.33)	3.34(9.54)	2.41(2.91)
2.0	1.71(1.59)	2.29(4.12)	1.94(1.40)

Table 10 **Average Detection Delays/ADDs(Standard Deviation of Detection Delays/STDs) for Gaussian data with $p = 100$, $m = 5$ for sensitivity analysis for w**

ξ	$w = 0.01$	$w = 0.1$	$w = 0.9$
0.0	200(183)	200(200)	200(219)
0.1	136(123)	142(130)	160(174)
0.2	43.1(31.8)	48.0(39.3)	50.6(49.3)
0.3	22.5(15.8)	23.0(14.5)	25.6(21.1)
0.4	17.5(11.1)	17.3(11.5)	19.6(15.5)
0.5	12.5(8.44)	14.2(9.78)	16.0(12.7)
0.6	12.9(8.85)	12.3(9.05)	14.6(12.1)
0.7	11.2(7.32)	11.3(8.24)	13.5(11.0)
0.8	10.7(7.82)	10.4(7.73)	12.2(10.5)
0.9	10.1(7.40)	10.0(7.49)	12.1(10.0)
1.0	9.76(6.89)	9.93(7.21)	11.4(9.66)

Table 11 **Average Detection Delays/ADDs(Standard Deviation of Detection Delays/STDs) for Gaussian data with $p = 100$, $m = 5$ for sensitivity analysis for σ_j**

ξ	$\sigma_j = 0.1$	$\sigma_j = 3$	$\sigma_j = 10$
0.0	200(232)	200(200)	200(194)
0.1	117(137)	142(130)	146(138)
0.2	36.6(43.1)	48.0(39.3)	46.9(39.9)
0.3	24.0(34.0)	23.0(14.5)	22.9(15.2)
0.4	19.6(26.5)	17.3(11.5)	17.2(11.4)
0.5	17.7(28.0)	14.2(9.78)	14.0(9.54)
0.6	15.9(33.1)	12.3(9.05)	12.4(8.70)
0.7	16.3(34.5)	11.3(8.24)	11.7(8.27)
0.8	13.7(23.4)	10.4(7.73)	11.0(7.99)
0.9	12.7(21.6)	10.0(7.49)	10.4(7.59)
1.0	11.2(17.0)	9.93(7.21)	9.86(7.51)

For mis-specification of type of distributions, actually, the forms of the variational Bayesian estimation and the test statistics are both derived based on the specific forms of distributions in the exponential family. So in practical use, we'd better assume the distribution type is known in advance and use the specific forms for the distribution. We can imagine that if the distribution is mis-specified, the detection results would degenerate. This is a well-known limitation for parametric statistical monitoring methods. Only nonparametric methods can solve this problem and achieve distribution-free performance(Härdle 1990). However, this would be a totally different story and we would like to put it as future work. For reviewer's information, we conduct such experiments that the distribution type is mis-specified. In particular, for sensitivity analysis, we only apply EF-BSSCD(S) for change detection, the regular patterns are the same for EF-BSSCD(G) and EF-BSSCD(P). We generate Poisson data according to Section 5.2 in our paper, and apply EF-BSSCD(S) derived from Gaussian distribution for change detection. The detection results are given as in Table 12. The detection results are shown in the column named "Poisson data using

Gaussian". As we can see, the ADDs are not as small as "Poisson data using Poisson" (the method specific for Poisson distribution), since the derivations for Gaussian are quite different from those for Poisson.

Table 12 Average Detection Delays/ADDs(Standard Deviation of Detection Delays/STDs) for Poisson data with $p = 10$, $m = 3$ for mis-specified distribution

ξ	Poisson data using Poisson	Poisson data using Gaussian
0.0	200(212)	200(204)
0.2	151(180)	134(148)
0.4	80.9(104)	129(119)
0.6	61.9(88.7)	116(116)
0.8	33.5(45.6)	103(106)
1.0	22.0(34.8)	58.3(62.5)
1.2	12.3(16.1)	51.6(61.4)
1.4	11.0(14.3)	51.1(56.1)
1.6	7.09(8.74)	34.1(44.8)
1.8	6.67(9.29)	31.8(39.7)
2.0	3.87(4.22)	16.2(17.7)

For the mis-specification of type or number of bases, in contrast, our method has robust performance to cases if the anomaly dictionary is mis-specified, which means we can not utilize the correlation information between variables correctly. To illustrate that, we conduct one experiment. We generate Poisson data as Section 5.2., where $\mathbf{B}_a \in \mathcal{R}^{10\times 7}$ are 7 four-order B-spline bases with 11 equally spaced knots. But we use $\mathbf{B}_a \in \mathcal{R}^{10\times 3}$ are 3 three-order B-spline bases with 6 equally spaced knots to detect the change, and the detection results are shown in the column named "Three-order B-spline $\mathbf{B}_a$" in Table 13. The other column is from the manuscript with the correctly specified anomaly bases. As we can see, since the correlation we utilize is not correct, the detection is a little slower than using correctly specified $\mathbf{B}_a$. But it is not influenced too much. So we can conclude that the mis-specification of basis function will not influence the efficiency of our proposed method too much.

