Bayesian Change Point Detection for Functional Data

Xiuqi Li and Subhashis Ghosal
Department of Statistics, North Carolina State University

Abstract

We propose a Bayesian method to detect change points in a sequence of functional observations that are signal functions observed with noises. Since functions have unlimited features, it is natural to think that the sequence of signal functions driving the underlying functional observations change through an evolution process, that is, different features change over time but possibly at different times. A change-point is then viewed as the cumulative effect of changes in many features, so that the dissimilarities in the signal functions before and after the change-points become overwhelming. In our setting, features are characterized by the wavelet coefficients in their expansion. We consider a Bayesian approach by putting priors independently on the wavelet coefficients of the underlying functions, allowing a change in their values over time. Then we compute the posterior distribution of change point for each sequence of wavelet coefficients, and obtain a measure of overall similarity between two signal functions in this sequence, which leads to the notion of an overall change-point by minimizing the similarity across the change point relative to the similarities within each segment. We study the performance of the proposed method through a simulation study and apply to a dataset on climate change.

Keywords:
Change point, Functional data, Wavelet transform, Posterior consistency

1. Introduction

Change point detection is an important aspect of data analysis. In recent years, detecting change points in functional data has received considerable attention. Berkes et al. [8] developed a method by projecting the difference of mean functions on the principal components for the data. Zhang et al.
developed a self-normalization based test to identify the potential change points. Aston and Kirch [4] introduced an estimator of a change point in a model for an epidemic, where a change in the state of the epidemic occurs, and then the observations return to baseline at a later time. Sharipov et al. [12] developed a test for structural changes in functional data taking values in a Hilbert space and determined critical values by bootstrap. Aue et al. [6] proposed a method to uncover structural breaks in functional data without using dimension reduction techniques.

In this paper, we propose a Bayesian method to detect change points in functional data. We consider the setup where functional observations appear in a sequence, and can be thought of as unknown signal functions observed with a noise process independent for each observation in the sequence. Our approach differs from others in a fundamental way, in that we do not assume that the true signals suddenly changes completely at a change point and are identical in all before the change point and also identical in all after the change point. Instead, as a function is an infinite-dimensional object containing many different features, it is natural to think that the signals evolve steadily changing some features of the signal functions at any time point. These changes cumulate and at a stage, their overall effect becomes substantial. Representing functions in a wavelet series, the features in a function can be quantified through the wavelet coefficients of its expansion in the wavelet basis. While any orthogonal basis can be used to expand a function, wavelets are particularly effective in picking up local features. We consider two settings of functional observations — either observed over a regular grid, or observed continuously over time. We extract the features of functional data by the discrete wavelet transform (DWT) in the former case, and using inner products with wavelet functions in the latter case. The signal-plus-noise model for functional data then reduces to the problem of detecting a change in the value of the mean vector in a multivariate (or an infinite-dimensional) normal model. Since there are no functional dependencies between the coefficients, we can treat each sequence of feature separately independently, and hence put independent priors on these coefficients. The prior for each coefficient allows for a change in the value of a wavelet coefficient before and after some time point, but this change point need not be the same or related across different wavelet coefficients. Therefore, the problem of the change point for functional data is reduced to the problem of multiple (or infinitely many) one-dimensional change points. Each of the latter can be solved very efficiently by a Bayesian method based on conjugate priors, allowing explicit
posterior computation without needing to sample from the posterior distribution. Finally, the change points for individual wavelet coefficients need to be synthesized to a single value that represents the most prominent change point for the entire signal. To this end, we introduce a similarity matrix whose \((i,i')\)th entry measures the fraction of wavelet coefficients of \(f_i\) and \(f_{i'}\) are tied together when the coefficients are sampled from their posterior distributions. The posterior expected value of this similarity can be obtained in terms of the posterior distributions of the change points for each wavelet coefficients, which are analytically computed in our setting. We define a synthetic change point for the overall change as the value that minimizes the ratio of the mean similarities between groups and within groups, where one group is formed by time points before and the other by time points after the given value. This process can also be repeated more than once to detect multiple change points. The number of change points to be found may also be adaptively obtained from the data using some stopping criterion involving these similarity measures.

The paper is organized as follows. In the next section, we formally introduce the model and describe the prior. Posterior computational formulas are presented in Section 3. Under a small noise regime, a posterior consistency result on estimation and identification of the change point is presented in Section 4. A simulation study demonstrating the accuracy of the proposed method, and an application to a dataset on climate change over a 250 years period are respectively presented in Sections 5 and 6. Proofs of the results are given in Section 7.

2. Setup, model description and prior distribution

We consider two different settings of functional data — observed over a regular discrete grid, or observed continuously over time. In both cases, the domain of the signal functions is taken to be the unit interval \([0,1]\), without loss of generality. The functional observations are viewed as signal functions corrupted by independent error processes. We follow the formulation of Suarez and Ghosal [13], who used wavelet coefficients of the underlying signal functions to describe their structure. Using these, they quantified the extent of similarities among different functional observations and used the information to cluster these observations in a few groups. Suarez and Ghosal [13] imposed a prior on wavelet coefficients which explicitly encourages forming ties in the values of these wavelet coefficients of different signal functions.
We modify their approach to change point detection for functional data. Actually, this problem can be regarded as a special case of clustering with the constraint that for each characteristic, there are at most two clusters and these are linearly ordered. Hence, the technique of Suarez and Ghosal [13] can be modified by imposing the additional restriction in the prior that ties occur over consecutive values only.

2.1. Model description

We model that the functional observations \( Y_i(t) \) arise from true signals \( f_i(t) \), corrupted by some noise process \( \varepsilon_i(t) \), \( t \in [0, 1], i = 1, \ldots, n \), where \( n \) denotes the sample size: \( Y_i(t) = f_i(t) + \varepsilon_i(t) \). As in Antoniadis et al. [3] and Suarez and Ghosal [13], we shall apply wavelet transforms on these functions to extract the relevant features. Let \( \{ \phi_0, \psi_{jk} : k = 0, \ldots, 2^j - 1; j = 0, 1, \ldots \} \) stand for a wavelet basis, and let \( (a_0^{(i)}, b_{jk}^{(i)} : k = 0, \ldots, 2^j - 1, j = 0, 1, \ldots), (\alpha_0^{(i)}, \beta_{jk}^{(i)} : k = 0, \ldots, 2^j - 1, j = 0, 1, \ldots), (e_0^{(i)}, e_{jk}^{(i)} : k = 0, \ldots, 2^j - 1, j = 0, 1, \ldots) \), stand respectively for the wavelet coefficients of \( Y_i, f_i \) and \( \varepsilon_i \) respectively, \( i = 1, \ldots, n \).

