# Supplementary to "Model-free Multivariate Change Point Detection and Localization with Statistical Guarantee"

Xin Xing, Zuofeng Shang, Hongyu Miao, Pang Du*

This document contains the proofs of the main results Theorems 1 and 2. Proofs of Lemmas 1 and 2, as well as some auxiliary results, are also included. Before proofs, we provide some technical preliminaries.

## S.1 Algorithms

We summarize the change point detection algorithm in Algorithm 1. This algorithm employs penalized maximum likelihood estimation to provide optimal solutions that balance goodness-of-fit and complexity. By systematically scanning through the sample set and utilizing a model-free approach, it can identify change points without making strong assumptions about the underlying data distribution. This makes it potentially effective in diverse scenarios, particularly in situations where traditional parametric models may not hold or might introduce biases.

We summarize the localization procedure in Algorithm 2.

*Corresponding author: Pang Du, pangdu@vt.edu

**Algorithm 1** Model-free Change Point Detection based on CUNSUM-PLR Test

1. Input: Sample X.

2. For $\tau = n_0, \ldots, n - n_0$ calculate

$$\widehat{\eta}_p = \text{argmin }_{\eta \in \mathcal{H}} \ell_p^\tau(\eta),\ \widehat{\eta}_q = \text{argmin }_{\eta \in \mathcal{H}} \ell_q^\tau(\eta), \text{ and } \widehat{\eta}_0^n = \text{argmin }_{\eta \in \mathcal{H}} \ell_0^n(\eta).$$

   by penalized maximum likelihood estimate. Then, calculate

$$PLR^\tau = \ell_p^\tau(\widehat{\eta}_p^\tau) + \ell_q^{n-\tau}(\widehat{\eta}_q^{n-\tau}) - \ell_0(\widehat{\eta}_0).$$

   End For loop.

3. Calculate the test statistic,

$$CPLR = \sum_{\tau=n_0}^{n-n_0} PLR^\tau.$$

4. if $\Phi(\alpha) = 1$ then output the existence of change point.

**Algorithm 2** Model-free Change Point Localization

1. If the null hypothesis is rejected based on the findings of Algorithm 1, compute the likelihood function $\ell^\tau(\widehat{\eta}_p^\tau, \widehat{\eta}_q^{n-\tau})$ across the spectrum of $\tau = 0, \ldots, n$.

2. Pinpoint $\widehat{\tau}$ (as deduced from (3.13)) to be the most probable change point location, occurring between time markers $t_1$ and $t_2$.

Algorithm 2 represents a departure from conventional model-based methodologies. Leveraging the capabilities of nonparametric penalized likelihood estimation, this algorithm systematically explores the dataset to identify optimal change point locations with a high degree of rigor. Its adaptability and precision render it well-suited for a wide spectrum of datasets, guaranteeing dependable outcomes irrespective of the intricacies within the data. Beyond merely detecting change points, this approach enriches our understanding of the underlying shifts within the data, proving to be an invaluable asset in both scientific research and practical industrial applications.

### S.1.1 Additional settings for Multivaraite Analysis

When dealing with multivariate signals, we add scenarios to detect pure location changes in multivariate Gaussian distributions in Setting 3, where we keep the Kullback-Leibler (KL) divergence between the alternative distribution and null distribution the same as in our Setting 1. Then we add Setting 4 to combine both location changes and covariance changes together. we let the sequence length, denoted by n, range from 500 to 2000, and we position the change point at $n\theta$, where $\theta$ is randomly sampled from a uniform distribution within the interval (0.2, 0.8). We perform 200 replications to compute the empirical power and size.

**Setting 3 (Only Mean Change):** In this setting, the sequence $X_i$, $i = 1, \dots, n$ are i.i.d. generated as $X_i \sim \mathcal{N}_d(\mathbf{0}, I_d)$ if $i \le n\theta$ and $X_i \sim \mathcal{N}_d(\delta \cdot \mathbf{1}_d, I_d)$ if $i > n\theta$, where $I_d$ is the identity matrix and $\delta = 0.1458$ for $d = 2$ and $\delta = 0.2507$ for $d = 10$ representing the magnitude of mean shift and $\mathbf{1}_d$ is a vector of ones. Such $\delta$ makes the KL divergence between $\mathcal{N}(\mathbf{0}, I_d)$ and $\mathcal{N}(\delta \cdot \mathbf{1}_d, I_d)$ equal to the KL divergence between $\mathcal{N}_d(\mathbf{0}, I_d)$ and $\mathcal{N}_d(\mathbf{0}, \Sigma_\rho)$ for $\rho = 0.2$. This setting examines the method's ability to detect pure location changes in multivariate Gaussian distributions while maintaining constant covariance structure.

**Setting 4 (Combined Mean and Covariance Change):** In this comprehensive setting, we consider simultaneous changes in both location and scale parameters. The sequence $X_i$, $i = 1, \dots, n$ are i.i.d. generated as $X_i \sim \mathcal{N}_d(\mathbf{0}, I_d)$ if $i \le n\theta$ and $X_i \sim \mathcal{N}_d(\delta \cdot \mathbf{1}_d, \Sigma_\rho)$ if $i > n\theta$, where $I_d$ is the identity matrix, $\delta = 0.1458$ for $d = 2$ and $\delta = 0.2507$ for $d = 10$ represents the mean shift, and $\Sigma_\rho$ has diagonal entries equal to 2.0 and off-diagonal entries $\rho = 0.2$. This setting reflects realistic scenarios where distributional changes affect multiple parameters simultaneously, as commonly observed in applications like financial time series or biological signals where both the average level and variability patterns shift during regime changes.

As shown in Figure S.1(a-b), under the same KL divergence as Setting 1, all methods have higher

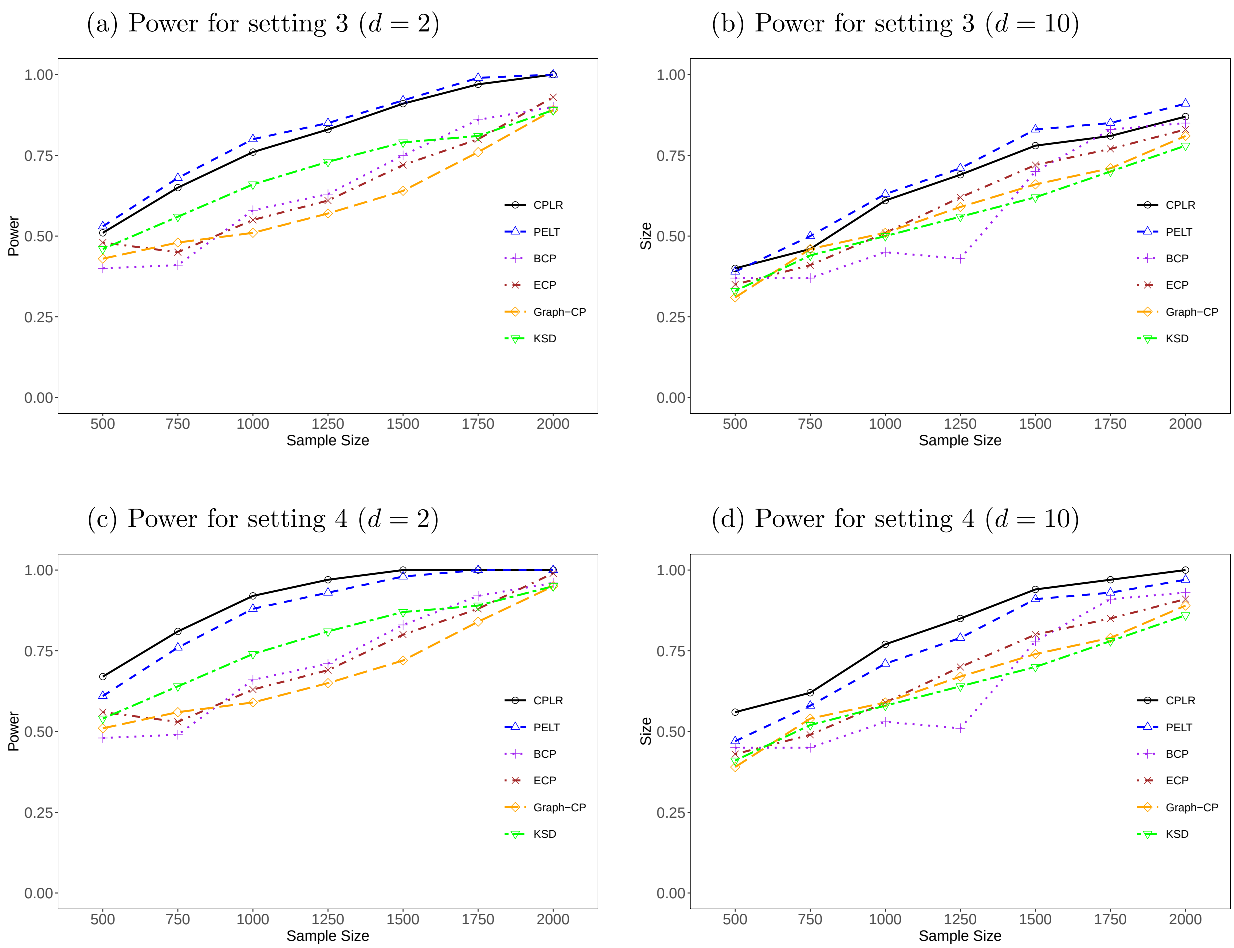


Figure S.1: Empirical power for Settings 3-4. Here the design matrix is generated from the multivariate normal distribution.

power compared with Setting 1. CPLR and PELT show comparably the highest power. When adding the covariance changes, shown in Figure S.1(c-d), CPLR's power increases significantly and becomes the best one, showing the ability of CPLR to be more sensitive in detecting complex distributional changes.

