A Preliminary

A.1 Change Point Detection Methods

WBSLSW: The WBSLSW method (Korkas & PryzlewiczV, 2017) incorporates the non-parametric locally stationary wavelet process with wild binary segmentation, and can detect the second-order structure of the sequence with an unknown number of change points. We implement the WBSLSW using the R package wbsts.

KCP: The KCP (Harchaoui & Cappé, 2007) method is a dynamic programming method with a known number of change points. This algorithm detects the change points by minimizing the kernel least-squares criterion. In our cases, we combine KCP with a linear penalty pruning the number of change points. This is done by using the python package ruptures (Truong et al., 2020) with a Gaussian kernel and default parameters.

DPHMM: The DPHMM method developed by Ko et al. (2015) is a combination of the Dirichlet process and the Hidden Markov model to detect change points using MCMC. This method allows the number of change points to be unknown. We implement this algorithm using the R package dirichletprocess with default parameters and set $\tilde{K} = 10$

ECP3O: Zhang et al. (2017) propose a new change point search framework called change point procedure via pruned objectives. The ECP3O method uses the new search frame with the E-statistics which measures the goodness-of-fit. This method is implemented using e.cp3o-delta function with default parameters in R package ecp. (Nicholas A. James and Wenyu Zhang and David S. Matteson, 2019). The maximum number of change points $\tilde{K}$ is 10.

$\mathcal{D}_m$-BOCD: $\mathcal{D}_m$-BOCD is an online method developed by Altamirano et al. (2023). This method is generalized from BOCD (Adams & MacKay, 2007) with diffusion score matching, which is robust to sequences with outliers. We implement this method using the code provided on the author's GitHub page.

A.2 Variational Inference

Variational inference (VI) Blei et al. (2017) works as a fast approximation method for Bayesian inference. Given the observation $\mathbf{x}$ and latent variable $\mathbf{z}$, VI uses a tractable variational distribution q drawn from the function class $\mathcal{F}$ to approach the complicated posterior $p(\mathbf{z} \mid \mathbf{x})$ by minimizing their KL divergence. However, the KL can not be computed analytically. Classical VI optimizes an alternative objective called Evidence Lower Bound (ELBO) that is equivalent to log marginal likelihood minus the KL:

$$\begin{aligned} \ln p(\mathbf{x}) &= \mathrm{ELBO}(q) + \mathrm{KL}(q \| p(\mathbf{z} \mid \mathbf{x})) \\ &= \int q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right\} \mathrm{d}\mathbf{z} - \int q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{z} \mid \mathbf{x})}{q(\mathbf{z})} \right\} \mathrm{d}\mathbf{z}. \end{aligned}$$

Under the common mean-field assumption $q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j)$, the maximizer of ELBO q_j^* has the analytical solution $q_j^*(z_j) \propto \exp\left\{\mathbb{E}_{\mathbf{z}_{-j}}[\log p(\mathbf{z}, \mathbf{x})]\right\}$, where $\mathbf{z}_{-j}$ denotes all variables z_i other than z_j, that can be solved by the coordinate ascent algorithm.

Recently, VI has been commonly applied in training deep generative models, including VAE (Kingma & Welling, 2013) and deep diffusion model (Ho et al., 2020) to approximate complex posterior distributions. VI plays a crucial role in approximating the posterior distribution over the latent variables, enabling efficient learning and generation of high-quality samples from complex data distributions.

B Normal Mean-Variance Shift Model

In this section, we derive updating formulas for the Normal Mean-Variance Shift model. Denoting the set of all latent variables as $\boldsymbol{\xi} = \{\{\mathbf{t}_k\}_{k=1}^{K}, \{\theta_{K+1}\}_{k=1}^{K+1}\}$ and the hyperparameters set $\alpha_k =$

$\{\beta, \nu_0, V_0\}$, we assume a constrained mean-field Q family in variational inference:

$$Q(\boldsymbol{\xi}) = Q(\mathbf{t}_1) \prod_{k=2}^{K} Q\left(\mathbf{t}_k | \mathbf{t}_{k-1}\right) \prod_{k}^{K+1} Q\left(u_k\right) Q(\Lambda_k).$$

Then the variational lower bound is given by

$$\begin{aligned}\mathcal{L}(Q) \;=\;& \sum_{k-1}^{K+1} \left[\sum_{t_k, t_{k-1}} Q\left(\mathbf{t}_k, \mathbf{t}_{k-1}\right) \int Q\left(u_k\right) Q(\Lambda_k) \ln p\left(\mathbf{Y}_k \mid \mathbf{t}_k, \mathbf{t}_{k-1}, u_k, \Lambda_k\right) du_k d\Lambda_k \right] \\ &+ \sum_{k=1}^{K+1} \left[\int Q\left(u_k\right) Q(\Lambda_k) \ln \frac{\mathcal{N}\left(u_k; 0, \beta^{-1} I\right) \mathcal{W}\left(\Lambda_k; \nu^0, V^0\right)}{Q\left(u_k\right) Q_T(\Lambda_k)} du_k d\Lambda_k \right] \\ &+ \sum_{k=1}^{K} \left[\sum_{t_k, t_{k-1}} Q_t\left(\mathbf{t}_k, \mathbf{t}_{k-1}\right) \ln \frac{p\left(\mathbf{t}_k \mid \mathbf{t}_{k-1}\right)}{Q_t\left(\mathbf{t}_k \mid \mathbf{t}_{k-1}\right)} \right].\end{aligned}$$

Minimizing KL divergence leads to an analytical solution. We can directly apply it to give the optimal solutions for the family of factors Q of variational posteriors

$$Q(\mathbf{t}_1) = \prod_{i=1}^{N} \tilde{\pi}_{1,i}^{\mathbf{t}_1(i)}, \quad Q\left(\mathbf{t}_k | \mathbf{t}_{k-1}\right) = \prod_{i=1}^{N} \prod_{j=1}^{N} \hat{\pi}_{k,i,j}^{\mathbf{t}_k(i) \times \mathbf{t}_{k-1}(j)},$$
$$Q(u_k) = \mathcal{N}\left(u_k \mid m_k, L_k^{-1}\right), \quad Q(\Lambda_k) = \mathcal{W}\left(\Lambda_k \mid \nu_k, V_k\right).$$