In the discrete model, we observe functional data \( Y_i(T_j), i = 1, \ldots, n, T_1, \ldots, T_m \in [0, 1] \), given by \( Y_i(T_i) = f_i(T_j) + \varepsilon_{ij} \), where \( \varepsilon_{ij} \) is assumed to follow a normal distribution with mean 0 and variance \( \sigma^2 \), and is assumed to be independent across \( i \) and \( j \). Define \( Y_i = (Y_i(T_1), \ldots, Y_i(T_m))^\top, f_i = (f_i(T_1), \ldots, f_i(T_m))^\top, \) and \( \varepsilon_i = (\varepsilon_i(T_1), \ldots, \varepsilon_i(T_m))^\top. \) If further \( m = 2^J \), for some \( J \) and \( T_1, \ldots, T_J \) are equidistant, then the wavelet coefficients up to the level \( J \) in the expansion of the functions may be computed much more efficiently by the discrete wavelet transform (DWT), mapping a vector in \( \mathbb{R}^m \) to another vector in \( \mathbb{R}^m \) by multiplying by a suitable orthogonal matrix; see Antoniadis et al. [3]). The orthogonality of the transformation preserves the independence and the normality of the errors, leading to the model

\[
a_0^{(i)} = a_0^{(i)} + e_0^{(i)}, \quad b_{jk}^{(i)} = \beta_{jk}^{(i)} + e_{jk}^{(i)}, \quad e_{jk}^{(i)} \sim N(0, \sigma^2), \quad (1)
\]

\( i = 1, \ldots, n, k = 0, \ldots, 2^j - 1, j = 0, 1, \ldots, J. \) If \( m \) is not a power of 2, a practically useful fix is to obtain a function in the continuous domain by a standard smoothing technique, and then evaluate the function at a set of equidistant points of cardinality a suitable power of 2.

When the functional data are (essentially) observed continuously in time, we consider a Gaussian white noise model \( dY_i(t) = f_i(t)dt + \sigma dB_i(t) \), where \( B_i(\cdot) \) are independent Brownian motions on \( [0, 1] \). Let \( a_0^{(i)} = \int_0^1 \phi_0(t)dY_i(t) \),
\[ \alpha^{(i)}_0 = \int_0^1 \phi_0(t) f_i(t) dt, \quad b^{(i)}_{jk} = \int_0^1 \psi_{jk}(t) dY_i(t), \quad \beta^{(i)}_{jk} = \int_0^1 \psi_{jk}(t) f_i(t) dt, \quad e^{(i)}_0 = \sigma \int_0^1 \phi_0(t) dB_i(t), \quad e^{(i)}_{jk} = \sigma \int_0^1 \psi_{jk}(t) dY_i(t), \]

\[ i = 1, \ldots, n, \quad k = 0, \ldots, 2^j - 1, \quad j = 0, 1, \ldots. \]

It follows that \( e^{(i)}_0 \) and \( e^{(i)}_{jk} \) are independent \( N(0, \sigma^2) \). Hence applying the wavelet transform, we can again express the model by the ‘independent normal sequence model’ (1), with the only difference that \( j = 0, 1, \ldots \) is not restricted by \( J \). The independence across different \((j, k)\) pairs allows us treating both the discrete and the continuous models within a single framework. Different components in the model stand for different features of the functions captured by the wavelet coefficients.

To detect the change point of this sequence of functional data, we first find the change in each component, that is, we detect the change for each feature \( \beta_{jk} \), and we decide the overall change point from them. The procedure of combining the information across different wavelet coefficients is formally described in Section 3.

### 2.2. Prior Distributions

For a given \((j, k)\), the change point prior on \( \beta_{jk} \) is regulated by its ‘change point parameter’ \( \tau_{jk} \) by the relations

\[ \beta^{(1)}_{jk} = \cdots = \beta^{(\tau_{jk} - 1)}_{jk} \neq \beta^{(\tau_{jk})}_{jk} = \cdots = \beta^{(n)}_{jk}, \tag{2} \]

\( \tau_{jk} = 1, \ldots, n \). Note that if \( \tau_{jk} = 1 \), there is actually no change point as all values are equal, which is permitted by our model. In principle, multiple change points may be possible in a component, but for simplicity of formulas, we only allow at most one change point for a component. Typically, this will suffice for applications since multiple change points can still be obtained for the overall functional observations as a result of the combined effect of component changes because of the difference of \( \tau_{jk} \) values across different pairs \((j, k)\). The prior on the leading wavelet coefficient \( \alpha^{(i)}_0 \) will also follow the scheme of (2) using a change point variable \( \tau_0 \).

Once the change point variable is given a prior on \( \{1, \ldots, n\} \), we can put independent priors on the common value \( \beta^{(i)}_{jk} \) of the ‘pre-change block’ \((\beta^{(i)}_{jk} : i = 1, \ldots, \tau_{jk} - 1)\) and that \( \beta^{(i)}_{jk} \) of the ‘post-change block’ \((\beta^{(i)}_{jk} : i = \tau_{jk}, \ldots, n)\). It is tempting to consider independent normal priors on the common values, as this will lead to posterior conjugacy given \( \tau_{jk} \). However, being a wavelet coefficient, it is natural for \( \beta_{jk} \) to assume the special value 0 commonly, as 0 stands for the absence of a feature in a function. Thus
it is imperative to allow the value 0 in the prior for a wavelet coefficient by attaching an additional point mass at 0, that is, replacing the normal prior by a spike-and-slab prior that is a convex combination of a point mass \( \delta_0 \) at 0 (a degenerate normal distribution) and a normal density: independently

\[
\beta_{jk}^{< \text{ind}} \sim \pi_j N(0, c_j^2 \sigma^2) + (1 - \pi_j) \delta_0, \quad \beta_{jk}^{> \text{ind}} \sim \pi_j N(0, c_j^2 \sigma^2) + (1 - \pi_j) \delta_0, \quad (3)
\]

where \( \pi_j \) and \( c_j \) are hyperparameters. The choices of these hyperparameters can be motivated by the degree of smoothness in the underlying functions. Suppose that \( c_j^2 = \nu_1 2^{-\gamma_1 j} \) and \( \pi_j = \min(1, \nu_2 2^{-\gamma_2 j}) \), and \( \nu_1, \nu_2, \gamma_1, \gamma_2 > 0 \). If the mother wavelet function \( \psi \) has regularity \( r \), then Abramovich et al. [1] observed that \( f \in B_{s,p,q} \) almost surely, where \( B_{s,p,q} \) denotes Besov space of index \( (p, q) \) and smoothness \( s \) with \( \max(0, 1/p - 1/2) < s < r \), \( 1 \leq p, q \leq \infty \), whenever \( s + \frac{1}{2} - \frac{2\gamma_2}{p} - \frac{\gamma_1}{2} < 0 \), or \( s + \frac{1}{2} - \frac{2\gamma_2}{p} - \frac{\gamma_1}{2} = 0 \) and \( 0 \leq \gamma_2 < 1, 1 \leq p < \infty, q = \infty \).

It may be noted that with positive probability (which is also increasing to one as \( j \) gets larger), the common value of \( \beta_{jk}^{(i)} \) before the change point \( \tau_{jk} \) and the common value after the change point are both 0, which negates an actual change even with a value \( \tau_{jk} > 1 \). In such a situation, the posterior distribution of \( \tau_{jk} \) is totally prior-driven, as the likelihood does not contain any information about \( \tau_{jk} \). This type of indeterminacy of value of the change point does not bother us, as we view \( \tau_{jk} \) only as a technical device to put priors on two blocks of common values of the coefficients using a hierarchical scheme. The posterior distribution of \( \tau_{jk} \) is not of primary interest, although it is used to compute the probability of ties in \( \beta_{jk}^{(i)} \) and \( \beta_{jk}^{(i')} \) for a given pair of indices \( i \neq i' \) (see the next section).