### S.1.2 Univariate Analysis

In this evaluation, we scrutinize the proficiency of the put forth test and the change point estimate, particularly in the scenario of a univariate signal. Our ambition is to discern the change points that have been provoked by a fluctuation in the distribution's aspects, such as variance, skewness, or the shape of the tail. Throughout these simulation scenarios, we handle sequence lengths $n$ spanning from 500 to 2000. We establish the change point's position at $n\theta$, where $\theta$ is a figure randomly extracted from a uniform distribution within the range (0.35, 0.65).

**Setting 5:** We consider the setting where the sequence of $X_i$, $i = 1, \ldots, n$ are from Gaussian distributions with different variances before and after the change point,

$$X_i \sim \begin{cases} N(0,1) & \text{if } i \leq n\theta, \\ N(0,\sigma_1^2) & \text{if } i > n\theta, \end{cases}$$

where the variance $\sigma_1^2$ takes values 1.2 .

**Setting 6:** We considered $X_i$, $i = 1, \ldots, n$ with the same mean and variance.

$$X_i = \begin{cases} N(0,1) & \text{if } i \leq n\theta, \\ Z_i^2 - 1 & \text{if } i > n\theta, \end{cases}$$

where $Z_i$s are i.i.d. generated from the standard normal distribution. The distributions of $X_i, i = 1, \ldots, n$ before and after the change point have the same mean and variance, but different skewnesses.

**Setting 7:** We consider a sequence whose distribution changes from a normal distribution to a heavy-tailed distribution at the change point, i.e.,

$$X_i \sim \begin{cases} N(0,1) & \text{if } i \leq n\theta, \\ \text{Cauchy}(0,1) & \text{if } i > n\theta, \end{cases}$$

where Cauchy$(0, 1)$ is the Cauchy distribution with the location parameter 0 and the scale parameter 1.

In all our simulation scenarios, we consistently set the significance level at 0.05. As clearly illustrated in Figure S.2(a), the size of the suggested test closely approximates 0.05 across a range

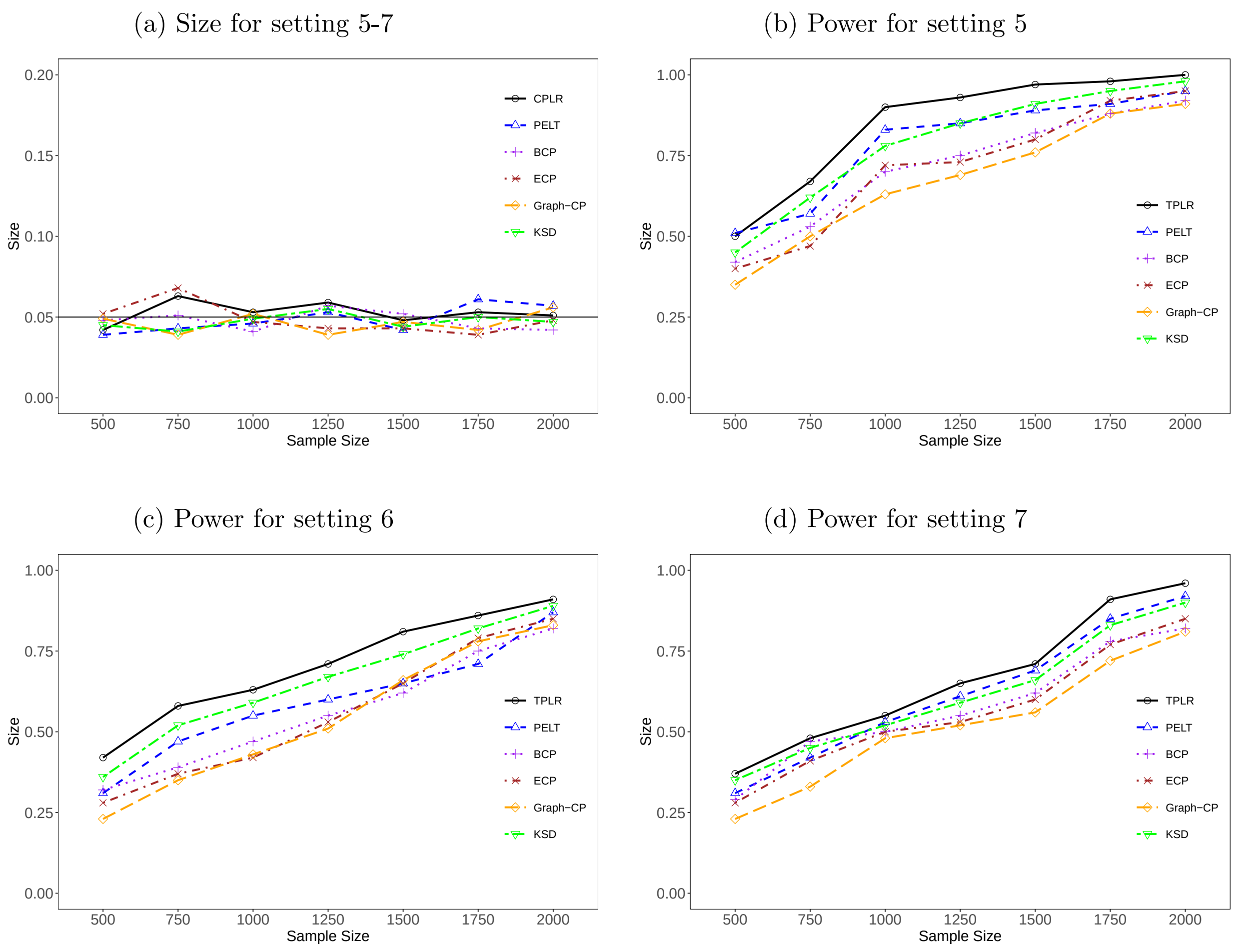


Figure S.2: Empirical power and FDP for setting 5-7. Here the design matrix is generated from the multivariate normal distribution.

of sequence lengths. This showcases the reliability of the test under varying data sizes. Notably, the efficacy of our proposed test becomes evident when observing its power. There's a rapid increase in the power as the sequence length grows, and it approaches a near-perfect value of 1 when the sequence length extends to 2000. Diving deeper into specific scenarios, Figure S.2(b) emphasizes the results under the Gaussian setting where a change point is introduced due to a variance shift. In this setting, both the PELT and CPLR tests emerge as the frontrunners, exhibiting the highest power among all tests considered. However, when we transition to non-Gaussian settings, specifically settings 6-7, the superiority of our proposed CPLR test becomes even more pronounced. It not only outperforms the PELT but also displays significantly higher power compared to all other competing methods.

Upon rejecting the null hypothesis, it becomes necessary to estimate the location of the change point. To demonstrate the convergence of our suggested change point estimate, we run simulations of setting 1 with $\sigma_1^2 = 1.2$, settings 4 and 5 (details of these settings are not provided here). We compute the relative distance $\widehat{r}$ between our estimate and the actual value as,

$$\widehat{r} = \frac{|\widehat{\tau} - \tau_0|}{n} \tag{S.1}$$

where $\widehat{\tau}$ is the proposed change point estimate, $\tau_0$ is the underlying truth and $n$ the total sequence length. In Table S.1, we report the mean and standard deviation of $\widehat{r}$ by repeating the simulation 200 times.

| Convergence Analysis for Univariate Settings | | | |
|---|---|---|---|
| n | Setting 5 | Setting 6 | Setting 7 |
| 500 | $0.051_{(0.053)}$ | $0.055_{(0.052)}$ | $0.053_{(0.055)}$ |
| 750 | $0.039_{(0.047)}$ | $0.035_{(0.035)}$ | $0.035_{(0.045)}$ |
| 1000 | $0.032_{(0.025)}$ | $0.030_{(0.027)}$ | $0.038_{(0.031)}$ |
| 1250 | $0.019_{(0.022)}$ | $0.024_{(0.025)}$ | $0.020_{(0.028)}$ |
| 1500 | $0.014_{(0.009)}$ | $0.012_{(0.012)}$ | $0.011_{(0.012)}$ |
| 1750 | $0.010_{(0.009)}$ | $0.006_{(0.008)}$ | $0.007_{(0.007)}$ |
| 2000 | $0.003_{(0.003)}$ | $0.003_{(0.001)}$ | $0.004_{(0.003)}$ |

Table S.1: Table lists the relative distances (and their standard deviations) of our proposed test for Settings 3-5 with sequence length ranging from 500 to 2000.

In Table S1, the convergence of our proposed change point estimate is demonstrated. Across all three scenarios, the estimation error diminishes quickly with the growth of the sample size, attesting to the robustness of our proposed change point location under various probabilistic distributions. When the sample size nears 2000, the estimation error is less than 0.005, empirically confirming our theoretical expectations regarding the convergence of the change point location.

