Changepoints in non-Gaussian autocorrelated time series

JSM 2025, Nashville, Tennessee

Thomas J. Fisher

Joint work with Robert Lund (UC-Santa Cruz), Norou Diawara (Old Dominion), Michael Wehner (Lawrence Berkeley National Laboratory)

Miami University, Oxford, OH

03 August 2025

MIAMI UNIVERSITY

OXFORD, OH • EST. 1809

Shameless Self-Promotion

These results are published and available:

Lund, R., Fisher, T.J., Diawara, N. and Wehner, M. (2025), "Multiple Changepoint Detection for Non-Gaussian Time Series." *J. Time Ser. Anal.* (`https://doi.org/10.1111/jtsa.12833`)

Slides and code available:

`https://tjfisher19.github.io/`

Github repo: `tjfisher19/non-gaussian_changepoints`

Email: **`fishert4@miamioh.edu`**

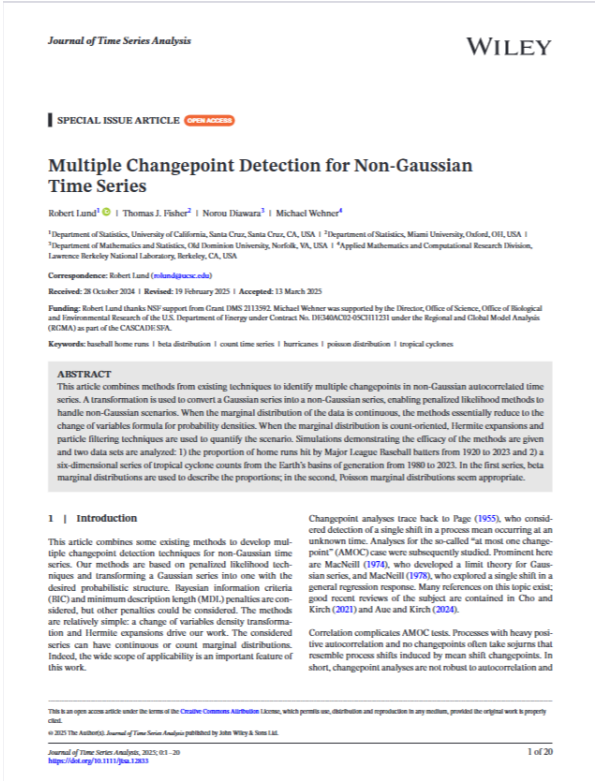
Some Highlights from the Paper

Changepoints using penalized likelihoods (model selection ideas!)

- BIC - Bayesian Information Criterion
- MDL - Minimum Description Length

Builds on Autocorrelated Gaussian Processes

- Transformed to non-Gaussian process
- Works for both continuous and discrete processes
  - Hermite expansions motivate retention of correlation structure
- Works in multivariate settings
  - Multivariate Poisson is explored in the paper



Journal of Time Series Analysis WILEY

SPECIAL ISSUE ARTICLE OPEN ACCESS

Multiple Changepoint Detection for Non-Gaussian Time Series

Robert Lund[1] | Thomas J. Fisher[2] | Norou Diawara[3] | Michael Wehner[4]

[1]Department of Statistics, University of California, Santa Cruz, Santa Cruz, CA, USA | [2]Department of Statistics, Miami University, Oxford, OH, USA | [3]Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA, USA | [4]Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Correspondence: Robert Lund (rolund@ucsc.edu)

Received: 28 October 2024 | **Revised:** 19 February 2025 | **Accepted:** 13 March 2025

Funding: Robert Lund thanks NSF support from Grant DMS 2113592. Michael Wehner was supported by the Director, Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy under Contract No. DE340AC02-05CH11231 under the Regional and Global Model Analysis (RGMA) as part of the CASCADE/SFA.