A prior on \( (\alpha_{0}^{(i)} : i = 1, \ldots, n) \) is put using a similar mechanism through its change point \( \tau \), except that the common values of the blocks \( (\alpha_{0}^{(i)} : i = 1, \ldots, \tau - 1) \) and \( (\alpha_{0}^{(i)} : i = \tau, \ldots, n) \) are given independent vague priors (normal with high variance) instead of spike-and-slab. All the change points \( \tau \) and \( \tau_{jk} \) are independently given a prior \( P(\tau = i) = \rho_i, i = 1, \ldots, n \), for some distribution \( (\rho_i : i = 1, \ldots, n) \), where \( \rho_i > 0, \sum_{i=1}^{n} \rho_i = 1 \). In numerical calculations, we take \( (\rho_i : i = 1, \ldots, n) \) to be discrete uniform on \( \{1, \ldots, n\} \).

Finally, we put an inverse-gamma prior \( \sigma^2 \sim IG(\theta, \lambda) \) on the global variance parameter \( \sigma^2 \).
3. Posterior computation

As the change points for the wavelet coefficients can potentially be different, a synthetic change point as an indicator of overall change is obtained by considering the degree of similarities between the wavelet coefficients. Following Suarez and Ghosal [13] who considered the problem of clustering of functional data, we quantify the similarity between the $i$th and $i'$th observation in the discrete model with $J$ levels by the proportion of ties between the wavelet coefficients of the underlying signals, which leads to a similarity matrix $S$ with entries

$$S(i, i') = S_J(i, i') = \frac{1}{2^J} \left[ \mathbb{1}(\alpha_0^{(i)} = \alpha_0^{(i')}) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \mathbb{1}(\beta_{jk}^{(i)} = \beta_{jk}^{(i')}) \right]. \quad (4)$$

In the continuous model (which corresponds to infinitely many levels), we may modify the measure by a weighted sum with a geometric discount sequence $S(i, i') = \sum_{J=1}^{\infty} 2^{-J} S_J(i, i')$. Due to the effect of the discounting factor, the infinite sum is easily approximated by a finite sum. The synthetic change point of a sequence of functional data standing for the aggregate effect of changes in all possible wavelet coefficients of the underlying signals is the index-value where the contrast between overall similarities of observations before and that after is the largest. For any $k$ that divides the data into two groups, we compute the ratio $C(k)$ of the mean similarity between group and the mean similarity within group:

$$C(k) = \frac{\sum_{1 \leq i \leq k-1, k \leq i' \leq n} S(i, i')/[(k-1)(n-k+1)]}{\left\{ \sum_{1 \leq i \leq k-1} S(i, i') + \sum_{k \leq i \leq n} S(i, i') \right\}/\left[\binom{k-1}{2} + \binom{n-k+1}{2}\right].} \quad (5)$$

The change point is where this ratio is the minimum: $\tau = \arg \min \{C(k) : k = 2, \ldots, n\}$. The change point $\tau$ thus defined is a functional of $(f_1, \ldots, f_n)$. Bayesian inference can be made on $\tau$ by computing its posterior distribution, which may be obtained from posterior samples. Posterior samples can be drawn using the conditional conjugacy of $\beta_{jk}^{(i)}$ given $\tau_{jk}$.

A Bayesian estimator can be obtained more easily without posterior sampling. Define an estimator $\hat{\tau} = \arg \min \{\hat{C}(k) : k = 2, \ldots, n\}$, where $\hat{C}(k)$ is
Similarly, \( P(\tau) \) can be obtained from the posterior distribution of \( \tau \) means there are at least two data points in each group.

In the actual implementation, we restrict the search to \( 3 \leq k \leq n - 1 \), which means there are at least two data points in each group.

Let \( i < i' \) without loss of generality. Observe that \( P(\beta_{jk}^{(i)} = \beta_{jk}^{(i')} | b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}) \) can be obtained from the posterior distribution of \( \tau_{jk} \):

\[
P(\beta_{jk}^{(i)} = \beta_{jk}^{(i')} | b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}) = P(\tau_{jk} \leq i | b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}) + P(\tau_{jk} \geq i' + 1 | b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}) + P(\beta_{jk}^{(i)} = \beta_{jk}^{(i')} = 0, i < \tau_{jk} \leq i' | b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}).
\]

Similarly, \( P(\alpha_{0}^{(i)} = \alpha_{0}^{(i')} | b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}) \) can be computed.

The posterior probability of \( \{\tau_{jk} = m\} \), \( m = 1, \ldots, n \), is

\[
P(\tau_{jk} = m | b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}) = \frac{p(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)} | \tau_{jk} = m) \rho_{m}}{\sum_{l=1}^{n} p(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)} | \tau_{jk} = l) \rho_{l}},
\]

where \( p \) stands for the joint density after integrating out the mean and the variance parameters. When \( \tau_{jk} = 1 \), there is only one block. Considering the possibilities that the common value is 0 which occurs with probability \((1 - \pi_{j})\), and that it is non-zero, which occurs with probability \(\pi_{j}\), we obtain

\[
p(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)} | \tau_{jk} = 1) = (1 - \pi_{j}) \int \left\{ \prod_{l=1}^{n} \phi(b_{jk}^{(l)}; 0, \sigma^2) \right\} g(\sigma^2; \theta, \lambda) d\sigma
\]

\[
+ \pi_{j} \int \left\{ \prod_{l=1}^{n} \phi(b_{jk}^{(l)}; \xi, \sigma^2) \right\} \phi(\xi; 0, c_{j}^{2} \sigma^2) g(\sigma^2; \theta, \lambda) d\lambda d\sigma
\]

\[
= (1 - \pi_{j}) (2\pi)^{-n/2} \int \left[ \frac{\lambda^\theta}{\Gamma(\theta)} \frac{\Gamma(n/2 + \theta)}{\left[ \frac{1}{2} \sum_{l=1}^{n} (a_{0}^{(l)})^2 + \lambda \right]^{(n/2 + \theta)}} \right]
\]

\[
+ \pi_{j} (2\pi)^{-n/2} \int \left[ \frac{\lambda^\theta}{\Gamma(\theta)} \frac{(c_{j}^2 n + 1)^{-1/2} \Gamma(n/2 + \theta)}{\left[ \frac{1}{2} \sum_{l=1}^{n} (a_{0}^{(l)})^2 + \frac{c_{j}^2}{2(c_{j}^2 n + 1)} \sum_{l=1}^{n} (a_{0}^{(l)})^2 + \lambda \right]^{(n/2 + \theta)}} \right].
\]
For $\tau_{jk} = 2, \ldots, n$, there are two blocks with a pair of common values. Writing $\ast$ for a generic non-zero value, this can be $(0, 0)$, $(0, \ast)$, $(\ast, 0)$ and $(\ast, \ast)$, leading to