## S.2 Temporal Dependence Test for Real Data

To validate the independence assumption critical to our statistical modeling framework, we conducted comprehensive empirical testing using the Ljung-Box test [Ljung and Box, 1978], which detects autocorrelation in time series by examining whether any group of autocorrelations differs significantly from zero. The original EEG recordings, collected at 2000Hz (0.5-millisecond intervals), introduce substantial temporal dependencies that violate standard statistical assumptions. However, since clinical seizure detection operates effectively within one-second temporal windows, we implemented systematic downsampling to balance statistical validity with clinical requirements. Our validation protocol progressively reduced the sampling frequency from 2000Hz through intermediate rates (1000Hz, 500Hz) to our target of 100Hz (10-millisecond intervals), using anti-aliasing filters to preserve essential frequency components. We analyzed two clinically validated EEG features: (1) Root Mean Squared (RMS) amplitude, which estimates cerebral function monitor (CFM) outputa standard metric in neonatal seizure monitoring [Shellhaas et al., 2011], and (2) Line Length, a fractal-based complexity measure proposed by Katz [1988] for adult epileptic seizure detection. These complementary features capture both amplitude variations and morphological characteristics critical for seizure identification. The temporal dependence assessment involved randomly selecting 100 non-overlapping 4-second windows from the EEG dataset and computing both features at each sampling rate.

As illustrated in Figure S.3, the Ljung-Box test results for the RMS amplitude feature reveal a statistically significant trend of decreasing temporal dependence with reduced sampling rates. At the original 2000Hz sampling rate, only 17% of windows showed non-significant autocorrelation (or 83% of windows showed significant autocorrelation) (mean p-value = 0.042), indicating strong

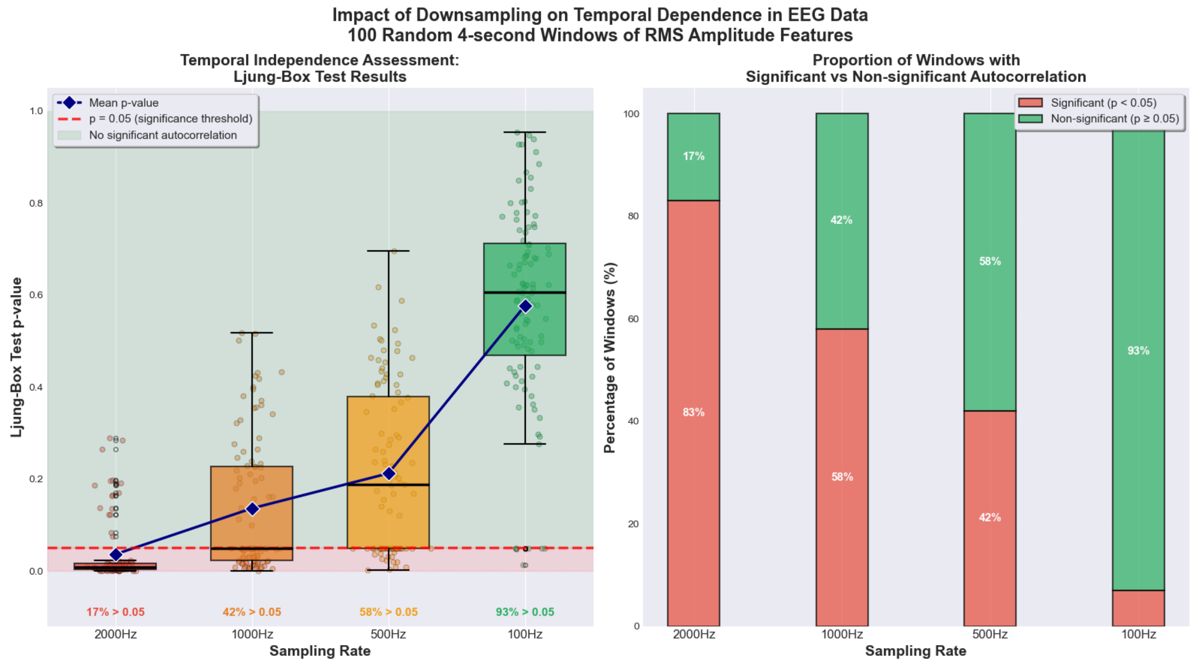


Figure S.3: Temporal independence assessment of RMS amplitude features using Ljung-Box tests across different sampling rates. Left panel: Distribution of Ljung-Box test p-values for 100 randomly sampled 4-second windows at each sampling rate (2000Hz, 1000Hz, 500Hz, 100Hz). The red dashed line indicates the significance threshold (p = 0.05), with the green shaded region representing non-significant autocorrelation. Right panel: Proportion of windows showing significant (red) versus non-significant (green) temporal autocorrelation at each sampling rate.

temporal dependence. Progressive downsampling yielded substantial improvements: 42% non-significant at 1000Hz (mean p-value = 0.129), 58% at 500Hz (mean p-value = 0.296), and ultimately 93% at our target 100Hz sampling rate (mean p-value = 0.598).

Figure S.4 demonstrates remarkably consistent results for the Line Length feature, exhibiting identical patterns of temporal independence across all sampling rates (21%, 45%, 72%, and 94% non-significant autocorrelation for 2000Hz, 1000Hz, 500Hz, and 100Hz respectively). This consistency across both features strengthens our conclusions regarding the effectiveness of the downsampling strategy.

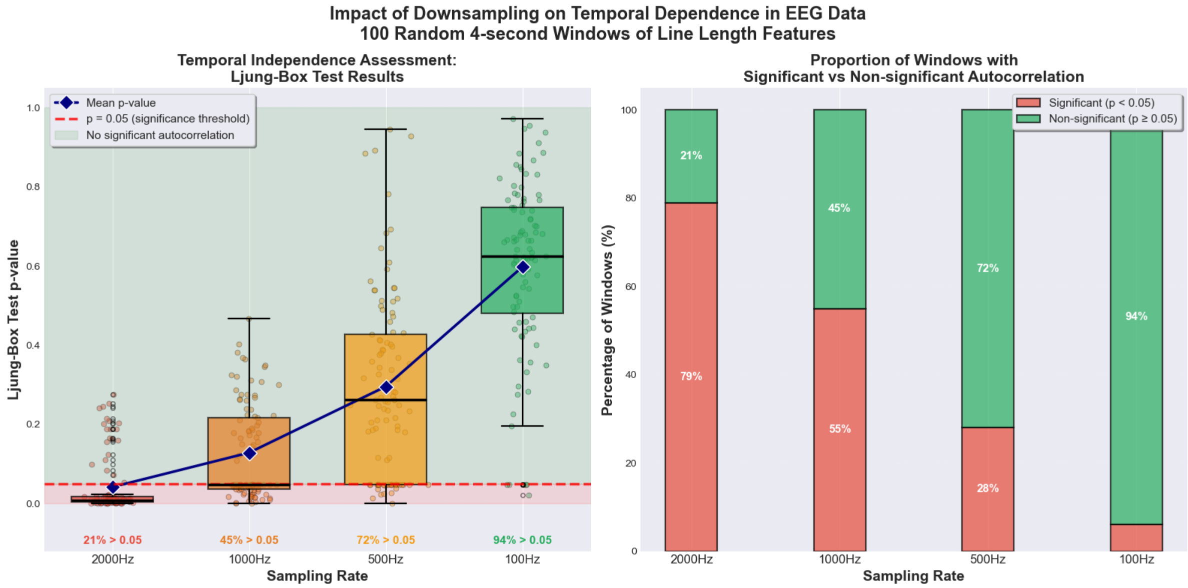


Figure S.4: Temporal independence assessment of Line Length features using Ljung-Box tests across different sampling rates. Left panel: Distribution of Ljung-Box test p-values for 100 randomly sampled 4-second windows at each sampling rate. The visualization follows the same format as Figure S.3, with the significance threshold at p = 0.05. Right panel: Proportion of windows with significant versus non-significant autocorrelation.

The empirical evidence from our Ljung-Box test analysis conclusively demonstrates that downsampling from 2000Hz to 100Hz effectively eliminates temporal autocorrelation while maintaining clinically sufficient temporal resolution. This sampling rate successfully balances statistical validity with clinical utility, providing adequate temporal resolution for accurate seizure detection within the required 1-second timeframe while ensuring the data meet the assumptions of our analytical framework.

## S.3 Additional Lemmas

**Lemma S.1.** *Under $H_1$, for any positive $M$, $r_n$, $h$ satisfying $c_K^2\sqrt{M}rh^{-1/2}A(h) \le 1/2$, we have*

$$\sup_{\eta\in\mathcal{H}} P\big(\|\widehat{\eta}_p^\tau - \eta_p^\tau\| \ge \delta_\tau(M,r)\big) \le 2\exp(-M\tau hr^2),$$

$$\sup_{\eta\in\mathcal{H}} P\big(\|\widehat{\eta}_q^{n-\tau} - \eta_q^{n-\tau}\| \ge \delta_{n-\tau}(M,r)\big) \le 2\exp(-M(n-\tau)hr^2),$$

*where $\delta_t(M,r) = 2h^m + c_K(\sqrt{2M}r + (th)^{-1/2})$, $t = \tau, n-\tau$.*

The proof of Lemma S.1 is similar to Lemma 1. This lemma gives a uniform bound for the MPLE under the alternative hypothesis $H_1$. Next, we use the score functional to approximate the penalized likelihood estimator.