Keywords: baseball home runs | beta distribution | count time series | hurricanes | poisson distribution | tropical cyclones

ABSTRACT

This article combines methods from existing techniques to identify multiple changepoints in non-Gaussian autocorrelated time series. A transformation is used to convert a Gaussian series into a non-Gaussian series, enabling penalized likelihood methods to handle non-Gaussian scenarios. When the marginal distribution of the data is continuous, the methods essentially reduce to the change of variables formula for probability densities. When the marginal distribution is count-oriented, Hermite expansions and particle filtering techniques are used to quantify the scenario. Simulations demonstrating the efficacy of the methods are given and two data sets are analyzed: 1) the proportion of home runs hit by Major League Baseball batters from 1920 to 2023 and 2) a six-dimensional series of tropical cyclone counts from the Earth's basins of generation from 1980 to 2023. In the first series, beta marginal distributions are used to describe the proportions; in the second, Poisson marginal distributions seem appropriate.

1 | Introduction

This article combines some existing methods to develop multiple changepoint detection techniques for non-Gaussian time series. Our methods are based on penalized likelihood techniques and transforming a Gaussian series into one with the desired probabilistic structure. Bayesian information criteria (BIC) and minimum description length (MDL) penalties are considered, but other penalties could be considered. The methods are relatively simple: a change of variables density transformation and Hermite expansions drive our work. The considered series can have continuous or count marginal distributions. Indeed, the wide scope of applicability is an important feature of this work.

Changepoint analyses trace back to Page (1955), who considered detection of a single shift in a process mean occurring at an unknown time. Analyses for the so-called "at most one changepoint" (AMOC) case were subsequently studied. Prominent here are MacNeill (1974), who developed a limit theory for Gaussian series, and MacNeill (1978), who explored a single shift in a general regression response. Many references on this topic exist; good recent reviews of the subject are contained in Cho and Kirch (2021) and Aue and Kirch (2024).

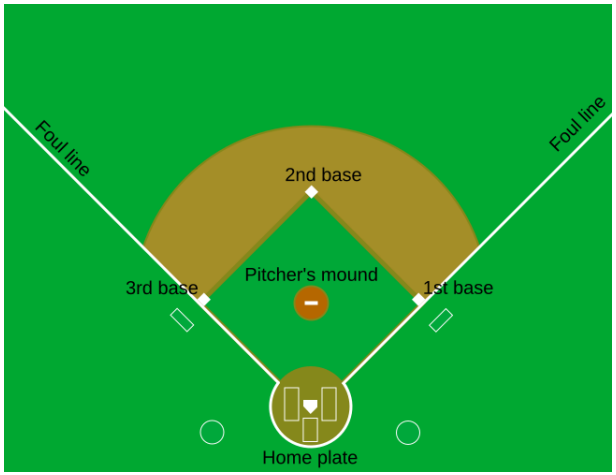
Correlation complicates AMOC tests. Processes with heavy positive autocorrelation and no changepoints often take sojurns that resemble process shifts induced by mean shift changepoints. In short, changepoint analyses are not robust to autocorrelation and

https://doi.org/10.1111/jtsa.12833

Talkin' Baseball

Baseball has been played professionally for over 150 years.

The game has gone through several distinct *eras*.



Currently, Major League Baseball (MLB)

- Consist of 30 teams, each with a roster of 26 players
- Separated into two *leagues* and six divisions
- From 1871 through 2024 there have been 21,271 players

Talkin' Baseball

The Basics of the game

- A *pitcher* stands on a *mound* and
- Throws (or pitches) the ball towards *homeplate*.

Talkin' Baseball

The Basics of the game

- A *batter* attempts to *hit* the pitched ball with a wooden bat,
- When hit, the ball is in play and the batter runs the bases.

Talkin' Baseball

The Basics of the game

- The defense tries to get the batter *out*
- While the batter tries to successfully reach base.

Harder than it sounds

The greatest hitters in history fail nearly 2 out of 3 times!!

Player	Hits	Average
Ty Cobb	4189	0.366
Rogers Hornsby	2930	0.358
Shoeless Joe Jackson	1772	0.356
Tris Speaker	3514	0.345
Ted Williams	2654	0.344

Harder than it sounds

The greatest hitters in history fail nearly 2 out of 3 times!!