$$P(b_{jk}^{(1)} \ldots b_{jk}^{(n)} | \tau_{jk} = m)$$

$$= (1 - \pi_j)^2 \int \left\{ \prod_{l=1}^{n} \phi(b_{jk}^{(l)}; 0, \sigma^2) \right\} g(\sigma^2; \theta, \lambda) d\sigma$$

$$+ (1 - \pi_j) \pi_j \int \left\{ \prod_{l=1}^{i-1} \phi(b_{jk}^{(l)}; 0, \sigma^2) \right\} \left\{ \prod_{l=i}^{n} \phi(b_{jk}^{(l)}; \xi, \sigma^2) \right\} \times \phi(\xi; 0, c_j^2 \sigma^2) g(\sigma^2; \theta, \lambda) d\xi d\sigma$$

$$+ \pi_j (1 - \pi_j) \int \left\{ \prod_{l=1}^{i-1} \phi(b_{jk}^{(l)}; \xi, \sigma^2) \right\} \left\{ \prod_{l=i}^{n} \phi(b_{jk}^{(l)}; 0, \sigma^2) \right\} \times \phi(\xi_1; 0, c_j^2 \sigma^2) \phi(\xi_2; 0, c_j^2 \sigma^2) g(\sigma^2; \theta, \lambda) d\xi_1 d\xi_2 d\sigma$$

$$= (1 - \pi_j)^2 (2\pi)^{-n/2} \frac{\lambda^\theta}{\Gamma(\theta)} \frac{\Gamma(n/2 + \theta)}{\left[ 1/2 \sum_{l=1}^{n} (b_{jk}^{(l)})^2 + \lambda \right]^{(n/2 + \theta)}}$$

$$+ \pi_j (1 - \pi_j) (2\pi)^{-n/2} \frac{\lambda^\theta}{\Gamma(\theta)} \frac{\Gamma(n/2 + \theta)}{\left[ B_{i-1} + \lambda \right]^{(n/2 + \theta)}}$$

$$+ (1 - \pi_j) \pi_j (2\pi)^{-n/2} \frac{\lambda^\theta}{\Gamma(\theta)} \frac{\Gamma(n/2 + \theta)}{\left[ B_{n-i+1} + \lambda \right]^{(n/2 + \theta)}}$$

$$+ \pi_j^2 (2\pi)^{-n/2} \frac{\lambda^\theta}{\Gamma(\theta)} \frac{\Gamma(n/2 + \theta)}{\left[ B_{i-1} + B_{n-i+1} - 1/2 \sum_{l=1}^{n} (b_{jk}^{(l)})^2 + \lambda \right]^{(n/2 + \theta)}} \]^{(n/2 + \theta)},$$

(10)

where

$$B_{i-1} = \frac{1}{2} \sum_{l=1}^{n} (b_{jk}^{(l)})^2 - \frac{c_j^2}{2(c_j^2(i - 1) + 1)} \left( \sum_{l=1}^{i-1} b_{jk}^{(l)} \right)^2,$$

(11)
and
\[ \tilde{B}_{n-i+1} = \frac{1}{2} \sum_{l=1}^{n} (b_{jk}^{(l)})^2 - \frac{c_j^2}{2(c_j^2(n-i+1)+1)} \left( \sum_{l=i}^{n} b_{jk}^{(l)})^2 \right). \] (12)

Finally note that
\[ P(\beta_{jk}^\leq = \beta_{jk}^\geq = 0, \tau_{jk} = m| b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}) \]
\[ = \rho_m P(\beta_{jk}^\leq = \beta_{jk}^\geq = 0| \tau_{jk} = m) \frac{p(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}| \tau_{jk} = m)}{\sum_{l=1}^{n} \rho_l p(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}| \tau_{jk} = l)}. \]

The terms in the denominator in (8) are obtained from (9) and (10) and the numerator is given by \( \rho_m \) multiplied by the first term on the right side of (10). Putting this information together, the third term in (7) can be computed.

Similarly, we can compute the marginal likelihood \( P(a_0^{(1)}, \ldots, a_0^{(n)}| \tau_\ell = l) \) and the posterior probability of \( \tau_\ell = i \) through the Bayes theorem.

4. Posterior Consistency

In this section, we state a posterior consistency result for the collection of underlying signals and the synthetic change point, under the regime that the noise variance \( \sigma^2 \to 0 \), while the number of subjects \( n \) remains fixed. This setup is equivalent with the number of independent and identically distributed replications \( r \to \infty \) keeping the variance fixed, i.e., \( \sigma_r^2 = \sigma^2/r \). Let \( D_r \) be the set of all observations. This allows an application of the general theory of posterior contraction (Ghosal and van der Vaart [9]). We shall present the consistency result in terms of the wavelet coefficients in the model (1), covering both the discrete model and the continuous model, respectively for \( J < \infty \) and \( J = \infty \). To simplify notation, we assume that \( \alpha_0^{(i)} = 0 \) for \( i = 1, \ldots, n \).

For a vector of functions \( f = (f_1, \ldots, f_n) \), the \( L_2 \)-norm \( \|f\| \) is defined by
\[ \|f\|^2 = \sum_{i=1}^{n} \|f_i\|^2 = \sum_{i=1}^{n} \sum_{j=0}^{2^j-1} \sum_{k=0}^{2^j-1} |\beta_{jk}^{(i)}|^2. \] (13)

For a smoothness level \( s > 0 \), we define the product Sobolev space \( \mathcal{H}_n^s \) as the collection of vectors \( f \) of functions \( f = (f_1, \ldots, f_n) \) with each component
s-times weakly differentiable and norm \( \| f \|_{H^n} \) defined by

\[
\| f \|_{H^n} = \sum_{i=1}^{n} \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} 2^{2js} |\beta_{jk}^{(i)}|^2;
\]

here a function with wavelet coefficients \( \{ \beta_{jk} : k = 0, \ldots, 2^j - 1, j = 1, 2, \ldots \} \) is said to have \( s \) weak derivatives if \( \sum_{j=1}^{\infty} \sum_{k=0}^{2^j-1} 2^{2js} |\beta_{jk}|^2 < \infty \).

Let \( f_0 \) be the true value of the signal functions \( f \) and let \( \beta_{jk,0}^{(i)}, i = 1, \ldots, n, k = 0, 1, \ldots, 2^j - 1, j = 1, 2, \ldots \). Let \( \tau_0 \) stand for the true value of the synthetic change point \( \tau \), obtained as \( \arg\min\{C(k) : k = 2, \ldots, n\} \) from (4) and (5) by replacing \( \beta_{jk}^{(i)} \) by \( \beta_{jk,0}^{(i)}, i = 1, \ldots, n, k = 0, 1, \ldots, 2^j - 1, j = 1, 2, \ldots \). The prior parameters are chosen to be \( c_j^2 = \nu_1 2^{-\gamma_1 j} \) and \( \pi_j = \min(1, \nu_2 2^{-\gamma_2 j}) \) for some \( \nu_1, \nu_2, \gamma_1, \gamma_2 > 0 \).

The following result shows the consistency of estimation and change point detection.