**Lemma S.2.** *Under $H_1$, for any positives $h, r, M$ satisfying $c_K^2\sqrt{M}rh^{-1/2}A(h) \le 1/2$, it holds that*

$$\sup_{\eta_p,\eta_q\in\mathcal{H}} P_f(\|\widehat{\eta}_p^\tau - \eta_p^\tau - (\frac{1}{\tau}\sum_i K_{\boldsymbol{X}_i} - \mathbb{E}[K_{\boldsymbol{X}_i}] - W_\lambda\eta_p^\tau)\| \ge \gamma_\tau(M,r))$$
$$\le 2\exp(-M\tau hr^2),$$
$$\sup_{\eta_p,\eta_q\in\mathcal{H}} P_f(\|\widehat{\eta}_q^{n-\tau} - \eta_q^{n-\tau} - (\frac{1}{n-\tau}\sum_i K_{\boldsymbol{X}_i} - \mathbb{E}[K_{\boldsymbol{X}_i}] - W_\lambda\eta_q^{n-\tau})\| \ge \gamma_{n-\tau}(M,r))$$
$$\le 2\exp(-M(n-\tau)hr^2)$$

*where $\gamma_t(M,r) = c_K^2\sqrt{M}rh^{-1/2}A(h)\sigma_t(M,r)$, $t = \tau, n-\tau$, and $\sigma_t(M,r)$ is defined in Lemma 2.*

## S.4 Proof of Theorem 1

Begin of the proof:

$$\begin{aligned} PLR^\tau &= \ell^\tau(\widehat{\eta}_p^\tau) + \ell^{n-\tau}(\widehat{\eta}_q^{n-\tau}) - \ell^0(\widehat{\eta}_0) \\ &= (\ell^\tau(\widehat{\eta}_p^\tau) - \ell_0^\tau(\widehat{\eta}_0)) + (\ell^{n-\tau}(\widehat{\eta}_q^{n-\tau}) - \ell_0^{n-\tau}(\widehat{\eta}_0)) \\ &\equiv I_1 + I_2 \end{aligned} \tag{S.2}$$

where $\ell_0^\tau(\widehat{\eta}_0) = \ell^\tau(\widehat{\eta}_0) + \ell^{n-\tau}(\widehat{\eta}_0)$. Using (9) and (10), we calculate the tylor expansion on the $I_1$. Let $g = \widehat{\eta}_0 - \widehat{\eta}^\tau$, we have

$$\begin{aligned}
I_1 &= D\ell^\tau(\widehat{\eta}_p^\tau)g + \int_0^1\int_0^1 sD^2\ell^\tau(\widehat{\eta}_p^\tau + ss'g)ggdsds' \\
&= \int_0^1\int_0^1 s\{D^2\ell^\tau(\widehat{\eta}_p^\tau + ss'g)gg - D^2\ell^\tau(\eta_0)gg\}dsds' + \frac{1}{2}D^2\ell^\tau(\eta_0)gg \\
&\equiv I_{11} + I_{12} \qquad\qquad \text{(S.3)}
\end{aligned}$$

where $\eta_0$ is the underlying truth. We will show that $I_{12}$ is a leading term compared with $I_{11}$. Let us frist analyze $I_1$. Let $\widetilde{g} = \widehat{\eta}_p^\tau + ss'g - \eta^*$, for any $0 \le s, s' \le 1$, we have

$$D^2\ell^\tau(\widehat{\eta}^\tau + ss'g)gg = D^2\ell^\tau(\widehat{\eta}^\tau + \eta^*)gg = \tau\int_{\mathcal{X}} g^2(\boldsymbol{x})e^{\widetilde{g}(\boldsymbol{x})+\eta_0(\boldsymbol{x})}d\boldsymbol{x} + \lambda J(g,g), \qquad \text{(S.4)}$$

and

$$D^2\ell_{n,\lambda}(\eta_0)gg = \tau\int_{\mathcal{X}} g^2(\boldsymbol{x})e^{\eta_0(\boldsymbol{x})}d\boldsymbol{x} + \tau\lambda J(g,g). \qquad \text{(S.5)}$$

Combining (S.4) and (S.5), we have

$$|D^2\ell_{n,\lambda}(\widehat{\eta}_p^\tau + ss'g)gg - D^2\ell_{n,\lambda}(\eta_0)gg| \le \tau\int_{\mathcal{X}} g^2(\boldsymbol{x})e^{\eta_0(\boldsymbol{x})}|e^{\widetilde{g}(\boldsymbol{x})} - 1|d\boldsymbol{x}.$$

By Taylor expansion of $e^{\widetilde{g}(\boldsymbol{x})+\eta_0(\boldsymbol{x})}$ at $\eta_0(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathcal{X}$, it trivially holds that $e^{\eta_0(\boldsymbol{x})}|e^{\widetilde{g}(\boldsymbol{x})} - 1| = e^{\eta_0(\boldsymbol{x})}O(|\widetilde{g}(\boldsymbol{x})|)$. Since $\sup_{\boldsymbol{x}\in\mathcal{X}} |\widetilde{g}(\boldsymbol{x})| \le c_m h^{-1/2}\|\widetilde{g}\|$ (Lemma 3.1), and $h^{-1/2}((nh)^{-1} + \lambda)^{1/2} = o(1)$, we have

$$|I_{11}| = O_P(h^{-1/2}(\tau\|\widehat{\eta}_{n,\lambda} - \eta_0\| + \|g\|)\cdot\|g\|^2) = o_P(\tau\|g\|^2).$$

From (10), we have that $I_{12} = \frac{\tau}{2}\|g\|^2 = \frac{\tau}{2}\|\widehat{\eta}_p^\tau - \widehat{\eta}_0\|^2$. Thus, we have

$$I_1 = \frac{\tau}{2}\|\widehat{\eta}_p^\tau - \widehat{\eta}_0\|^2 + o_P(\tau\|\widehat{\eta}_p - \widehat{\eta}_0\|^2). \qquad \text{(S.6)}$$

By the similar derivation, we have the expansion of $I_2$ as

$$I_2 = \frac{n-\tau}{2}\|\widehat{\eta}_q - \widehat{\eta}_0\|^2 + o_P((n-\tau)\|\widehat{\eta}_q - \widehat{\eta}_0\|^2). \qquad \text{(S.7)}$$

Then, we combine (S.6) and (S.7) and get

$$PLR^{\tau} = \frac{\tau}{2}\|\widehat{\eta}_p^{\tau} - \widehat{\eta}_0\|^2 + \frac{n-\tau}{2}\|\widehat{\eta}_q - \widehat{\eta}_0\|^2 + o_P(\max\{\tau\|\widehat{\eta}_p - \widehat{\eta}_0\|^2, (n-\tau)\|\widehat{\eta}_q - \widehat{\eta}_0\|^2\}) \tag{S.8}$$

Next, we use Lemmas 1 and 2 to approximate the two norms in (S.8). Since Lemmas 1 and 2 holds non-asymptotically, uniformly for any $\tau \in (n_0, n-n_0)$, we have

$$\tau^{1/2}\|\widehat{\eta}_p^{\tau} - \widehat{\eta}_0 - S^{\tau}(\eta_0) + S^n(\eta_0)\| = O_P(\tau^{1/2}h^{-3/2}((\tau h)^{-1} + h^{2m})) = o_P(1),$$

$$(n-\tau)^{1/2}\|\widehat{\eta}_q^{n-\tau} - \widehat{\eta}_0 - S^{(n-\tau)}(\eta_0) + S^n(\eta_0)\|$$

$$= O_P((n-\tau)^{1/2}h^{-3/2}(((n-\tau)h)^{-1} + h^{2m})) = o_P(1).$$

where $S^{\tau}(\eta_0) = -\frac{1}{\tau}\sum_{i=1}^{\tau} K_{\boldsymbol{X}_i} + E_{\eta_0}K_{\boldsymbol{X}} + W_\lambda\eta_0$, $S^{n-\tau}(\eta_0) = -\frac{1}{n-\tau}\sum_{i=1}^{n-\tau} K_{\boldsymbol{X}_i} + E_{\eta_0}K_{\boldsymbol{X}} + W_\lambda\eta_0$, and $S^n(\eta^*) = -\frac{1}{n}\sum_{i=1}^{n} K_{\boldsymbol{X}_i} + E_{\eta_0}K_{\boldsymbol{X}} + W_\lambda\eta_0$. Thus we only need to focus on $n^{1/2}\|S^{\tau}(\eta_0) - S^n(\eta_0)\|$ and $n^{1/2}\|S^{n-\tau}(\eta_0) - S^n(\eta_0)\|$. Then, the dominate term in (S.8) could be written as,

$$\begin{aligned}
&\tau\|\widehat{\eta}^{\tau} - \widehat{\eta}_0\|^2 + (n-\tau)\|\widehat{\eta}^{n-\tau} - \widehat{\eta}_0\| \\
=&\tau\|(\frac{1}{n} - \frac{1}{\tau})\sum_{i=1}^{\tau} K_{X_i} + \frac{1}{n}K_{X_i}\|^2 + (n-\tau)\|\frac{1}{n}\sum_{i=1}^{\tau} K_{X_i} + (\frac{1}{n} - \frac{1}{n-\tau})\sum_{i=\tau_1}^{n} K_{X_i}\|^2 \\
=&\frac{n-\tau}{n\tau}\|\sum_{i=1}^{\tau} K_{X_i}\|^2 + \frac{\tau}{n(n-\tau)}\|\sum_{i=\tau+1}^{n} K_{X_i}\| - \frac{2}{n}\left\langle \sum_{i=1}^{\tau} K_{X_i}, \sum_{i=\tau+1}^{n} K_{X_i}\right\rangle \\
=&\frac{1}{n}\|\sqrt{\frac{n-\tau}{\tau}}\sum_{i=1}^{\tau} K_{X_i} - \frac{\tau}{n-\tau}\sum_{i=\tau+1}^{n} K_{X_i}\|^2 \\
=&\frac{1}{n}\|\sum_{i=1}^{n} c_i^{\tau} K_{X_i}\|^2.
\end{aligned} \tag{S.9}$$

where $c_i^{\tau} = \sqrt{(n-\tau)/\tau}$ for $1 \le i \le \tau$ and $c_i^{\tau} = -\sqrt{\tau/(n-\tau)}$ for $\tau + 1 \le i \le n$.