Given prior distributions defined above, solutions for variational parameters are given by

$$\begin{aligned} m_k \;=\;& \left[\langle \Lambda_k \rangle \sum_{n=1}^{N} \sum_{m \geq n}^{N} Q\left(\mathbf{t}_k(m), \mathbf{t}_{k-1}(n)\right) \sum_{j=n}^{m} \mathbf{1} + \mathbf{I}/\beta \right]^{-1} \\ & \times \left[\langle \Lambda_k \rangle \sum_{n=1}^{N} \sum_{m \geq n}^{N} Q\left(\mathbf{t}_k(m), \mathbf{t}_{k-1}(n)\right) \sum_{j=n}^{m} y_j \right], \\ L_k \;=\;& \langle \Lambda_k \rangle \sum_{n=1}^{N} \sum_{m \geq n}^{N} Q\left(\mathbf{t}_i(m), \mathbf{t}_{i-1}(n)\right) \sum_{j=n}^{m} \mathbf{1} + \alpha^{-1} \mathbf{I}, \\ \nu_i \;=\;& \nu_0 + \sum_{n=1}^{N} \sum_{m \geq n}^{N} Q\left(\mathbf{t}_i(m), \mathbf{t}_{i-1}(n)\right) \sum_{j=n}^{m} \mathbf{1}, \end{aligned}$$

and

$$V_k^{-1} \;=\; V_0^{-1} + \sum_{n=1}^{N} \sum_{m \geq n}^{N} Q\left(\mathbf{t}_k(m), \mathbf{t}_{k-1}(n)\right) \left[\sum_{j=n}^{m} y_j y_j^{\top} - 2 \sum_{j=n}^{m} y_j \left\langle u_k \right\rangle^{\top} + \sum_{j=n}^{m} \langle u_k u_k^{\top} \rangle \right].$$

As we mentioned in **Section** 2.2, solutions for $Q_t\left(\mathbf{t}_k \mid \mathbf{t}_{k-1}\right)$ can be obtained through sum-product algorithm. The updating formulas are given by

$$Q\left(\mathbf{t}_k(n) = 1\right) \;=\; \mu_{\rightarrow \mathbf{t}_k}(n) \cdot \mu_{\mathbf{t}_k \leftarrow}(n) = \tilde{\pi}_{k,n},$$

and

$$\begin{aligned} & Q\left(\mathbf{t}_{k-1}(m) = 1, \mathbf{t}_k(n) = 1\right) \\ =\; & \mu_{\rightarrow \mathbf{t}_{k-1}}(m) \cdot \pi_{i,m,n} \cdot \exp\left(\mathrm{E}_{Q(u_k)Q(\Lambda_k)} \ln p(\mathbf{Y}_k, \mathbf{t}_k, u_k, \Lambda_k \mid \mathbf{t}_{k-1})\right) \cdot \mu_{t_k \leftarrow}(n), \end{aligned}$$

where messages are obtained recursively. For $k = 1, \ldots, K$,

$$\begin{aligned} \mu_{\rightarrow \mathbf{t}_k}(n) \;=\;& \sum_{m=1}^{n} \Bigg\{ \mu_{\rightarrow \mathbf{t}_{k-1}}(m) \cdot \pi_{k,m,n} \cdot \exp \Bigg\{ \sum_{j=m}^{n} \frac{1}{2} \langle \ln |\Lambda_k| \rangle \\ & - \frac{1}{2} \sum_{j=m}^{n} \operatorname{Tr}\left[\left(y_j \cdot y_j^{\top} - y_j \cdot \left\langle u_k^{\top} \right\rangle - \langle u_k \rangle \cdot y_j^{\top} + \left\langle u_k, u_k^{\top} \right\rangle \right) \cdot \langle \Lambda_k \rangle \right] \Bigg\} \Bigg\}, \end{aligned}$$

and

$$\begin{aligned} \mu_{\mathbf{t}_{k-1}\leftarrow}(m) \;&=\; \sum_{n=m}^{N}\Bigg\{\mu_{\mathbf{t}_{k-1}\leftarrow}(n)\cdot\pi_{k,m,n}\cdot\exp\Bigg\{\sum_{j=m}^{n}\frac{1}{2}\left\langle \ln|\Lambda_k|\right\rangle \\ &-\frac{1}{2}\sum_{j=m}^{n}\mathrm{Tr}\left[\left(y_j\cdot y_j^{\top}-y_j\cdot\langle u_k^{\top}\rangle+\langle u_k,u_k^{\top}\rangle\right)\cdot\langle\Lambda_k\rangle\right]\Bigg\}\Bigg\}. \end{aligned}$$

To start recursion, the initial message state $\mu_{\rightarrow\mathbf{t}_1}$ and $\mu_{\mathbf{t}_K\leftarrow}$ are given by

$$\begin{aligned} \mu_{\rightarrow\mathbf{t}_1}(m) \;&=\; \pi_{1,m}\exp\Bigg\{\sum_{j=m}^{n}\frac{1}{2}\left\langle \ln|\Lambda_1|\right\rangle \\ &-\frac{1}{2}\sum_{j=m}^{n}\mathrm{Tr}\left[\left(y_j\cdot y_j^{\top}-y_j\cdot\langle u_1^{\top}\rangle-\langle u_1\rangle\cdot y_j^{\top}+\langle u_1,u_1^{\top}\rangle\right)\cdot\langle\Lambda_1\rangle\right]\Bigg\}, \end{aligned}$$

and

$$\begin{aligned} \mu_{\mathbf{t}_K\leftarrow}(m) \;&=\; \exp\Bigg\{\sum_{j=m}^{n}\frac{1}{2}\left\langle \ln|\Lambda_{K+1}|\right\rangle \\ &-\frac{1}{2}\sum_{j=m}^{n}\mathrm{Tr}\left[\left(y_j\cdot y_j^{\top}-y_j\cdot\langle u_{K+1}^{\top}\rangle+\langle u_{K+1},u_{K+1}^{\top}\rangle\right)\cdot\langle\Lambda_{K+1}\rangle\right]\Bigg\}, \end{aligned}$$

where we have assumed:

$$\langle u_k\rangle = m_k,\quad \langle u_k u_k^{\top}\rangle = m_k m_k^{\top}+L_k^{-1},\quad \langle\Lambda_k\rangle=\nu_k V_k,$$

$$\langle \ln|\Lambda_k|\rangle=\sum_{j=1}^{D}\psi\left(\frac{u_k+1-j}{2}\right)+D\ln 2+\ln|V_k|,$$

and $\psi(\cdot)$ is the digamma function.