**Theorem 1.** Let \( \gamma_1 > 2s + 1 \), and \( f_0 \in H^n \) be the vector of true functions. Then the posterior is consistent for estimation of the signals and detection of the change point, i.e., for any \( \epsilon > 0 \), \( \Pi(\| f - f_0 \| < \epsilon | D_r) \rightarrow 1 \) and \( \Pi(\tau = \tau_0 | D_r) \rightarrow 1 \) in probability as \( r \rightarrow \infty \).

Posterior consistency of the function will be established by applying Schwartz’s theory and its extensions (see Ghosal and van der Vaart [9]). To apply Schwartz’s posterior consistency theorem in this context, we need to establish that the parameter space is effectively compact. By direct moment calculations, we shall show that there exists a sufficiently large Sobolev norm-bounded set that contains almost all posterior probability, with high true probability.

After the posterior consistency of the function is established, we show that with high posterior probability, the similarity matrix identically agrees with the true similarity matrix of the collection \( \{ \beta_{jk}^{(i)} \} \), in true probability. This analysis comes down to showing that, for every \( k = 0, 1, \ldots, 2^j - 1, j = 1, 2, \ldots \), the posterior probability of forming blocks with ties in the values of \( \beta_{jk}^{(i)}, i = 1, \ldots, n \), agreeing with the blocks formed by the ties in the values of \( \beta_{jk,0} \), tends to one, in the true probability. This is shown by a combined application of posterior consistency of the function established in the first part of the theorem, and a consistency property of Bayes factors testing between possible structures established by a fine analysis of marginal...
likelihoods in conjugate normal models. In particular, the posterior consistency allows us to rule out certain structures as unlikely in the posterior and concentrate on structures that are only more liberal than the true structure, formed by these ties in the values of the wavelet coefficients. The argument is in a spirit similar to that used by Suarez and Ghosal [13] to establish posterior consistency of functional clustering in a similar setting, but it seems that they missed a growing factor. This factor needs to be properly controlled and hence their proof needed to be modified. We take care of the issue and our argument can also be used to rectify the proof of the posterior consistency result of Suarez and Ghosal [13].

To describe the model selection property, consider the following possible structures for consecutive values of $\beta_{jk}$:

(i) Change from nonzero to nonzero at $\tau_{jk} = t$:
$$S^{*,*}_{jk}(t) = \{\beta^{(1)}_{jk} = \cdots = \beta^{(t-1)}_{jk} = \xi_1, \beta^{(t)}_{jk} = \cdots = \beta^{(n)}_{jk} = \xi_2, \xi_1 \neq \xi_2 \neq 0\},$$
$t = 2, \ldots, n$;

(ii) Change from nonzero to zero at $\tau_{jk} = t$:
$$S^{*,0}_{jk}(t) = \{\beta^{(1)}_{jk} = \cdots = \beta^{(t-1)}_{jk} = \xi, \beta^{(t)}_{jk} = \cdots = \beta^{(n)}_{jk} = 0, \xi \neq 0\},$$
$t = 2, \ldots, n$;

(iii) Change from zero to nonzero at $\tau_{jk} = t$:
$$S^{0,*}_{jk}(t) = \{\beta^{(1)}_{jk} = \cdots = \beta^{(t-1)}_{jk} = 0, \beta^{(t)}_{jk} = \cdots = \beta^{(n)}_{jk} = \xi, \xi \neq 0\},$$
$t = 2, \ldots, n$;

(iv) No change and the value is nonzero:
$$S^{*,*}_{jk} = \{\beta^{(1)}_{jk} = \cdots = \beta^{(n)}_{jk} = \xi, \xi \neq 0\};$

(v) No change and the value is zero:
$$S^{0,0}_{jk} = \{\beta^{(1)}_{jk} = \cdots = \beta^{(n)}_{jk} = 0\}.$$

The structure of a random $(\beta^{(i)}_{jk} : i = 1, \ldots, n)$ drawn from its posterior distribution will be denoted by $S_{jk}$. The true model $S_{jk,0}$ is defined as the structure $(\beta_{jk,0}^{(i)} : i = 1, \ldots, n)$ follows. A model $S_{jk}$ is called incompatible if its closure does not contain $S_{jk,0}$, i.e., $(\beta_{jk,0}^{(i)} : i = 1, \ldots, n)$ cannot be approached by a sequence of values within $S_{jk}$. For instance, if $\beta_{jk,0}^{(1)} = \cdots = \beta_{jk,0}^{(t-1)} = \xi_1 \neq \xi_2 = \beta_{jk,0}^{(t)} = \cdots = \beta_{jk,0}^{(n)}$, then the models $S_{jk}^{*,0}(t), S_{jk}^{0,*}(t), S_{jk}$ and $S_{jk}^{0,0}$ are all incompatible, and so are (i)–(iii) for a wrong value of $t$. By posterior consistency for $\beta_{jk}$, posterior probability of any incompatible model converges to zero in probability. A compatible model can, however, be strictly bigger than the true model. For instance,
all models are compatible with $S^0_{jk}$, since zero can be approximated by non-zero sequences arbitrarily well. Possibility of the occurrence of a compatible model strictly bigger than the true model cannot be ruled out by posterior consistency. The following lemma, a key ingredient in the proof of posterior consistency of the change point, shows that the true model is indeed selected with high posterior probability, under the true distribution.

**Lemma 1.** Let $S_{jk,0}$ denote the true structure for given $j,k$. Then $\Pi(S_{jk} = S_{jk,0}|D_r) \to 1$ in probability as $r \to \infty$.

Proofs of Theorem 1 and Lemma 1 are given in Section 7.

5. **Simulation**

In order to study the performance of our method, we implement it on a set of simulated data. Since we detect the change point through the features extracted by the DWT, we first generate the wavelet coefficients and then apply the inverse discrete wavelet transform to get the functional data. We generate 16 features for the first data point by a uniform distribution on $[0,0.5]$. To make the change distinguishable, we generate 16 features for the last data point by a uniform distribution on $[0.5,1]$. Suppose that we have 100 data points, and there is one change in each feature. We randomly sample 16 numbers from 3 to 98 and regard them as the change point for these 16 features. To generate the sequence of 100 data units, we repeat the feature of the first data point and change it to that of the last data point after the change point. Thus, we have a sequence of data represented by the true feature values. After applying the inverse discrete wavelet transform, we get a sequence of 100 true signals. To generate features for the data, we sample from the normal distribution with the true feature values as the mean value and variance of 0.01, 0.1, and 1, respectively. Hence we get three sequences of 100 functional observations after applying the inverse discrete wavelet transform to them.