Recall, our test statistics is constructed by the sum of $PLR^{\tau}$ for $\tau = n_0, \ldots, n-n_0$, i.e.

$$CPLR = \sum_{\tau=n_0}^{n-n_0} PLR^{\tau} = \sum_{\tau=n_0}^{n-n_0} \frac{1}{n}\|\sum_{i=1}^{n} c_i^{\tau} K_{X_i}\|^2$$

Recall that $K_{X_i} = \sum_{\nu=1}^{\infty} \frac{\phi_\nu(X_i)}{1+\lambda\rho_\nu}\phi_\nu$. Let $Y_i^p = \frac{\phi_\nu(X_i)}{\sqrt{1+\lambda\rho_\nu}}$, and $\boldsymbol{Y}_i = \{Y_i^{\nu}\}_{\nu=1}^{\infty}$. So we have $K(X_i) =$

$\sum_{\nu=1}^{\infty} \frac{Y_i^{\nu}\phi_\nu}{\sqrt{1+\lambda\rho_\nu}}$. Then we could rewrite $\|\sum_{i=1}^n K_{X_i}\|^2$ as,

$$\begin{aligned}\|\sum_{i=1}^n K_{X_i}\|^2 &= \|\sum_{i=1}^n \sum_{\nu=1}^{\infty} \frac{Y_i^{\nu}\phi_\nu}{\sqrt{1+\lambda\rho_\nu}}\| = ||\sum_{\nu=1}^{\infty} \frac{\sum_{i=1}^n Y_i^{\nu}}{\sqrt{1+\lambda\rho_\nu}}\phi_\nu||^2 \\ &= \sum_{\nu=1}^{\infty} |\frac{\sum_{i=1}^n Y_i^{\nu}}{\sqrt{1+\lambda\rho_\nu}}|^2(1+\lambda\rho_\nu) = \sum_{\nu=1}^{\infty}|\sum_{i=1}^n Y_i^{\nu}|^2 = \|\sum_{i=1}^n \mathbf{Y}_i\|_2^2\end{aligned}$$

Also, we have

$$L_\gamma = \sum_{i=1}^n E\|x_i\|_2^\gamma \le \sum_{i=1}^n C_\phi^\gamma h^{-\gamma/2} = C_\phi^\gamma n h^{-\gamma/2}$$

Then, we can construct $\boldsymbol{Z}_i = \{Z_i^\nu\}_{\nu=1}^{\infty}$ with $cov(\boldsymbol{Z}_i) = cov(\mathbf{Y}_i)$ and $EY_i = 0$ such that

$$E(\max_{1\le l\le n} \|\sum_{i=1}^l \mathbf{Y}_i - \sum_{i=1}^l \boldsymbol{Z}_i\|_2)^\gamma) \le L_\gamma.$$

Thus $\max_{1\le l\le n}\|\sum_{i=1}^l \mathbf{Y}_i - \sum_{i=1}^l \boldsymbol{Z}_i\|_2 \le O_p(L_\gamma^{1/\gamma}) = O_p(n^{1/\gamma}h^{-1/2})$. We define $\zeta_i = \sum_{\nu=1}^{\infty}\frac{Z_i^\nu\phi_\nu}{\sqrt{1+\lambda\rho_\nu}}$. By Götze and Zaitsev [2011], we have

$$\|\sum_{i=1}^l K_{X_i} - \sum_{i=1}^l \zeta_i\|_2 = \|\sum_{i=1}^l X_i - \sum_{i=1}^l Y_i\|_2 = O_p(n^{1/\gamma}h^{-1/2})$$

Thus, our test statistics has the same distribution with

$$CPLR \overset{d}{\sim} T = \sum_{\nu=1}^{\infty} \frac{\sum_{\tau=n_0}^{n-n_0}(\sum_{i=1}^n c_i^\tau \zeta_\tau)^2}{1+\lambda\rho_\nu} = \sum_{\nu=1}^{\infty}\frac{\xi_\nu}{1+\lambda\rho_\nu}$$

where $\xi_\nu \equiv \sum_{\tau=n_0}^{n-n_0}(\sum_{i=1}^n c_i^\tau\zeta_\tau)^2$ with $\mu_\xi = \mathrm{tr}(\Delta)$ and $\sigma_\xi^2 = 2\,\mathrm{tr}(\Delta)$, $\Delta = \sum_{\tau=n_0}^{n-n_0}\boldsymbol{c}^\tau(\boldsymbol{c}^\tau)^T$, and $\boldsymbol{c}^\tau = \{c_i^\tau\}_{i=1}^n$.

By using the characteritic function, we will show that the standarized version $T$, $(T-\mathrm{tr}(\Delta)h^{-1})/2\,\mathrm{tr}(\Delta)h^{-1}$ asymptotically follows Gaussian distribution. Let $\Delta = P^T\Lambda P$ where $\Lambda = diag(\lambda_1,\ldots,\lambda_n)$. The characteristic function of standanderized $T$ is

$$\begin{aligned}Ee^{ix\frac{T-\mu_T}{\sigma_T}} &= e^{-ix\frac{\mu_T}{\sigma_T}}Ee^{ix\frac{T}{\sigma_T}} \\ &= e^{-ix\frac{\mu_T}{\sigma_T}}\prod_{\nu=1}^{\infty}[\det(I-\frac{2ix}{\sigma_T(1+\lambda\rho_\nu\Delta)})]^{-1/2} \\ &= e^{-ix\frac{\mu_T}{\sigma_T}}\prod_{\nu=1}^{\infty}\prod_{l=1}^{n}(1-\frac{2ix}{\sigma_T(1+\lambda\rho_\nu)}\lambda_l)^{-1/2}\end{aligned}$$

Then we take the log-transform of characteristic function as,

$$\log Ee^{ix\frac{T-\mu_T}{\sigma_T}} = -ix\frac{\mu_T}{\sigma_T} - \frac{1}{2}\sum_{\nu=1}^{\infty}\sum_{l=1}^{n}\log(1-\frac{2ix}{\sigma_T(1+\lambda\rho_\nu)}\lambda_l)$$
$$= -ix\frac{\mu_T}{\sigma_T} - \frac{1}{2}\sum_{\nu=1}^{\infty}\sum_{l=1}^{n}\sum_{k=1}^{\infty}-\frac{1}{k}(\frac{2ix}{\sigma_T(1+\lambda\rho_\nu)})^k \tag{S.10}$$

Note that, for $k=1,2$, we have

$$\sum_{\nu=1}^{\infty}-(\frac{2ix}{\sigma_T(1+\lambda\rho_\nu)}) = \frac{2ix\mu_T}{\sigma_T}, \tag{S.11}$$

$$\sum_{\nu=1}^{\infty}-\frac{1}{2}(\frac{2ix}{\sigma_T(1+\lambda\rho_\nu)})^2 = \frac{x^2}{2}. \tag{S.12}$$

For $k>3$, we have

$$|\sum_{\nu=1}^{\infty}\sum_{l=1}^{n}-\frac{1}{k}(\frac{2ix}{\sigma_T(1+\lambda\rho_\nu)})^k| = |\frac{(2ix)^k}{k\sigma_T^k}\sum_{\nu=1}^{\infty}\sum_{l=1}^{n}\frac{\lambda_l^k}{1+\lambda\rho_\nu}| \le \frac{(2x)^k}{k}h^{(k-2)/2}$$

where the second inequatility is due to that $\sigma_T^k \asymp (2\sum_{\nu=1}^{\infty}\frac{1}{(1+\lambda\rho_\nu)^2})^{k/2}(\sum_{l=1}^{n}\lambda_l^2)^{k/2} \asymp h^{-k/2}(\sum_{l=1}^{n}\lambda_l^2)^{-k/2} \ge h^{-k/2}\sum_{l=1}^{n}\lambda_l^k$. Thus we have,

$$|\sum_{\nu=1}^{\infty}\sum_{l=1}^{n}\sum_{k=3}^{\infty}-\frac{1}{k}(\frac{2ix}{\sigma_T(1+\lambda\rho_\nu)})^k| \le \sum_{k=3}^{\infty}\frac{(2x)^k}{k}h^{(k-2)/2} = (2x\sqrt{h})^3h^{-1}\frac{1}{1-2x\sqrt{h}} \to 0. \tag{S.13}$$