C Proof of **Theorem** 1

To start up, it's worth mentioning that practically for a sequence of time T, we observe finite data points $\{y_1,...,y_T\}$ at each time stamp $t=1,...,T$, which is the input for the proposed algorithm. However, in theory, we consider a continuous timeline and there are infinitely many data points between any time intervals $[m,n]\subseteq[0,T]$. Thus before discussing our theoretical results, we first list our setup and assumptions:

A1: The underlying sequence on time interval $[0,T]$ consists of K change points $0<T_1<...<T_K<T$ with $T_0=0$ and $T_{K+1}=T$. For any time stamp $T_{k-1}<t<T_k$, the random function $\mathrm{y}(t):\mathbb{R}\rightarrow\mathbb{R}^D$ represents the sample drawn from $\mathcal{N}(y\mid u_k,S_k)$ at time t.

A2: The total number of collected observations is N. For any time interval $[m,n]\subseteq[0,T]$, the number of observations within this interval equals $O(N^{\frac{n-m}{T}})$.

A3: The algorithm initializes $M_{K+1}>K+1$ regimes corresponding to $\{\mathbf{t}_i\}_{i=1}^{M_{K+1}-1}$ change points. The regimes are segmented by a time subset $\{t_1,...,t_{M_{K+1}-1}\}$ with equidistance, such that $t_{i+1}-t_i=\frac{T}{M_{K+1}}$. Based on the characteristic of the regime between $[t_i,t_{i+1}]$, we can further categorize $\{\mathbf{t}_i\}_{i=1}^{M_{K+1}-1}$ into two subsets:

- Any $\mathbf{t}_i\in\{\mathbf{t}_{M_k}\}_{k=1}^{K}$ denotes the junction points, e.g In initialization, there is a true change point T_k located within the interval $[t_{i-1},t_i]$ and $\mathrm{y}(t)$ for $t\in[t_i,t_{k+1}]$ does not identically distributes.
- For $k=1,...,K+1$, any $i\in\{M_{k-1}+1,...,M_k-1\}$ denotes the non-junction index and we have $T_{k-1}<t_i<T_k$, where we let $M_0=0$. Every $\mathrm{y}(t)$ for $t\in[t_i,t_{k+1}]$ distributes equivalently with those in $[T_k,T_{k+1}]$.

A4: The row of transition matrix Π_k is a uniform distribution, such that $\pi_{k,i,j} = N^{\frac{i_T}{T}}$. The observation dimension D, the number of change point K and initialized change point number $M_{K+1} - 1$ is fixed.

We further define the random functions of $a(t)$, $b(t)$ and $c(t)$ for time interval $[m, n]$ as following:

$$\int_m^n a(t)dt \quad = \quad \begin{cases} \int_0^{m-n} -b(t)dt & \text{if } n \leq m, \\ \\ \int_0^{n-m} c(t)dt & \text{if } n > m. \end{cases}$$

with

$$\begin{aligned} b(t) &= \max_k \Big[\ln |S_k| /2 - (y(t) - u_k)^\top S_k \left(y(t) - u_k\right)/2 \mid t \in [T_{k-1}, T_k]\Big], \\ c(t) &= \max_k \Big[\ln |S_k| /2 - (y(t) - u_k)^\top S_k \left(y(t) - u_k\right)/2 \mid t \notin [T_{k-1}, T_k]\Big]. \end{aligned}$$

Intuitively, the defined $b(t)$ is the maximum likelihood value at time t, where the likelihood function is parameterized with true u and S, while $c(t)$ is the maximum likelihood value associated with false parameters u and S. Thus, the integral range $[m, n]$ of $b(t)$ and $c(t)$ indicates the sequence length that is correctly aligned or not, respectively.

Corollary 1. *As N approaches infinity, for any time interval $[m, n]$,the random variables, we have:*

$$\int_m^n \left(c(t) - b(t)\right) dt = \mathcal{O}_p(N^{\frac{n-m}{T}}) < 0.$$

Proof: *First using the Lemma from (Bishop et al., 2007):*

Lemma 1. *Let $\{X_n\}$ be a stochastic sequence with $\mu_n = \mathbb{E}\left(X_n\right)$ and $\sigma_n^2 = \operatorname{Var}\left(X_n\right) < \infty$, then $X_n = \mu_n + O_p\left(\sigma_n\right)$.*

Thus, based on the ***Lemma*** *1 and our assumptions, it's easy to see the following results hold:*

1)*: The value of $\int_m^n b(t)dt$ equals to:*

$$\begin{aligned} &N^{\frac{n-m}{T}} \max_k \left(\frac{1}{2}\ln |S_k| - \frac{1}{2}\operatorname{Tr}\left(\left[\mathbb{E}\left[y(t)y(t)^\top\right] - \mathbb{E}[y(t)]u_k^\top - u_k\mathbb{E}[y(t)^T] + u_k u_k^T\right] \cdot S_k\right)\right) \\ &= N^{\frac{n-m}{T}} \max_k \left(\frac{1}{2}\ln |S_k|\right) + \mathcal{O}_p(N^{\frac{n-m}{2T}}). \end{aligned}$$

2)*: The value of $\int_m^n c(t)dt$ eqauls to:*

$$N^{\frac{n-m}{T}} \max_k \max_{k' \neq k} \left(\frac{1}{2}\ln |S_k| - \frac{1}{2}\left[\left(\mu_k - \mu_{k'}\right)^\top S_k \left(\mu_k - \mu_{k'}\right) + \operatorname{Tr}(S_{k'}^{-1} \cdot S_k)\right]\right) + \mathcal{O}_p(N^{\frac{n-m}{2T}}).$$