We apply our method to the observations. For the true signals, we use (4) to compute the similarity. The change point is the value of $k$ where $C(k)$ in (5) is the minimum. Once we detect the first change point, we divide the sequence of data into two subgroups. Furthermore, we can find the change point in these two groups. We can continue the process to divide the data into more subgroups, and stop either of the following conditions are met:
Table 1: Change point detection results comparison

<table>
<thead>
<tr>
<th>Method</th>
<th>Variance</th>
<th>Change points</th>
</tr>
</thead>
<tbody>
<tr>
<td>True signal</td>
<td>0.00</td>
<td>26(^{(2)}), 49(^{(1)}), 69(^{(2)})</td>
</tr>
<tr>
<td>Our</td>
<td>0.01</td>
<td>26(^{(2)}), 49(^{(1)}), 69(^{(2)})</td>
</tr>
<tr>
<td>E-Divisive</td>
<td>0.01</td>
<td>25(^{(2)}), 49(^{(1)}), 75(^{(2)})</td>
</tr>
<tr>
<td>Our</td>
<td>0.10</td>
<td>26(^{(2)}), 49(^{(1)}), 68(^{(3)}), 75(^{(2)}), 92(^{(3)})</td>
</tr>
<tr>
<td>E-Divisive</td>
<td>0.10</td>
<td>25(^{(2)}), 49(^{(1)}), 62(^{(3)}), 75(^{(2)}), 88(^{(3)})</td>
</tr>
<tr>
<td>Our</td>
<td>1.00</td>
<td>5(^{(3)}), 24(^{(2)}), 40(^{(3)}), 46(^{(1)}), 94(^{(3)}), 99(^{(2)})</td>
</tr>
<tr>
<td>E-Divisive</td>
<td>1.00</td>
<td>42(^{(1)}), 70(^{(2)})</td>
</tr>
</tbody>
</table>

- the plot of \(C(k)\) versus \(k\) is relatively flat which means that there is not much difference in these data;
- the minimum number of data points is reached;
- the maximum step of the resulting binary tree is reached.

In this study, we stop either if \(\max(C(k)) - \min(C(k)) < 0.1\), or there are less than 10 data points in the group, or the resulting binary tree has 3 steps. We compare our results with the E-Divisive method (James and Matterson [10]) in the R-package ecp, which also estimates multiple change points by iteratively applying a procedure for locating a single change point. We apply the E-Divisive method on the wavelet transform of the observations. Table 1 shows the change points for different sequences of observations detected by our method and E-Divisive method. The numbers in the parentheses denote the hierarchical order of the change points. When the variance is small (\(\sigma^2 = 0.01\)), the change points our method detects are exactly the same as the true change points. With a larger variance (\(\sigma^2 = 0.1\)), our method still can detect most change points correctly. When the variance is large (\(\sigma^2 = 1\)), naturally it would be difficult to detect the change points by any method.

6. Application to a climate change dataset

On Berkeley Earth (http://berkeleyearth.org/data/), we can find the land-surface monthly average temperature between 1753–2016. These temperatures are in degrees Celsius and reported as anomalies relative to the average temperature from January 1951 to December 1980. We can construct a set of functional data by the 12 monthly average temperatures
in each year. We smooth the data by the basis expansion. Thus we get 264 functional data ordered by the year. We suspect that there is a change in these functional data. Therefore we apply our method and find a change point of this sequence of functional data at the year 1914 by minimizing $C(k)$ with respect to $k$ in (5), the mean similarity across groups on the two sides of the change point, relative to the average values within these groups.

We apply the procedure again on these two segments 1753–1913 and 1914–2016. The corresponding change points are found to be at the years 1839 and 1969. The plots of these functional datasets for each blocks 1753–1838, 1839–1913, 1914–1968, and 1969–2016 are plotted in Figure 1.

![Figure 1: The plot on the left is the land-surface average temperature curves between given year ranges.](image)

The procedure can be continued longer to split into further subgroups, and stop subdividing a group further if $\max(C(k)) - \min(C(k)) < 0.1$ in that group. This will lead to 15 subgroups formed by the change points 1761, 1767, 1801, 1812, 1821, 1839, 1858, 1878, 1898, 1914, 1942, 1969 and 1994.

Posterior consistency results in Section 4 require replication of functional data, which we cannot have in this context. A possible way to address to issue is to artificially create replication by average daily temperatures over a block of years, say over a decade. Then change points will be identified as certain decades. We do not carry out this modification on this dataset.

### 7. Proof of the theorem

*Proof of Theorem 1.* We treat only the continuous case $J = \infty$. For the discrete case, $J$ is an increasing sequence which in the bounds may be replaced by infinity and the arguments of the continuous case can then apply.
First we show that
\[ \lim_{B \to \infty} \sup_{r > 0} E_{f_0} \Pi(\mathcal{H}_n^s(B)^c | D_r) = 0. \]  

(15)

This will imply that the parameter space can be effectively taken to be \( \mathcal{H}_n^s(B) \) for a sufficiently large \( B > 0 \), on which Schwartz’ theory for posterior consistency can be applied.

To prove (15), we apply Markov’s inequality to estimate
\[ \Pi(\mathcal{H}_n^s(B)^c | D_r) \leq B^{-2} \left\{ \sum_{i=1}^{n} \sum_{j=0}^{\infty} 2^{2j} \sum_{k=0}^{2j-1} E(|\beta_{jk}(i)|^2 | D_r) \right\}. \]

(16)

The expectation \( E(|\beta_{jk}(i)|^2 | D_r) \) above is given by
\[ \sum_{t=1}^{n} E\left(|\beta_{jk}(i)|^2 | \tau_{jk} = t, D_r\right) \Pi(\tau_{jk} = t | D_r) \leq \max_{1 \leq t \leq n} E\left(|\beta_{jk}(i)|^2 | \tau_{jk} = t, D_r\right). \]

The posterior density of the common value \( \xi \) of \( (\beta_{jk}(i) : i = 1, \ldots, t-1) \) given \( (\beta_{jk} \neq 0, \tau_{jk} = t, D_r) \) is proportional to \( \exp \left\{ -\left( r/2\sigma^2 \right) \sum_{l=1}^{t-1} (b_{jk}^{(l)})^2 - \xi^2/(2c_j^2\sigma^2) \right\} \), giving the posterior distribution \( N\left( \frac{c_j^2 \sum_{l=1}^{t-1} b_{jk}^{(l)}}{(t-1)c_j^2 + 1/r}, \frac{c_j^2\sigma^2}{1+(t-1)c_j^2 r} \right) \). Thus \( E(|\beta_{jk}(i)|^2 | \tau_{jk} = t, D_r) \) is given by the expressions
\[ \frac{c_j^2\sigma^2}{1 + (t-1)c_j^2 r} + \left( \frac{c_j^2}{(t-1)c_j^2 + 1/r} \right)^2 \left( \sum_{l=1}^{t-1} b_{jk}^{(l)} \right)^2 \quad \text{if } i < t, \]
\[ \frac{c_j^2\sigma^2}{1 + (n-t+1)c_j^2 r} + \left( \frac{c_j^2}{(n-t+1)c_j^2 + 1/r} \right)^2 \left( \sum_{l=t}^{n} b_{jk}^{(l)} \right)^2 \quad \text{if } i \geq t. \]