By pluging (S.11), (S.12) and (S.13) into (S.10), we have

$$\frac{T-\mu_T}{\sigma_T} \xrightarrow{d} N(0,1).$$

### S.4.1 Proof of Theorem 2

Without loss of generality, we assume $E(\eta_p - \eta_q) = c > 0$. When $n_0 \le \tau \le \tau_0$, the total likelihood functional are written as,

$$\begin{aligned}
&\ell^{\tau}(\widehat{\eta}_p^{\tau}, \widehat{\eta}_q^{n-\tau}) = \ell^{\tau}(\widehat{\eta}_p^{\tau}) + \ell^{n-\tau}(\widehat{\eta}_q^{n-\tau}) \\
&= -\sum_{i=1}^{\tau_0}(\widehat{\eta}_p^{\tau}(\boldsymbol{x}_i) + 1) + \tau_0 \frac{\lambda}{2} J(\widehat{\eta}_p^{\tau}) - \sum_{i=\tau_0+1}^{n} (\widehat{\eta}_q^{n-\tau}(\boldsymbol{x}_i) - 1) + (n - \tau_0)\frac{\lambda}{2} J(\widehat{\eta}_q^{n-\tau}) \\
&+ \sum_{i=\tau+1}^{\tau_0} (\widehat{\eta}_q^{n-\tau}(\boldsymbol{x}_i) - 1) - (\tau_0 - \tau)\frac{\lambda}{2} J(\widehat{\eta}_p^{\tau}) - \sum_{i=\tau+1}^{\tau_0} (\widehat{\eta}_q^{n-\tau}(\boldsymbol{x}_i) - 1) + (\tau_0 - \tau)\frac{\lambda}{2} J(\widehat{\eta}_q^{n-\tau}) \\
&= \ell^{\tau_0}(\widehat{\eta}_p^{\tau}) + \ell^{n-\tau_0}(\widehat{\eta}_q^{\tau}) + \sum_{i=\tau+1}^{\tau_0} (\widehat{\eta}_p^{\tau}(\boldsymbol{x}_i) - \widehat{\eta}_q^{n-\tau}(\boldsymbol{x}_i)) - (\tau_0 - \tau)\frac{\lambda}{2}(J(\widehat{\eta}_p^{\tau}) - J(\widehat{\eta}_q^{n-\tau}))
\end{aligned}$$

Thus we have,

$$\ell^{\tau}(\widehat{\eta}_p^{\tau}, \widehat{\eta}_q^{n-\tau}) - \ell^{\tau_0}(\widehat{\eta}_p^{\tau_0}, \widehat{\eta}_q^{\tau_0})$$

$$= \frac{\tau_0}{2}\|\widehat{\eta}_p^{\tau} - \widehat{\eta}_p^{\tau_0}\|^2 + \frac{n-\tau_0}{2}\|\widehat{\eta}_q^{n-\tau} - \widehat{\eta}_q^{\tau_0}\|^2 \tag{S.14}$$

$$+ \sum_{i=\tau+1}^{\tau_0} (\widehat{\eta}_p^{\tau}(\boldsymbol{x}_i) - \widehat{\eta}_q^{n-\tau}(\boldsymbol{x}_i)) - (\tau_0 - \tau)\frac{\lambda}{2}(J(\widehat{\eta}_p^{\tau}) - J(\widehat{\eta}_q^{n-\tau})) \tag{S.15}$$

We will consider the four terms in (S.15) seperatly. We first define $\eta^{\tau_0}$ as,

$$\eta^{\tau_0} = \log(\frac{\tau_0 - \tau}{n - \tau} e^{\eta_p^0(x)} + \frac{n - \tau_0}{n - \tau} e^{\eta_q^0(x)})$$

By Lemma 1, we have

$$\|\widehat{\eta}_p^{\tau} - \widehat{\eta}_p^{\tau_0}\|^2 = \|\widehat{\eta}_p^{\tau} - \eta_p + \eta_p - \widehat{\eta}_p^{\tau_0}\|^2 = O((nh)^{-1} + \lambda), \tag{S.16}$$

and

$$\|\widehat{\eta}_q^{n-\tau} - \widehat{\eta}_q^{\tau_0}\|^2 = \|\widehat{\eta}_q^{n-\tau} - \eta_q^{n-\tau} + \eta_q^{n-\tau} - \eta_q + \eta_q - \widehat{\eta}_q^{\tau_0}\|^2 = \|\eta_q - \eta_q^{n-\tau}\|^2 + O((nh)^{-1} + \lambda),$$

The third term is

$$\begin{aligned}
&\sum_{i=\tau+1}^{\tau_0} (\widehat{\eta}_p^{\tau}(\boldsymbol{x}_i) - \widehat{\eta}_q^{n-\tau}(\boldsymbol{x}_i)) \\
=&\sum_{i=\tau+1}^{\tau_0} (\widehat{\eta}_p^{\tau}(\boldsymbol{x}_i) - \eta_p^{\tau}(\boldsymbol{x}_i) - \eta_q^{n-\tau}(\boldsymbol{x}_i) - \widehat{\eta}_q^{\tau}(\boldsymbol{x}_i)) + \sum_{i=\tau_0+1}^{\tau} (\eta_p^{\tau}(\boldsymbol{x}_i) - \eta_q^{n-\tau}(\boldsymbol{x}_i)) \\
=&\sum_{i=\tau_0+1}^{\tau} (\eta_p^{\tau}(\boldsymbol{x}_i) - \eta_q^{n-\tau}(\boldsymbol{x}_i)) + (\tau - \tau_0) O_p(h^m + \sqrt{\log(n)/(n-\tau)h})
\end{aligned} \tag{S.17}$$

The forth term of (S.15) is

$$(\tau_0 - \tau)\frac{\lambda}{2}(J(\widehat{\eta}_p^{\tau}) - J(\widehat{\eta}_q^{n-\tau})) = (\tau_0 - \tau) O_p(\lambda) \tag{S.18}$$

Through plugging (S.16), (S.17), (S.17), and ($S.18$) into (S.15), we have

$$\begin{aligned}
&\ell^{\tau}(\widehat{\eta}_p^{\tau}, \widehat{\eta}_q^{n-\tau}) - \ell^{\tau_0}(\widehat{\eta}_p^{\tau_0}, \widehat{\eta}_q^{\tau_0}) \\
=&\frac{\tau - \tau_0}{2}\|\eta_q - \eta_q^{n-\tau}\|^2 + \sum_{i=\tau_0+1}^{\tau} (\eta_p^{\tau}(\boldsymbol{x}_i) - \eta_q^{n-\tau}(\boldsymbol{x}_i)) + O_p((h)^{-1} + n\lambda)
\end{aligned} \tag{S.19}$$

By Hoeffding's inequality, we have

$$\mathrm{P}(\frac{1}{\tau - \tau_0}\sum_{i=\tau_0+1}^{\tau} (\eta_p^{\tau}(\boldsymbol{x}_i) - \eta_q^{n-\tau}(\boldsymbol{x}_i)) - \mathbb{E}[\eta_p^{\tau}(\boldsymbol{x}_i) - \eta_q^{n-\tau}] \geq t) \leq e^{-2(\tau-\tau_0)t^2}$$

Where $t \leq 0$. In order to proof $\ell^{\tau}(\widehat{\eta}_p^{\tau}, \widehat{\eta}_q^{n-\tau}) - \ell^{\tau_0}(\widehat{\eta}_p^{\tau_0}, \widehat{\eta}_q^{\tau_0}) > 0$, we need to show that $\frac{\tau-\tau_0}{2}\|\eta^{\tau_0} - \eta_q^0\|^2 - (\tau - \tau_0)\mathbb{E}[\eta_p^0(\boldsymbol{x}_i) - \eta^{\tau_0}] > 0$. We could further rewrite the main term of ($S.19$) as

$$\begin{aligned}
&\frac{n - \tau_0}{2}\|\eta_q^{n-\tau} - \eta_q^{\tau_0}\|^2 + (\tau_0 - \tau)\mathbb{E}_{\eta_p}(\eta_p^{\tau}(x) - \eta_q^{n-\tau}(x)) \\
=&\frac{n - \tau}{2}[(1-\theta)\mathbb{E}_{\eta_p^0}[|\log(\theta E^{\Delta(x)} + 1 - \theta)|^2] + 2\theta\mathbb{E}_{\eta_p^0}(\Delta(x)) - 2\theta\mathbb{E}_{\eta_p}(\log(\theta e^{\Delta(x)} + 1 - \theta))] \\
\geq&(1-\theta)T^2(\tau) + 2\theta\mathbb{E}_{\eta_q^0}(\Delta(x)) - 2\theta T(\tau) \\
=&(T(\tau) - \frac{\theta}{1-\theta})^2 - (\frac{\theta}{1-\theta})^2 + \frac{2\theta}{1-\theta}E_{\eta_p}(\Delta(\boldsymbol{x})) \\
\geq&(2E_{\eta_q^0}[\Delta(x)] - \frac{\theta}{1-\theta})\frac{\theta}{1-\theta}
\end{aligned} \tag{S.20}$$

where $T^2(\tau) = E_{\eta_p^0}[|\log(\theta E^{\Delta(x)} + 1 - \theta)|^2]$ and $\theta = \frac{\tau_0 - \tau}{n - \tau}$. Since we have $E_{\eta_p^0}[\Delta(x)] \geq c \geq \frac{\theta}{n-\theta}$, we have (S.20) larger than zero. Thus, if $\|\tau - \tau_0\| > c'(h^{-1} + n\lambda)$, we have $\frac{\tau_0}{2}\|\eta_p - \eta_p^{\tau}\|^2 > 0$.