Then the value of $\int_m^n \left(c(t) - b(t)\right) dt$ is given by:

$$\begin{aligned} &\int_m^n \left(c(t) - b(t)\right) dt \\ &= N^{\frac{n-m}{T}} \max_k \max_{k' \neq k} -\frac{1}{2}\left[\left(\mu_k - \mu_{k'}\right)^\top S_k \left(\mu_k - \mu_{k'}\right) + \operatorname{Tr}(S_{k'}^{-1} \cdot S_k)\right] + \mathcal{O}_p(N^{\frac{n-m}{2T}}) < 0. \end{aligned}$$

□

Based on the above assumptions, we can present our results in the following:

Theorem 2. *For $t_i \in \{t_{M_{k-1}+1}, ..., t_{M_k}\}$, the value of each forward message is given by:*

$$\mu_{\to \mathbf{t}_i}(n) = \frac{N^{\frac{(i-1)n}{T}}}{N^i (\ln N)^{i-1}} \exp\left(\int_0^{T_k} b(t)dt\right) \exp\left(\int_0^{n-T_k} \alpha(t)dt\right).$$

Proof: We will use the method of induction to drive the general formula of the forward message. By considering each data $y(t)$ as the continuous function of time t, the initial message is given by:

$$\begin{aligned}
\mu_{\rightarrow \mathbf{t}_1}(m) &= \pi_{1,m} \cdot \exp\left\{ \int_0^m \left\{ \frac{1}{2} \left\langle \ln \left| \Lambda_1^0 \right| \right\rangle \right.\right. \\
&\quad \left.\left. - \frac{1}{2} \operatorname{Tr}\left[\left(y(t) \cdot y(t)^\top - y(t) \cdot \left\langle u_1^0 \right\rangle^\top - \left\langle u_1^0 \right\rangle \cdot y(t)^\top + \left\langle u_1^0, u_1^{0\top} \right\rangle \right) \cdot \left\langle \Lambda_1^0 \right\rangle \right] \right\} \right\} \\
&= \begin{cases} \mathcal{O}_p(\pi_{1,m} \cdot \exp\left\{ \int_0^m b(t)dt \right\}) & \text{if } m \leq T_1, \\ \\ \mathcal{O}_p(\pi_{1,m} \cdot \exp\left\{ \int_0^{T_1} b(t)dt + \int_{T_1}^m c(t)dt) \right\} & \text{if } m \geq T_1. \end{cases} \\
&= \mathcal{O}_p\left(\frac{1}{N} \exp\left(\int_0^{T_1} b(t)dt \right) \cdot \exp\left(\int_{T_1}^m \alpha(t)dt \right) \right),
\end{aligned}$$

where we use the fact that $\pi_{i,m} = 1/N$ and the initial segment is a subset of the first regime $[0, t_1] \subset [0, T_1]$ and the initialized parameters are consistent with the true value S_1 and u_1, such that:

$$\left\langle u_1^0 \right\rangle = \hat{u}_1 \xrightarrow{\text{p}} u_1, \quad \left\langle u_1^0, u_1^{0\top} \right\rangle = \hat{u}_1 \cdot \hat{u}_1^\top, \quad \left\langle \Lambda_1^0 \right\rangle = \hat{S}_1 \xrightarrow{\text{p}} S_1, \quad \left\langle \ln \left| \Lambda_1^0 \right| \right\rangle = \ln \left| \hat{S}_1 \right| \xrightarrow{\text{p}} \ln \left| S_1 \right|.$$

Now consider the next message using the updated formula:

$$\begin{aligned}
\mu_{\rightarrow \mathbf{t}_2}(n) = &\int_0^n \left\{ \mu_{f_0 \rightarrow \mathbf{t}_1}(m) \cdot \pi_{2,m,n} \cdot \exp\left\{ \int_m^n \frac{1}{2} \left\langle \ln \left| \Lambda_2^0 \right| \right\rangle \right.\right. \\
&\left.\left. - \frac{1}{2} \operatorname{Tr}\left[\left(y(t) \cdot y(t)^\top - y(t) \cdot \left\langle u_2^0 \right\rangle^\top - \left\langle u_2^0 \right\rangle y(t)^\top + \left\langle u_2^0, u_2^{0\top} \right\rangle \right) \cdot \left\langle \Lambda_2^0 \right\rangle \right] dt \right\} \right\} dm \\
= &\begin{cases} \mathcal{O}_p\left(\frac{1}{N} (\int_0^n N^{\frac{m-T}{T}} dm) \cdot \exp\left\{ \int_0^n b(t)dt \right\} \right) & \text{if } n \leq T_1, \\ \\ \mathcal{O}_p\left(\frac{1}{N} (\int_0^n N^{\frac{m-T}{T}} dm) \cdot \exp\left\{ \int_0^{T_1} b(t)dt + \int_{T_1}^n c(t)dt) \right\} \right) & \text{if } n \geq T_1. \end{cases} \\
= &\mathcal{O}_p\left(\frac{N^{\frac{n}{T}}}{N^2 \cdot \ln N} \exp\left(\int_0^{T_1} b(t)dt \right) \cdot \exp\left(\int_{T_1}^n \alpha(t)dt \right) \right).
\end{aligned}$$

Therefore, the exponential term is exactly the same as the initial message. It's easy to see as long as $i \in \{1, ..., M_1 - 1\}$, the message is given by:

$$\mu_{\rightarrow \mathbf{t}_i}(n) = \frac{N^{\frac{(i-1)n}{T}}}{N^i (\ln N)^{i-1}} \exp\left(\int_0^{T_k} b(t)dt \right) \exp\left(\int_0^{n-T_k} \alpha(t)dt \right).$$

Now consider the first junction point $i = M_1$, the message is given by:

$$\begin{aligned}
\mu_{\rightarrow \mathbf{t}_{M_1}}(n) &= \mathcal{O}_p \Bigg(\int_0^n \Bigg\{ \frac{N^{\frac{(M_1-1)m}{T}}}{N^{M_1} \cdot (\ln N)^{(M_1-2)}} \exp\Big(\int_0^{T_1} b(t)dt \Big) \\
&\qquad \times \exp\Big(\int_{T_1}^m \alpha(t)dt \Big) \cdot \exp\Big(\int_0^{n-m} c(t)dt \Big) \Bigg\} dm \Bigg).
\end{aligned}$$

We can discuss in part:

1. for $n \geq m \geq T_1$:

$$\mu_{\rightarrow \mathbf{t}_{M_1}}(n) = \mathcal{O}_p\left(\frac{N^{\frac{(M_1-1)n}{T}}}{N^{M_1} (\ln N)^{(M_1-1)}} \exp\Big(\int_0^{T_1} b(t)dt \Big) \exp\Big(\int_0^{n-T_1} c(t)dt \Big) \right).$$

2. for $n \geq T_1 \geq m$:

$$\begin{aligned}\mu_{\rightarrow \mathbf{t}_{M_1}}(n) &= \mathcal{O}_p\Bigg(\frac{N^{\frac{(M_1-1)n}{T}}}{N^{M_1}(\ln N)^{(M_1-1)}} \int_0^n \exp\Big(\int_0^{T_1} b(t)dt\Big) \\ &\quad \times \exp\Big(\int_0^{T_1-m} (c(t)-b(t))dt\Big) \exp\Big(\int_0^{n-T_1} c(t)dt\Big) dm \Bigg).\end{aligned}$$

3. for $T_1 \geq n \geq m$:

$$\begin{aligned}\mu_{\rightarrow \mathbf{t}_{M_1}}(n) &= \mathcal{O}_p\Bigg(\frac{N^{\frac{(M_1-1)n}{T}}}{N^{M_1}(\ln N)^{(M_1-1)}} \int_0^n \exp\Big(\int_0^{T_1} b(t)dt\Big) \\ &\quad \times \exp\Big(\int_0^{n-m} (c(t)-b(t))dt\Big) \exp\Big(\int_0^{T_1-n} -b(t)dt\Big) dm \Bigg).\end{aligned}$$

Thus,

$$\mu_{\rightarrow \mathbf{t}_{M_1}}(n) = \mathcal{O}_p\left(\frac{N^{\frac{(M_1-1)n}{T}}}{N^{M_1}(\ln N)^{(M_1-1)}} \exp\Big(\int_0^{T_1} b(t)dt\Big) \exp\Big(\int_{T_1}^{n} \alpha(t)dt\Big) \right).$$

Then we evaluate the first non-junction point $i = M_1 + 1$. By discussing it by part, we show that:

1. if $n \geq T_1$:

$$\begin{aligned}\mu_{\rightarrow \mathbf{t}_{M_1+1}}(n) &= \mathcal{O}_p\Bigg(\exp\Big(\int_0^{T_2} b(t)dt\Big) \exp\Big(\int_{T_2}^{n} \alpha(t)dt\Big) \\ &\quad \times \int_0^n \left\{ \frac{N^{\frac{M_1 m}{T}}}{N^{M_1+1}(\ln N)^{(M_1-1)}} \exp\Big(\int^{T_1-m} (c(t)-b(t))dt\Big) \right\} dm \Bigg).\end{aligned}$$

2. if $n \leq T_1$:

$$\begin{aligned}\mu_{\rightarrow \mathbf{t}_{M_1+1}}(n) &= \mathcal{O}_p\Bigg(\exp\Big(\int_0^{T_2} b(t)dt\Big) \exp\Big(\int_{T_2}^{n} \alpha(t)dt\Big) \\ &\quad \times \int_0^n \left\{ \frac{N^{\frac{M_1 m}{T}}}{N^{M_1+1}(\ln N)^{(M_1-1)}} \exp\Big(\int^{n-m} (c(t)-b(t))dt\Big) \right\} dm \Bigg).\end{aligned}$$

In both cases, we can rewrite the message as:

$$\mu_{\rightarrow \mathbf{t}_{M_1+1}}(n) = \mathcal{O}_p\left(\frac{N^{\frac{M_1 m}{T}}}{N^{M_1+1} \cdot (\ln N)^{M_1}} \exp\Big(\int_0^{T_2} b(t)dt\Big) \exp\Big(\int_{T_2}^{n} \alpha(t)dt\Big) \right),$$

which returns to the initial message $\mu_{\rightarrow \mathbf{t}_1}(n)$ with the same exponential terms. □

By the same induction procedure, it's easy to see **Theorem** 2 holds for all $i \in \{t_{M_{k-1}+1}, ..., t_{M_k}\}$.

Theorem 3. *For $t_i \in \{t_{M_{k-1}}, ..., t_{M_k-1}\}$, the value of each backward message is given by:*

$$\mu_{\mathbf{t}_i \leftarrow}(m) = \exp\left(\int_0^{T-T_{k-1}} b(t)dt\right) \exp\left(\int_0^{T_{k-1}-m} \alpha(t)dt\right).$$

Proof: The proof of **Theorem** 3 is similar to that of **Theorem** 2. The initial backward message is given by:

$$
\begin{aligned}
\mu_{\mathbf{t}_{M_{K+1}-1}\leftarrow}(m) &= \exp\Bigg\{\int_m^T \Big\{\frac{1}{2}\left\langle \ln\left|\Lambda_{K+1}^0\right|\right\rangle - \frac{1}{2}\operatorname{Tr}\Big[\Big(y(t)y(t)^\top - y(t)\cdot\left\langle u_{K+1}^0\right\rangle^\top \\
&\qquad - \left\langle u_{K+1}^0\right\rangle^\top\cdot y(t) + \left\langle u_{K+1}^0, u_{K+1}^{0\top}\right\rangle\Big)\cdot\left\langle\Lambda_{K+1}^0\right\rangle\Big]dt\Big\}\Bigg\} \\
&= \begin{cases} \mathcal{O}_p(\exp\left\{\int_m^T b(t)dt\right\}) & \text{if } m \geq T_K, \\ \mathcal{O}_p(\exp\left\{\int_{T_K}^T b(t)dt + \int_m^{T_K} c(t)dt)\right\} & \text{if } m \leq T_K. \end{cases} \\
&= \mathcal{O}_p\left(\exp\Big(\int_0^{T-T_K} b(t)dt\Big)\exp\Big(\int_0^{T_K-m}\alpha(t)dt\Big)\right),
\end{aligned}
$$

where the initialized parameters are consistent estimators of true u_{K+1} and S_{K+1}.Now consider the next backward message using the updated formula:

$$
\begin{aligned}
\mu_{\to\mathbf{t}_{M_{K+1}-2}}(m) &= \int_m^T \Bigg\{\mu_{\to\mathbf{t}_{M_{K+1}-1}}(n)\cdot\pi_{M_{K+1}-1,m,n} \\
&\quad \times\exp\Bigg\{\int_m^n\Big\{\frac{1}{2}\left\langle\ln\left|\Lambda_{K+1}^0\right|\right\rangle - \frac{1}{2}\operatorname{Tr}\big[\big(y(t)\cdot y(t)^\top - 2y(t)\cdot\left\langle u_{K+1}^0\right\rangle \\
&\quad + \left\langle u_{K+1}^0, u_{K+1}^{0\top}\right\rangle\big)\cdot\left\langle\Lambda_{K+1}^0\right\rangle\big]dt\Big\}\Bigg\}\Bigg\}dn \\
&= \begin{cases} \mathcal{O}_p\left(N^{\frac{m-T}{T}}\cdot\exp\left\{\int_0^{T-m} b(t)dt\right\}\right) & \text{if } m \geq T_K, \\ \mathcal{O}_p\left(N^{\frac{m-T}{T}}\cdot\exp\left\{\int_0^{T-T_K} b(t)dt + \int_0^{T_K-m} c(t)dt)\right\}\right) & \text{if } m \leq T_K. \end{cases} \\
&= \mathcal{O}_p\left(\exp\Big(\int_0^{T-T_K} b(t)dt\Big)\exp\Big(\int_0^{T_K-m}\alpha(t)dt\Big)\right).
\end{aligned}
$$

Thus, it's easy to show for all $i \in \{M_K+1, ..., M_{K+1}-2\}$,

$$
\mu_{\mathbf{t}_i\leftarrow}(m) = \mathcal{O}_p\left(\exp\Big(\int_0^{T-T_K} b(t)dt\Big)\exp\Big(\int_0^{T_K-m}\alpha(t)dt\Big)\right).
$$

Then we consider the first junction point $i = M_K$:

$$
\begin{aligned}
\mu_{\mathbf{t}_{M_K}\leftarrow}(m) &= N^{\frac{m-T}{T}}\int_m^T\Bigg\{\exp\Big(\int_0^{T-T_K} b(t)dt\Big) \\
&\qquad \times\exp\Big(\int_0^{T_K-n}\alpha(t)dt\Big)\exp\Big(\int_0^{n-m} c(t)dt\Big)\Bigg\}dn.
\end{aligned}
$$

Consider three cases:

1. If $n \geq m \geq T_k$:

$$
\begin{aligned}
\mu_{\mathbf{t}_{M_K}\leftarrow}(m) &= N^{\frac{m-T}{T}}\int_m^T\Bigg\{\exp\Big(\int_0^{T-T_K} b(t)dt\Big) \\
&\qquad \times\exp\Big(\int_0^{m-T_K} -b(t)dt\Big)\exp\Big(\int_0^{n-m}(c(t)-b(t))dt\Big)\Bigg\}dn.
\end{aligned}
$$

2. If $n \geq T_K \geq m$:

$$\begin{aligned}\mu_{\mathbf{t}_{M_K}\leftarrow}(m) \quad &= \quad N^{\frac{m-T}{T}} \int_m^T \Bigg\{ \exp\left(\int_0^{T-T_K} b(t)dt\right) \\ &\times \exp\left(\int_0^{T_K-m} c(t)dt\right) \exp\Big(\int_0^{n-T_K} \big(c(t)-b(t)\big)dt\Big)\Bigg\}dn.\end{aligned}$$

3. If $T_K \geq n \geq m$:

$$\mu_{\mathbf{t}_{M_K}\leftarrow}(m) = N^{\frac{m-T}{T}} \int_m^T \left\{ \exp\left(\int_0^{T-T_K} b(t)dt\right) \cdot \exp\left(\int_0^{T_K-m} c(t)dt\right)\right\}dn.$$

Thus we can sum it up as:

$$\mu_{\mathbf{t}_i\leftarrow}(m) = \mathcal{O}_p\left(\exp\Big(\int_0^{T-T_K} b(t)dt\Big)\exp\Big(\int_0^{T_K-m} \alpha(t)dt\Big)\right).$$

When it comes to the new point in the previous segment $i = M_K - 1$:

1. If $n \leq T_K$:

$$\begin{aligned}\mu_{\mathbf{t}_{M_{K-1}}\leftarrow}(m) \quad &= \quad N^{\frac{m-T}{T}} \int_m^T \Bigg\{ \exp\Big(\int_0^{T-T_{K-1}} b(t)dt\Big) \\ &\times \exp\Big(\int_0^{T_K-n} c(t)-b(t)dt\Big) \exp\Big(\int_0^{T_{K-1}-m} \alpha(t)dt\Big)\Bigg\}dn.\end{aligned}$$

2. If $n \geq T_K$:

$$\begin{aligned}\mu_{\mathbf{t}_{M_{K-1}}\leftarrow}(m) \quad &= \quad N^{\frac{m-T}{T}} \exp\Big(\int_0^{T-T_{K-1}} b(t)dt\Big) \\ &\times \int_m^T \Bigg\{ \exp\Big(\int_0^{\min\{n-T_K, n-m\}} \big(c(t)-b(t)\big)dt\Big) \\ &\times \exp\Big(\int_0^{T_{K-1}-m} \alpha(t)dt\Big)\Bigg\}dn.\end{aligned}$$

Thus, following the same procedure in the proof of **Theorem** 2, we can derive that for all i, the recursive formula holds.