As for the influence of colinearity in $\mathbf{B}_a$, when different bases in $\mathbf{B}_a$ have strong colinearity, the elements of $\boldsymbol{\theta}_a$ will be strongly correlated. Then when a change happens on one of these colinear bases, their coefficients would be nonzero values at the same time, which can be used jointly to detect the change. So the change detection in this circumstance will not be delayed, but the later diagnosis of the change would be not accurate, i.e., determining which variables are anomalous according to the estimated values on the bases. Here we use one experiment to show that. We consider two cases, 1) generating Poisson data by setting $\mathbf{B}_a \in \mathcal{R}^{10\times 4}$ as 4 four-order B-spline bases with 8 knots. These four bases have strong colinearity; 2) generating Poisson data by selecting $\mathbf{B}_a \in \mathcal{R}^{10\times 4}$ as $1^{st}, 3^{rd}, 10^{th}, 12^{th}$ bases from 12 four-order B-spline bases with 16 knots. These four

Table 13 **Average Detection Delays/ADDs(Standard Deviation of Detection Delays/STDs) for Poisson data with $p=10$, $m=3$ for mis-specified anomaly dictionary**

ξ	Three-order B-spline $\mathbf{B}_a$	Four-order B-spline $\mathbf{B}_a$
0.0	200(241)	200(212)
0.2	150(180)	151(180)
0.4	117(145)	80.9(104)
0.6	76.7(98.3)	61.9(88.7)
0.8	45.3(71.0)	33.5(45.6)
1.0	30.2(42.8)	22.0(34.8)
1.2	19.1(27.9)	12.3(16.1)
1.4	13.7(24.1)	11.0(14.3)
1.6	11.4(17.6)	7.09(8.74)
1.8	6.84(10.5)	6.67(9.29)
2.0	5.91(8.55)	3.87(4.22)

bases are almost orthogonal. The detection results are shown in Table 14. As we can see, the colinearity in $\mathbf{B}_a$ does not influence the ADDs, but achieves even smaller ADDs. However, in practice, for efficiently diagnosing the change in the downstream task, it is still better to remove the colinearity and try to construct almost orthogonal $\mathbf{B}_a$ to capture different change features.

Table 14 **Average Detection Delays/ADDs (Standard Deviation of Detection Delays/STDs) for Poisson data with $p=10$, $m=3$ for the colinearity of $\mathbf{B}_a$**

ξ	Highly colinear $\mathbf{B}_a$	Almost orthogonal $\mathbf{B}_a$
0.0	200(246)	200(232)
0.2	119(184)	149(184)
0.4	45.4(80.3)	58.2(86.7)
0.6	23.3(37.5)	54.2(77.1)
0.8	12.2(22.5)	15.0(21.6)
1.0	8.46(14.6)	12.3(20.3)
1.2	5.33(9.40)	7.43(10.6)
1.4	3.67(4.81)	4.20(5.43)
1.6	2.40(3.15)	5.19(6.36)
1.8	2.53(3.11)	3.53(3.66)
2.0	2.09(2.64)	3.14(3.95)

References

Chang SI, Yadama S (2010) Statistical process control for monitoring non-linear profiles using wavelet filtering and b-spline approximation. *International Journal of Production Research* 48(4):1049–1068.

Collins M, Dasgupta S, Schapire RE (2001) A generalization of principal components analysis to the exponential family. *Advances in neural information processing systems* 14.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) Bayesian data analysis .

George EI, McCulloch RE (1993) Variable selection via gibbs sampling. *Journal of the American Statistical Association* 88(423):881–889.

George EI, McCulloch RE (1997) Approaches for bayesian variable selection. *Statistica sinica* 339–373.

Hardin JW, Hilbe JW, et al. (2012) Generalized linear models and extensions. *Stata Press books* .

Härdle W (1990) *Applied nonparametric regression*. Number 19 (Cambridge university press).

Ishwaran H, Rao JS, et al. (2005) Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics* 33(2):730–773.

Montgomery DC (2007) *Introduction to statistical quality control* (John Wiley & Sons).

Paynabar K, Jin J (2011) Characterization of non-linear profiles variations using mixed-effect models and wavelets. *Iie transactions* 43(4):275–290.

Raudenbush SW, Yang ML, Yosef M (2000) Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. *Journal of computational and Graphical Statistics* 9(1):141–157.

Ročková V, George EI (2014) Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association* 109(506):828–846.

Wang Y, Blei DM (2019) Frequentist consistency of variational bayes. *Journal of the American Statistical Association* 114(527):1147–1161.

Yan H, Paynabar K, Shi J (2017) Anomaly detection in images with smooth background via smooth-sparse decomposition. *Technometrics* 59(1):102–114.

Zhang C, Chen N, Wu J (2020) Spatial rank-based high-dimensional monitoring through random projection. *Journal of Quality Technology* 52(2):111–127.

Zhou Q, Zou C, Wang Z, Jiang W (2012) Likelihood-based ewma charts for monitoring poisson count data with time-varying sample sizes. *Journal of the American Statistical Association* 107(499):1049–1062.

Zou C, Wang Z, Tsung F (2012) A spatial rank-based multivariate ewma control chart. *Naval Research Logistics (NRL)* 59(2):91–110.