For $\tau_0 < \tau < n$, following the similar derivation, we have the difference of the log-likelihood functionals at $\tau$ and $\tau_0$ expanded as,

$$\begin{aligned}
&\ell^{\tau}(\widehat{\eta}_p^{\tau}, \widehat{\eta}_q^{n-\tau}) - \ell^{\tau_0}(\widehat{\eta}_p^{\tau_0}, \widehat{\eta}_q^{\tau_0}) \\
=&\frac{\tau_0}{2}\|\widehat{\eta}_p^{\tau} - \widehat{\eta}_p^{\tau_0}\|^2 + \frac{n-\tau_0}{2}\|\widehat{\eta}_q^{n-\tau} - \widehat{\eta}_q^{\tau_0}\|^2 + \sum_{i=\tau+1}^{\tau_0}(\widehat{\eta}_q^{n-\tau}(\boldsymbol{x}_i) - \widehat{\eta}_p^{\tau}(\boldsymbol{x}_i)) - (\tau_0 - \tau)\frac{\lambda}{2}(J(\widehat{\eta}_p^{\tau}) - J(\widehat{\eta}_q^{n-\tau})) \\
=&\frac{\tau_0}{2}\|\eta_p - \eta_p^{\tau}\|^2 + (\tau - \tau_0)\mathbb{E}[\eta_q^{n-\tau}(\boldsymbol{x}_i) - \eta_p^{\tau}(\boldsymbol{x}_i)] + O_p((h)^{-1} + n\lambda) \qquad \text{(S.21)}
\end{aligned}$$

The first term of (S.21) can be written as

$$\frac{\tau_0}{2}\|\eta_p - \eta_p^{\tau}\|^2 = \frac{\tau_0}{2}\mathbb{E}[\log(\frac{\tau_0}{\tau} + \frac{\tau - \tau_0}{\tau}e^{-\Delta(x)})]$$

The second term of (S.21) can be written as

$$\sum_{i=\tau_0+1}^{\tau}(\eta_q^{n-\tau}(\boldsymbol{x}_i) - \eta_p^{\tau}(\boldsymbol{x}_i)) = -(\tau - \tau_0)\mathbb{E}[\log(\frac{\tau_0}{\tau}e^{\Delta(\boldsymbol{x})} + \frac{\tau - \tau_0}{\tau})]$$

Since $(\tau - \tau_0)/\tau$ goes to 0, we have the first term dominate the second term. Thus, if $\|\tau - \tau_0\| > c'(h^{-1} + n\lambda)$, we have $\frac{\tau_0}{2}\|\eta_p - \eta_p^{\tau}\|^2 > 0$.

## S.4.2 Proof for Lemmas 1 and 2

*Proof.* Without loss of generality, we only consider the situation when $n_0 < \tau < \tau_0 < n - n_0$ and prove the probability bound of $\|\widehat{\eta}_q^{n-\tau} - \eta_q^{n-\tau}\|$. We define the averaged first order Fréchet derivative as

$$S_{\tau}(\eta)\Delta\eta = -\frac{1}{n-\tau}\sum_{i=\tau+1}^{n}\Delta\eta(\boldsymbol{x}_i) + \mathbb{E}(\Delta\eta) + \lambda J(\Delta\eta, \eta)$$

The expectation of $S_\tau(\eta)$ is

$$
\begin{aligned}
&\mathbb{E}S_\tau(\eta)\Delta\eta\\
=&-\frac{1}{n-\tau}\sum_{i=\tau+1}^{n}\mathbb{E}\Delta\eta(\boldsymbol{x}_i)+\int_{\mathcal{X}} e^{\eta(\boldsymbol{x})}\Delta\eta(\boldsymbol{x})+\lambda J(\Delta\eta,\eta)\\
=&-\frac{1}{n-\tau}(\sum_{i=\tau+1}^{n}\int_{\mathcal{X}}\Delta\eta(\boldsymbol{x})e^{\eta_p(\boldsymbol{x})}d\boldsymbol{x}-\sum_{i=\tau_0+1}^{n}\int_{\mathcal{X}}\Delta\eta(\boldsymbol{x})e^{\eta_q(\boldsymbol{x})}d\boldsymbol{x})\\
&+\int_{\mathcal{X}} e^{\eta(x)}\Delta\eta(\boldsymbol{x})d\boldsymbol{x}+\lambda J(\Delta\eta,\eta)\\
=&\int_{\mathcal{X}}\Delta\eta(\boldsymbol{x})(e^{\eta(\boldsymbol{x})}-\frac{\tau_0-\tau}{n-\tau}e^{\eta_p(\boldsymbol{x})}-\frac{n-\tau_0}{n-\tau}e^{\eta_q(\boldsymbol{x})})d\boldsymbol{x}+\lambda J(\Delta\eta,\eta)
\end{aligned}
$$

Let $\eta_q^{n-\tau}=\frac{\tau_0-\tau}{n-\tau}e^{\eta_p(\boldsymbol{x})}+\frac{n-\tau_0}{n-\tau}e^{\eta_q(\boldsymbol{x})}$, and $T_1(\eta)=-\eta+\mathbb{E}S_\tau(\eta_q^{n-\tau}+\eta)$. Let $\mathbb{B}(r_{1n})=\{g\in\mathcal{H}:\|g\|\le r_{1n}\}$ be the $r_{1n}$-ball. For any $g\in\mathbb{B}(r_{1n})$, using $DS_\tau(\eta_q^{n-\tau})=id$ and $\|\eta\|_{\sup}\le c_K h^{-1/2}r_{1n}=2c_K h^{m-1/2}\le 1$, it is easy to see that

$$
\begin{aligned}
\|T_1(\eta)\|=&\|\eta-\mathbb{E}S_\tau(\eta_q^{n-\tau}+\eta)+\mathbb{E}S_\tau(\eta_q^{n-\tau})-\mathbb{E}S_\tau(\eta_q^{n-\tau})\|\\
=&\|\eta-\mathbb{E}S_\tau(\eta_q^{n-\tau}+\eta)+\mathbb{E}S_\tau(\eta_q^{n-\tau})\|+\|\mathbb{E}S_\tau(\eta_q^{n-\tau})\|\\
=&\|\eta-DS_\tau(\eta_q^{n-\tau})+\int_{\mathcal{X}}\int_{\mathcal{X}}\boldsymbol{s}D^2S_\tau(\eta_q^{n-\tau}+\boldsymbol{s}\boldsymbol{s}'\eta)\eta\eta d\boldsymbol{s}d\boldsymbol{s}'\|+\|S_\tau(\eta_q^{n-\tau})\|\\
=&\|\int_{\mathcal{X}}\int_{\mathcal{X}}s\int_{\mathcal{X}}e^{\eta_q^{n-\tau}(\boldsymbol{x})+ss'\eta(\boldsymbol{x})}K_{\boldsymbol{x}}\eta(\boldsymbol{x})^2d\boldsymbol{x}d\boldsymbol{s}d\boldsymbol{s}'\|+r_n/2\\
=&c_K h^{-1/2}\|g\|^2+r_n/2
\end{aligned}
$$

Therefore, $T_1$ maps $B(r_n)$ to itself. Next we will show that $T_1$ is an contraction mapping. Let $g_1$, $g_2\in B(r_n)$, and $g=g_1-g_2$. Note that for any $0\le s\le 1$, $\|g_2+\boldsymbol{s}g\|_{\sup}\le 2c_K h^{-1/2}r_n=$

$4c_K h^{-1/2}h^m \leq 2/3$. Therefore, we have

$$
\begin{aligned}
&\|T_1(g_1) - T_1(g_2)\| \\
=&\|g_1 - g_2 + S_\tau(\eta_q^{n-\tau} + g_1) - S_\tau(\eta_q^{n-\tau} + g_2)\| \\
=&\|g_1 - g_2 + \int_{\mathcal{X}} DS_\tau(\eta_q^{n-\tau} + g_2 + \boldsymbol{s}g)d\boldsymbol{s}\| \\
=&\| \int_{\mathcal{X}} [DS_\tau(\eta_q^{n-\tau} + g_2 + sg) - DS_\tau(\eta_q^{n-\tau})]gd\boldsymbol{s}\| \\
=&\| \int_{\mathcal{X}} \int_{\mathcal{X}} D^2 S_\tau(\eta_q^{n-\tau} + \boldsymbol{s}'(g_2 + sg))(g_2 + sg)gd\boldsymbol{s}d\boldsymbol{s}'\| \\
=&\| \int_{\mathcal{X}} \int_{\mathcal{X}} \int_{\mathcal{X}} e^{\eta_q^{n-\tau}(\boldsymbol{x}) + \boldsymbol{s}'(g_2(\boldsymbol{x}) + \boldsymbol{s}g(\boldsymbol{x}))}(g_2(\boldsymbol{x}) + \boldsymbol{s}g(\boldsymbol{x}))g(\boldsymbol{x})K_{\boldsymbol{x}}d\boldsymbol{x}d\boldsymbol{s}d\boldsymbol{s}'\| \\
\leq& c_K h^{-1/2} \int_{\mathcal{X}} \|g_2 + \boldsymbol{s}g\|d\boldsymbol{s}\|g\| \\
=&4c_K h^{m-1/2}\|g\|
\end{aligned}
$$

Since $4c_K h^{m-1/2} < 1$, this shows $T_1$ is is a contraction mapping. By , $T_1$ has a unique fixed point $g' \in \mathbb{B}(r_n)$ satifying $T_1(g') = g'$. Let $\eta' = \eta_q^{n-\tau} + g'$. Then $S_\tau(\eta') = 0$ and $\|\eta' - \eta_q^{n-\tau}\| \leq r_n$.