We are now ready to prove the location consistency. First consider the change point $t_i \in \{t_{M_{k-1}+1}, ..., t_{M_k-1}\}$. The unnormalized marginal probability $\tilde{Q}(t_i = m)$ is given by:

$$\frac{N^{\frac{in}{T}}}{N^i(\ln N)^{i-1}} \exp\left(\int_0^{T+T_k-T_{k-1}} b(t)dt\right) \exp\left(\int_0^{m-T_k} \alpha(t)dt\right) \exp\left(\int_0^{T_{k-1}-m} \alpha(t)dt\right).$$

Thus we can discuss all possible values of location m:

1. When $T_{k-1} \leq m \leq T_k$: It's easy to show

$$\int_0^{m-T_k} \alpha(t)dt + \int_0^{T_{k-1}-m} \alpha(t)dt = -\int_0^{T_k-m} b(t)dt - \int_0^{m-T_{k-1}} b(t)dt.$$

Thus,

$$\begin{aligned}\tilde{Q}(t_i = m) &= \frac{N^{\frac{im}{T}}}{N^i(\ln N)^{i-1}} \exp\left(\int_0^{T+T_k-T_{k-1}} b(t)dt\right) \exp\left(-\int_0^{T_k-T_{k-1}} b(t)dt\right) \\ &= \frac{N^{\frac{in}{T}}}{N^i(\ln N)^{i-1}} \exp\left(\int_0^T b(t)dt\right).\end{aligned}$$

2. When $m \geq T_k$: It's easy to show

$$\int_0^{m-T_k} \alpha(t)dt + \int_0^{T_{k-1}-m} \alpha(t)dt = \int_0^{m-T_k} \left(c(t) - b(t)\right) dt - \int_0^{T_k - T_{k-1}} b(t)dt.$$

Thus

$$\tilde{Q}(t_i = m) = \frac{N^{\frac{im}{T}}}{N^i (\ln N)^{i-1}} \exp\left(\int_0^T b(t)dt\right) \exp\left(\int_0^{m-T_k} \big(c(t) - b(t)\big)dt\right).$$

3. When $m \leq T_{k-1}$: It's easy to show

$$\int_0^{m-T_k} \alpha(t)dt + \int_0^{T_{k-1}-m} \alpha(t)dt = \int_0^{T_{k-1}-m} \big(c(t) - b(t)\big)dt - \int_0^{T_k - T_{k-1}} b(t)dt,$$

Thus

$$\tilde{Q}(t_i = m) = \frac{N^{\frac{im}{T}}}{N^i (\ln N)^{i-1}} \exp\left(\int_0^T b(t)dt\right) \exp\left(\int_0^{T_{k-1}-m} \big(c(t) - b(t)\big)dt\right).$$

The value of $Q(t_i = m)$ requires normalization. The normalization constant is given by:

$$\begin{aligned} C &= \int_0^N \tilde{Q}(t_i = m)dm \\ &= \mathcal{O}_p\left(\frac{N^{\frac{iT_k}{T}}}{N^i (\ln N)^{i-1}} \exp\Big(\int_0^T b(t)dt\Big)\right). \end{aligned}$$

Thus, the value of $Q(t_i = m) = \tilde{Q}(t_i = m)/C$ is given by:

$$\begin{aligned} Q(t_i = m) &= \begin{cases} N^{\frac{im - iT_k}{T}} & \text{if } m \in [T_{k-1}, T_k), \\ N^{\frac{im - iT_k}{T}} \exp\left(\int_0^{m-T_k} \left(c(t) - b(t)\right) dt\right) & \text{if } m \geq T_k, \\ N^{\frac{im - iT_k}{T}} \exp\left(\int_0^{T_{k-1}-m} \left(c(t) - b(t)\right) dt\right) & \text{if } m \leq T_{k-1}. \end{cases} \\ &= \begin{cases} 1 & \text{if } m = T_k, \\ \mathcal{O}_p(N^{\frac{m-T_k}{T}}) & \text{if } m \in [T_{k-1}, T_k), \\ \mathcal{O}_p(\exp(N^{\frac{\min\{|m-T_k|, |m-T_{k-1}|\}}{T}})) & \text{if } m \notin [T_{k-1}, T_k]. \end{cases} \end{aligned}$$

For junction points T_i with $i \in \{M_k\}_{k=1}^K$, the unormalized probability is given by:

$$\tilde{Q}(t_i = m) = \frac{N^{\frac{in}{T}}}{N^i (\ln N)^{i-1}} \exp\left(\int_0^T b(t)dt\right) \exp\left(\int_0^{|T_{k-1}-m|} \big(c(t) - b(t)\big)dt\right).$$

Then, the normalization constant is

$$\begin{aligned} C &= \int_0^N \tilde{Q}(t_i = m)dm \\ &= \mathcal{O}_p\left(\frac{N^{\frac{iT_k}{T}}}{N^i (\ln N)^{i-1}} \exp\Big(\int_0^T b(t)dt\Big)\right). \end{aligned}$$

Since $Q(t_i = m) = \tilde{Q}(t_i = m)/C$, we have

$$\begin{aligned} Q(t_i = m) &= N^{im - iT_k} \exp\left(\int_0^{|m-T_k|} \big(c(t) - b(t)\big)dt\right) \\ &= \mathcal{O}_p\left(\exp\Big(-N^{\frac{|m-T_k|}{T}}\Big)\right). \end{aligned}$$

Hence the proof is finished.

□

D Simulation settings

D.1 initialization and Hyperparameters Setting

The hierarchical model given in **Equation** 1 has hyperparameters $\alpha = \{\beta, \nu^0, V^0\}$. To implement **Algorithm** 1, the hyperparameters in the conjugate prior defined in **Equation** 1 set as follows: the parameter β in all Gaussian prior distributions $\mathcal{N}\left(0, \beta^{-1} I\right)$ are set to the data dimension D. Similarly the prior Wishart distribution $\mathcal{W}(\nu^0, V^0)$ is assigned with $\nu^0 = D, V^0 = D \cdot \mathbf{I}$ where $\mathbf{I}$ is identity matrix of dimension D. The Gaussian-Wishart prior has been studied for low-rank matrix completion, which imposes an appropriate penalty, and encourages sparse solutions with promising convergence.

Throughout all the experiments, the initialization of **Algorithm** 1 follows the description of **A3** in **Section** 2.3 where we evenly divide the entire sequence into $\tilde{K}$ segments. Then $\{Q(\theta)\}_{k=1}^{\tilde{K}}$ are initialized using the statistical moments (mean and variance for the Gaussian distribution) from these segments.

D.2 Numerical demonstration

In this section, We evaluate the performance by varying values of sequence length N. In particular, we consider sampling the 1-dimensional mean-variance shift sequence with five equally spaced segments. The length of each segment is $N/5$ and N varies in the set $\{50, 100, 200, 400, 600, 800, 1000, 1200, 1400, 1600\}$. In each segment, samples are drawn from a normal distribution with the following parameters.

$$\boldsymbol{u} = [0, 3, 2, 4, 4] \qquad \Lambda = [1, 0.25, 1, 1, 4]$$

Elements in $\boldsymbol{u}$ and Λ represent the mean and precision of a particular segment. For example, samples in the first segment follow a $\mathcal{N}(0, 1)$ distribution. We initialize our algorithm with $\tilde{K} = 10$ and set the iteration number to 30. The simulations are repeated 100 times and the average number of change points and average mean absolute error are reported in Figure 2.