Note that if \( t = 1 \), we only need to consider the former case. Either expression can be bounded by \( c_j^2\sigma^2 + \left( \frac{c_j^2}{c_j^2 + 1/r} \right)^2 \left( \sum_{l=1}^{n} b_{jk}^{(l)} \right)^2 \), which leads to the estimate
\[ \Pi(\mathcal{H}_n^s(B)^c | D_r) \leq B^{-2} \left\{ \sum_{i=1}^{n} \sum_{j=0}^{\infty} 2^{2js} \sum_{k=0}^{2j-1} \left[ c_j^2\sigma^2 + \left( \frac{c_j^2}{c_j^2 + 1/r} \right)^2 \left( \sum_{l=1}^{n} b_{jk}^{(l)} \right)^2 \right] \right\}. \]
To prove the claim, it therefore suffices to show that $E_{f_0} \Pi(\mathcal{H}_n^s(B) | D_r)$, which is bounded by

$$B^{-2} \left\{ n \sum_{j=0}^\infty 2^{2j} \sigma^2 \sum_{k=0}^{2^j-1} \left[ c_j^2 \sigma^2 + \left( \frac{c_j^2}{c_j^2 + 1/r} \right)^2 \left( \frac{n \sigma^2}{r} + \left( \sum_{l=1}^n \beta_{jk,0}^{(l)} \right)^2 \right) \right] \right\} \leq B^{-2} \left\{ n \sum_{j=0}^\infty 2^{2j} \sigma^2 \sum_{k=0}^{2^j-1} \left[ c_j^2 \sigma^2 + \sum_{l=1}^n \beta_{jk,0}^{(l)} \right] \right\},$$

is arbitrarily small for sufficiently large $B > 0$. Recalling that $c_j^2 = \nu_1 2^{-\gamma_1 j}$, the term inside the bracket reduces to

$$n(n+1) \sigma^2 \nu_1 \sum_{j=0}^\infty 2^{(2s+1-\gamma_1)j} n^2 \sum_{j=0}^\infty 2^{2j} \sigma^2 \sum_{k=0}^{2^j-1} \sum_{l=1}^n |\beta_{jk,0}^{(l)}|^2$$

$$= n(n+1) \sigma^2 \nu_1 \sum_{j=0}^\infty 2^{(2s+1-\gamma_1)j} n^2 \|f_0\|_{\mathcal{H}_n^s}^2 < \infty$$

provided that $\gamma_1 > 2s + 1$, completing the proof of the claim.

In view of (15), for posterior consistency, we may assume that the parameter space is $\mathcal{H}_n^s(B)$ for some $B > 0$, which is norm-compact (see, example, Lemma 6.4 of Belitser and Ghosal [7], from where finiteness of covering numbers, and hence compactness, follow). Since the weak topology on the corresponding distributions is weaker and is Hausdorff, the weak and norm topologies coincide on norm-compact sets. It therefore suffices to prove posterior consistency with respect to the weak topology only. According to Example 6.20 in Ghosal and van der Vaart [9], if the prior satisfies the Kullback–Leibler property (Definition 6.15 of Ghosal and van der Vaart [9]), then the posterior distribution is consistent in the weak topology. In our setting, the Kullback–Leibler divergence is easily seen to be proportional to $\|f - f_0\|^2$, and hence we need to verify that for all $\epsilon > 0$, $\Pi \left( \sum_{j=1}^n \sum_{i=0}^\infty \sum_{k=0}^{2^j-1} |\beta_{jk}^{(i)} - \beta_{jk,0}^{(i)}|^2 < \epsilon^2 \right) > 0$, which follows from Lemma 1 and 2 of Lian [11]. This completes the proof of posterior consistency of the function $f$ at its true value $f_0$.

To show the posterior consistency of the synthetic change point $\tau$ at its true value $\tau_0$, as the weights over levels $2^{-j}$ are summable, it suffices to prove posterior consistency of

$$\{ \beta_{jk}^{(i)} = \beta_{jk,0}^{(i)} \} : k = 0, \ldots, 2^j - 1, j = 1, \ldots, J, i \neq i', i, i' = 1, \ldots, n \}$$
at
\[ \{ \beta_{jk,0}^{(i)} = \beta_{jk,0}^{(i')} : k = 0, \ldots, 2^j - 1, j = 1, \ldots, J, i \neq i', i, i' = 1, \ldots, n \} \]
for any fixed \( J \). This is equivalent to showing that for all fixed \( j, k \), \( \Pi(S_{jk} = S_{jk,0} | D_r) \rightarrow 1 \) in probability, which is obtained by Lemma 1.

**Proof of Lemma 1.** Because incompatible structures have asymptotically negligible posterior probabilities in view of posterior consistency, it suffices to show that the ratio of the marginal likelihood of a compatible structure other than the true structure and that of the true structure goes to zero in probability. If the true structure is \( S_{jk}^* \), then the result is immediate, as there are no other compatible structures. In this proof, we only deal with the cases when the true structure is \( S_{jk}^* \) or \( S_{jk}^0 \). The proofs for other cases follow by similar arguments.

The marginal likelihood \( p(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)} | S_{jk}^*(t)) \) of \( S_{jk}^*(t) \) is
\[
\int \int \prod_{l=1}^{t-1} \phi(b_{jk}^{(l)}; \xi_1, \frac{\sigma^2}{r}) \prod_{l=t}^{n} \phi(b_{jk}^{(l)}; \xi_2, \frac{\sigma^2}{r}) \phi(\xi_1; 0, c_j^2 \sigma^2) \phi(\xi_2; 0, c_j^2 \sigma^2) d\xi_1 d\xi_2
\]
\[
= (c_j^2 r(t-1) + 1)^{-1/2} (c_j^2 r(n-t+1) + 1)^{-1/2} \left( \frac{2\sigma^2}{r} \right)^{-n/2}
\]
\[
\times \exp \left\{ \frac{r}{2\sigma^2} \left[ \frac{c_j^2}{c_j^2(t-1) + 1/r} \left( \sum_{l=1}^{t-1} b_{jk}^{(l)} \right)^2 
\right.
\]
\[
\left. + \frac{c_j^2}{c_j^2(n-t+1) + 1/r} \left( \sum_{l=t}^{n} b_{jk}^{(l)} \right)^2 - \sum_{l=1}^{n} (b_{jk}^{(l)})^2 \right] \right\}.
\]

The marginal likelihood \( p(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)} | S_{jk}^0(t)) \) of \( S_{jk}^0(t) \) is
\[
\int \prod_{l=1}^{t-1} \phi(b_{jk}^{(l)}; \xi, \sigma^2/r) \prod_{l=t}^{n} \phi(b_{jk}^{(l)}; 0, \sigma^2/r) \phi(\xi; 0, c_j^2 \sigma^2) d\xi
\]
\[
= (c_j^2 r(t-1) + 1)^{-1/2} \left( \frac{2\sigma^2}{r} \right)^{-n/2}
\]
\[
\times \exp \left\{ \frac{r}{2\sigma^2} \left[ \frac{c_j^2}{c_j^2(t-1) + 1/r} \left( \sum_{l=1}^{t-1} b_{jk}^{(l)} \right)^2 
\right. \right.
\]
\[
\left. - \sum_{l=1}^{n} (b_{jk}^{(l)})^2 \right] \right\}.
\]
The marginal likelihood \( p(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}|S_{jk}^{0,*}(t)) \) of \( S_{jk}^{0,*}(t) \) is

\[
\int \prod_{l=1}^{t-1} \phi(b_{jk}^{(l)}, 0, \sigma^2/r) \prod_{l=t}^{n} \phi(b_{jk}^{(l)}, \xi, \sigma^2/r) \phi(\xi; 0, c_j^2 \sigma^2) d\xi
\]

\[
= (c_j^2 r (n - t + 1) + 1)^{-1/2} (2\pi \sigma^2/r)^{-n/2}
\]

\[
\times \exp \left\{ \frac{r}{2\sigma^2} \left[ \frac{c_j^2}{c_j^2 (n - t + 1) + 1/r} \left( \sum_{l=t}^{n} b_{jk}^{(l)} \right)^2 - \sum_{l=1}^{n} (b_{jk}^{(l)})^2 \right] \right\}.
\]