Then we show that the operator $\|DS_\tau(\eta')\|$ is inveratable. For any $g_1, g_2 \in \mathcal{H}$, we have

$$
|[DS_\tau(\eta') - DS_\tau(\eta_q^{n-\tau})]g_1 g_2| = |\int_{\mathcal{X}} (e^{\eta'(\boldsymbol{x})} - e^{\eta_q^{n-\tau}(\boldsymbol{x})})g_1(\boldsymbol{x})g_2(\boldsymbol{x}) \leq \|g_1\|\|g_2\|/2.
$$

Togeter with the fact that $DS_\tau = id$, we get that the operator norm satisfies $\|DS_\tau(\eta') - id\| \leq 1/2$. This implies that $DS_\tau(\eta')$ is inveratable. Define an operator $T_2(\eta)$, $\eta \in \mathcal{H}$ and rewrite it as

$$
\begin{aligned}
T_2(\eta) =& \eta - [DS_\tau(\eta')]^{-1} S_\tau(\eta' + \eta) \\
=& - DS_\tau(\eta')^{-1}[S_\tau(\eta' + \eta) - S_\tau(\eta') - DS_\tau(\eta')] \\
&- DS_\tau(\eta')^{-1} S_\tau(\eta')
\end{aligned}
$$

For $i = 1, \ldots, n$, let $R_i = K_{\boldsymbol{X}_i} - \mathbb{E}_{\eta'}[K_{\boldsymbol{X}}]$. By direct calculation, we have

$$
\begin{aligned}
\|\mathbb{E}_{\eta'}[K_X]\| &= \sup_{\|g\|=1} \langle \mathbb{E}_{\eta'}[K_X], g\rangle = \sup_{\|g\|=1} |\mathbb{E}_{\eta'} g(\boldsymbol{X})| \\
&= |\int_{\mathcal{X}} e^{\eta' - \eta_q^{n-\tau} + \eta_{\eta_0}} g(X)| \leq \|\eta' - \eta_q^{n-\tau}\| \leq r_n
\end{aligned}
$$

Therefore, we have $\|R_i\| \le h^{-1/2} + r_n$. By using Cauchy-Schwartz inequality,

$$\mathbb{E}[\exp(\frac{\|R_i\|}{2h^{-1/2}})] \le C_1 \exp(2)$$

Let $\delta = hr_n$. Recall that $h^{1/2} r_n \le 1$ which implies $\delta \le h^{1/2}$. Therefore, $\mathbb{E}[\exp(\delta\|R_i\|)] \le C_1 \exp(2)$. Moreover, for $x \ge 0$ and any constant $c > 0$, $\exp(M) - 1 - M \le M^2 \exp(M)$ and $M^{-2}\exp(cx) \ge c^2 \exp(2)/4$. Let $c = (2h^{-1/2})^{-1} - \delta$ and $c \ge (4h^{-1/2})^{-1}$. So, we have

$$\mathbb{E}[\exp(\delta\|R_i\|) - 1 - \delta\|R_i\|] \le \frac{4\delta^2 \exp(-2)}{c^2} \mathbb{E}[\exp(\delta\|R_i\|)] \le 64 C_1 h^{-1} \delta^2$$

It follows by Theorem in Pinelis [1994] that, for $L(M) = 4(4C_1 + M)$,

$$\mathrm{P}(\|\sum_{i=1}^{n} R_i\| \ge L(M) n r_n) \le 2\exp(-L(M)\delta n r_n + 64 C_1 h^{-1} \delta^2) = 2\exp(-M n h r_n^2)$$

We note that the right-hand side in the above inequality does not depend on $\eta_q^{n-\tau}$. Then we consider the three terms in (S.4.2). Let $\mathcal{E}_1 = \{\|S_\tau(\eta')\| \le L(M) r_n\}$, then $\sup_{\eta\in\mathcal{H}} \mathrm{P}(\mathcal{E}_1^c) \le 2\exp(-M n h r_n^2)$

$$\begin{aligned}
&\|S_\tau(\eta' + \eta) - S_\tau(\eta') - DS_\tau(\eta')\eta\| \\
=&\|\int_{\mathcal{X}}\int_{\mathcal{X}} D^2 S_\tau(\eta' + \boldsymbol{s}'\boldsymbol{s}\eta)\eta^2 d\boldsymbol{s} d\boldsymbol{s}'\| \\
=&\int_{\mathcal{X}}\int_{\mathcal{X}} \|\int_{\mathcal{X}} e^{\eta'(\boldsymbol{x}) + \boldsymbol{s}'\boldsymbol{s}\eta(\boldsymbol{x})} \eta(\boldsymbol{x})\eta(\boldsymbol{x}) K_{\boldsymbol{x}} d\boldsymbol{x}\| \\
\le& \sup_{\boldsymbol{x}} \{e^{\eta'(\boldsymbol{x}) - \eta_q^{n-\tau} + \boldsymbol{s}'\boldsymbol{s}\eta(\boldsymbol{x})}\} \|\eta\|_{\sup} \|\eta\| \|K_{\boldsymbol{x}}\| \\
\le& c_K h^{-1/2} r_n (h^{-1/2})^2 \|\eta\| \le \|\eta\|/2
\end{aligned} \tag{S.22}$$

It follows by (S.22) on $\mathcal{E}_1$, for any $\eta \in \mathbb{B}(2L(M)r_n)$, $\|T_2(\eta)\| \le (\|\eta\|/2 + L(M)r_n) = 2L(M)r_n$, meanwhile, by replacing $\eta$ by $\eta_1 - \eta_2$ in (S.22), we get $\|T_2(\eta_1) - T_2(\eta_2)\| \le L(M) r_n$. Therefore, for any $\eta \in \mathcal{H}$, on $\mathcal{E}_1$, $T_2$ is a contraction mapping from $\mathbb{B}(2L(M)r_n)$ to itself.itself. By contraction mapping theorem, there exists uniquely an element $\eta''$ on $\mathbb{B}(2L(M)r_n)$ such that $T_2(\eta'') = \eta''$. Let $\widehat{\eta}_q^{n-\tau} = \eta' + \eta''$. Then on $\mathcal{E}_1$, $\|\widehat{\eta}_q^{n-\tau} - \eta_q^{n-\tau}\| \le \|\widehat{\eta}_q^{n-\tau} - \eta'\| + \|\eta' - eta_{\tau_0}\| \le r_n + 2L(M) r_n$. □

### S.4.3 Proof for Lemmas S.1 and S.2

*Proof.* Let $\widehat{\eta}_q^{n-\tau}$ be the penalized MLE of $\eta_q^{n-\tau}$, and $g = \eta_q^{n-\tau} - \widehat{\eta}_q^{n-\tau}$. We have the functional expansion of $S_\tau(\eta_q^{n-\tau} + g)$ as

$$S_\tau(\eta_q^{n-\tau} - g) = S_\tau(\eta_q^{n-\tau}) - DS_\tau(\eta_q^{n-\tau})g + \int_{\mathcal{X}}\int_{\mathcal{X}} sD^2S_\tau(f - \boldsymbol{ss}'g)ggd\boldsymbol{ss}'$$

Since $S_\tau(\eta_q^{n-\tau}) = 0$ and $DS_\tau(\eta_q^{n-\tau})g = id$, we have

$$\begin{aligned}\|S_\tau(\eta_q^{n-\tau}) - (\widehat{\eta}_q^{n-\tau} - \eta_q^{n-\tau})\| =& \|\int_{\mathcal{X}}\int_{\mathcal{X}} sD^2S_\tau(\eta_q^{n-\tau} - \boldsymbol{ss}'g)ggd\boldsymbol{ss}'\| \\ &\leq \int_{\mathcal{X}}\int_{\mathcal{X}} \|\int_{\mathcal{X}} g(\boldsymbol{x})^2 K_{\boldsymbol{x}} e^{\eta_q^{n-\tau} - \boldsymbol{ss}'g} d\boldsymbol{x}\| d\boldsymbol{ss}' \\ &\leq Cc_k h^{-1/2}\|g\|^2 < Cc_k h^{-1/2}(2h^m + 2L(M))\end{aligned}$$

Since $\sup_{\eta\in\mathcal{H}} \mathrm{P}(\mathcal{E}_1^c) \leq 2\exp(-Mnhr_n^2)$, the proof is completed. □

## References


Friedrich Götze and A Yu Zaitsev. Estimates for the rate of strong approximation in hilbert space. *Siberian mathematical journal*, 52(4):628, 2011.

Michael J Katz. Fractals and the analysis of waveforms. *Computers in Biology and Medicine*, 18 (3):145–156, 1988.

Greta M Ljung and George EP Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978.

Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

Renée A Shellhaas, Taeun Chang, Tammy Tsuchida, et al. The american clinical neurophysiology society's guideline on continuous electroencephalography monitoring in neonates. *Journal of Clinical Neurophysiology*, 28(6):611–617, 2011.