D.3 Location and parameter estimation

In this subsection, we consider one mixed distribution sequence and two normal sequences. Model 1 is a variance shift sequence model. The five ordered segments are sample from $Binomial(10, 0.3)$, $\mathcal{N}(3, 4)$, $Poisson(3)$, and $Binomial(15, 0.2)$, each with 100 samples. For Model 2 and Model 3, the multivariate normal sequences, parameters are specified in the following Table 3. Let

$$\boldsymbol{u}_1^{(2)} = [0, 0, 0, 0, 0], \quad \boldsymbol{u}_2^{(2)} = \boldsymbol{u}_3^{(2)} = [0, 2, 0, 1, 2], \quad \boldsymbol{u}_4^{(2)} = \boldsymbol{u}_5^{(2)} = [4, 0, 2, 0, 4],$$

$$\boldsymbol{u}_1^{(3)} = [\mathbf{0}_{10}], \quad \boldsymbol{u}_2^{(3)} = \boldsymbol{u}_3^{(3)} = [0, 2, 0, 1, 0, 1, 0, 0, 0, 1], \quad \boldsymbol{u}_4^{(3)} = \boldsymbol{u}_5^{(3)} = [1, 0, 2, 0, 4, 0, 0, 4, 0, 1].$$

$\mathbf{I}$ is the identity matrix of size D and $\mathbf{I}_{0.8}$ is an identity matrix with the non-diagonal elements equal to 0.8. $\mathbf{0}_{10}$ is a zero vector in 10-d. Specifically, all three Models are subject to four change points occurring at $\tau = \{100, 200, 300, 400\}$, each representing a change in distribution. Clearly, Model 2 and 3 incorporate the mean shift or correlation shift around change points. For all three models, N, the length of the sequence is 500, and the upper bound of the number of change points $\tilde{K}$ is 10. The iteration number is set to 30. For each model, the repetition of simulation is 100 and the average Rand index of each model is reported in Table 1.

Table 3: Normal Mean correlation-Shift of Model 2 and 3

	$\boldsymbol{u}$	Λ	τ	D	N
Model 2	$[\boldsymbol{u}_1^{(2)}, \boldsymbol{u}_2^{(2)}, \boldsymbol{u}_3^{(2)}, \boldsymbol{u}_4^{(2)}, \boldsymbol{u}_5^{(2)}]$	$\mathbf{I}_{0.8}^{-1}, \mathbf{I}_{0.8}^{-1}, \mathbf{I}, \mathbf{I}, \mathbf{I}_{0.8}^{-1}$	$100, 200, 300, 400$	5	500
Model 3	$[\boldsymbol{u}_1^{(3)}, \boldsymbol{u}_2^{(3)}, \boldsymbol{u}_3^{(3)}, \boldsymbol{u}_4^{(3)}, \boldsymbol{u}_5^{(3)}]$	$\mathbf{I}_{0.8}^{-1}, \mathbf{I}_{0.8}^{-1}, \mathbf{I}, \mathbf{I}, \mathbf{I}_{0.8}^{-1}$	$100, 200, 300, 400$	10	500

In this part, we evaluate the accuracy of the posterior parameter estimation. Here we only consider normal sequence cases for estimation. The parameters are summarized in Table 4

Table 4: Normal Mean correlation-Shift in Case 1, 2 and 3

	$\boldsymbol{u}$	Λ	τ	D	N
Case 1	$[0, 3, 2, 4, 4]$	$1, 0.25, 1, 1, 4$	$100, 200, 300, 400$	1	500
Case 2	$[\boldsymbol{u}_1^{(2)}, \boldsymbol{u}_2^{(2)}, \boldsymbol{u}_3^{(2)}, \boldsymbol{u}_4^{(2)}, \boldsymbol{u}_5^{(2)}]$	$\mathbf{I}_{0.8}^{-1}, \mathbf{I}_{0.8}^{-1}, \mathbf{I}, \mathbf{I}, \mathbf{I}_{0.8}^{-1}$	$100, 200, 300, 400$	5	500
Case 3	$[\boldsymbol{u}_1^{(3)}, \boldsymbol{u}_2^{(3)}, \boldsymbol{u}_3^{(3)}, \boldsymbol{u}_4^{(3)}, \boldsymbol{u}_5^{(3)}]$	$\mathbf{I}_{0.8}^{-1}, \mathbf{I}_{0.8}^{-1}, \mathbf{I}, \mathbf{I}, \mathbf{I}_{0.8}^{-1}$	$100, 200, 300, 400$	10	500

The included symbols are the same as above. The estimation error (Mean square error) is measured by taking l^2 norm of the difference between the estimated mean and ground truth, while the MSE.SD is the ordinary standard deviation of estimation. Notice that in the estimation of the covariance matrix, the estimation error is further divided by data dimension D to maintain numerical consistency. These simulations are also repeated 100 times and the average MSE is shown in Table 2:

$$\text{Mean square error } = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{\mathbf{0}}\right\|_2^2, \tag{5}$$

and

$$\text{MSE.SD } = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\left\|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{\mathbf{0}}\right\|_2^2 - \frac{1}{N}\sum_{j=1}^{N}\left\|\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_{\mathbf{0}}\right\|_2^2\right)^2}. \tag{6}$$

D.4 NON-GAUSSIAN EXAMPLES SETTINGS

In these non-Gaussian examples, we consider testing the performance on Poisson, chi-squared, or the exponential random sequences. For the Poisson sequence, the rate parameters $\boldsymbol{\lambda}$ of five segments are $\boldsymbol{\lambda} = [1, 5, 2, 10, 3]$. $\mathbf{df} = [1, 5, 2, 4, 1]$ are set to be the parameters of chi-squared sequence. The scale parameters of the exponential distribution are $\boldsymbol{\beta} = [1, 5, 0.5, 4, 1]$. N, the length of the sequence is 500 and each segment contains 100 samples. $\tilde{K}$, the upper bound of the number of change points is 8.