The marginal likelihood \( p(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}|S_{jk}^*) \) of \( S_{jk}^* \) is

\[
\int \prod_{l=1}^{n} \phi(b_{jk}^{(l)}; \xi, \sigma^2/r) \phi(\xi; 0, c_j^2 \sigma^2) d\xi
\]

\[
= [(c_j^2 r n + 1)(2\pi \sigma^2/r)^n]^{-1/2} \exp \left\{ \frac{r}{2\sigma^2} \left[ \frac{c_j^2}{c_j^2 n + 1/r} \left( \sum_{l=1}^{n} b_{jk}^{(l)} \right)^2 - \sum_{l=1}^{n} (b_{jk}^{(l)})^2 \right] \right\}.
\]

Finally the marginal likelihood \( p(b_{jk}^{(1)}, \ldots, b_{jk}^{(n)}|S_{jk}^0) \) of \( S_{jk}^0 \) is

\[
\prod_{l=1}^{n} \phi(b_{jk}^{(l)}, 0, \sigma^2/r) = (2\pi \sigma^2/r)^{-n/2} \exp \left\{ -\frac{r}{2\sigma^2} \sum_{l=1}^{n} (b_{jk}^{(l)})^2 \right\}.
\]

Let the true model \( S_{jk,0} \) be \( S_{jk}^0 \). To prove model selection consistency in this case, it suffices to show that the ratio of the marginal likelihoods of each models \( S_{jk}^{*\ast}(t), S_{jk}^{0,*}, \) and \( S_{jk}^{0,*}, t = 2, \ldots, n, \) to that of \( S_{jk}^* \), converges to 0. In the first case, the ratio is given by

\[
\sqrt{\frac{c_j^2 r n + 1}{(c_j^2 r (t-1) + 1)(c_j^2 r (n-t+1) + 1)}} \exp \left\{ \frac{r}{2\sigma^2} \left[ \frac{c_j^2}{c_j^2 (t-1) + 1/r} \left( \sum_{l=1}^{t-1} b_{jk}^{(l)} \right)^2 + \frac{c_j^2}{c_j^2 (n-t+1) + 1/r} \left( \sum_{l=t}^{n} b_{jk}^{(l)} \right)^2 \right] \right\}.
\]

The first term in the square root goes to 0 as \( r \to \infty \). Hence it suffices to
show that the form inside the exponential is bounded. As \( r \to \infty \),

\[
\frac{r}{2\sigma^2} \left[ \frac{c_j^2}{c_j} (t - 1) + \frac{1}{r} \right] + \frac{c_j^2}{c_j^2} (n - t + 1) + \frac{1}{r} - \frac{c_j^2}{c_j^2 n + \frac{1}{r}} \left( \sum_{l=1}^{n} b_{jk}^{(l)} \right)^2
\]

\[
\begin{align*}
= \frac{r}{2\sigma^2} \left[ \frac{\sum_{l=1}^{t-1} b_{jk}^{(l)}}{(t - 1) + O(\frac{1}{r})} + \frac{\sum_{l=1}^{n} b_{jk}^{(l)}}{(n - t + 1) + O(\frac{1}{r})} - \frac{\sum_{l=1}^{n} b_{jk}^{(l)}}{n + O(\frac{1}{r})} + O(1) \right]
\end{align*}
\]

\[
= \frac{r}{2\sigma^2} \left[ \frac{\sum_{l=1}^{t-1} b_{jk}^{(l)}}{(t - 1)} + \frac{\sum_{l=1}^{n} b_{jk}^{(l)}}{(n - t + 1)} - \frac{\sum_{l=1}^{n} b_{jk}^{(l)}}{n} \right] + O(1). \tag{17}
\]

Let \( U = \sum_{l=1}^{t-1} b_{jk}^{(l)} \) and \( V = \sum_{l=1}^{n} b_{jk}^{(l)} \). Consider a random variable \( W \) which is \( U/(t-1) \) with probability \((t-1)/n\) and is \( V/(n-t+1) \) with probability \((n-t+1)/n\). Then Jensen’s inequality implies that \( U^2/(t-1) + V^2/(n-t+1) \geq (U + V)^2/n \). Thus, the term in the brackets of the exponential in (17) is nonnegative. Hence it suffices to control its expectation and show that it remains bounded as \( r \to \infty \). Suppose that the true value is \( \beta_{jk,0}^{(i)} = \xi \). Then the expectation of \( O(1) \) of the first term of (17) with respect to the true value is

\[
\frac{r}{2\sigma^2} \left[ \frac{(t - 1)^2 \xi^2 + \frac{(t-1)^2}{r}}{(t - 1)} + \frac{(n - t + 1)^2 \xi^2 + \frac{(n-t+1)^2}{r}}{(n - t + 1)} - \frac{n^2 \xi^2 + \frac{n}{r}}{n} \right],
\]

which reduces to \( \frac{r^2}{2\sigma^2} \left[ (t - 1) + (n - t + 1) - n \right] + O(1) = O(1) \), as the first term vanishes. This completes the convergence proof in the present case.

If the true structure is \( S_{jk}^0 \) and the compatible structure is \( S_{jk}^{*,*}(t) \), the marginal likelihood ratio is

\[
\left( c_j^2 r (t - 1) + 1 \right)^{-1/2} \left( c_j^2 r (n - t + 1) + 1 \right)^{-1/2}
\]

\[
\times \exp \left\{ \frac{r}{2\sigma^2} \left[ \frac{c_j^2}{c_j^2} (t - 1) + \frac{1}{r} \left( \sum_{l=1}^{t-1} b_{jk}^{(l)} \right)^2 + \frac{c_j^2}{c_j^2} (n - t + 1) + \frac{1}{r} \left( \sum_{l=t}^{n} b_{jk}^{(l)} \right)^2 \right] \right\}.
\]

The first two terms in the square root go to 0 as \( r \to \infty \). The exponential term is \( \frac{r}{2\sigma^2} \left[ \frac{\sum_{l=1}^{t-1} b_{jk}^{(l)}}{(t-1)} + \frac{\sum_{l=t}^{n} b_{jk}^{(l)}}{(n-t+1)} \right] + O(1) \). As \( \beta_{jk,0}^{(i)} = 0 \) and the first term is always nonnegative, its expectation is

\[
\frac{r}{2\sigma^2} \left[ \frac{(t - 1)^2}{(t - 1)} + \frac{(n - t + 1)^2}{(n - t + 1)} \right] = 1.
\]

This concludes the proof of the present case.

Other cases can be handled similarly. \( \square \)
References