Player	Hits	Average
Ty Cobb	4189	0.366
Rogers Hornsby	2930	0.358
Shoeless Joe Jackson	1772	0.356
Tris Speaker	3514	0.345
Ted Williams	2654	0.344

When hit over the outfield fence, known as a homerun

Homerun

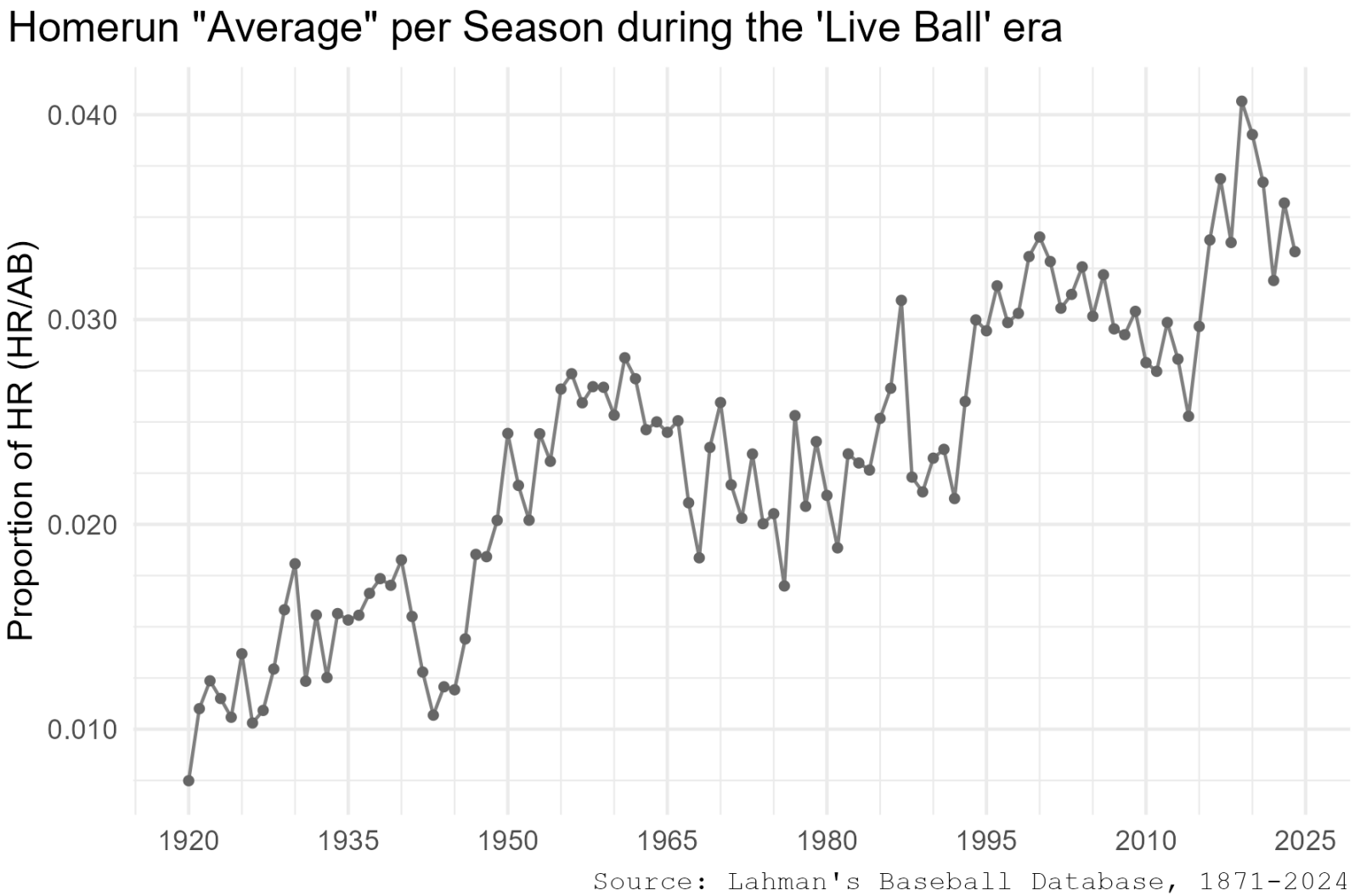
Link

Greatest HR Hitters

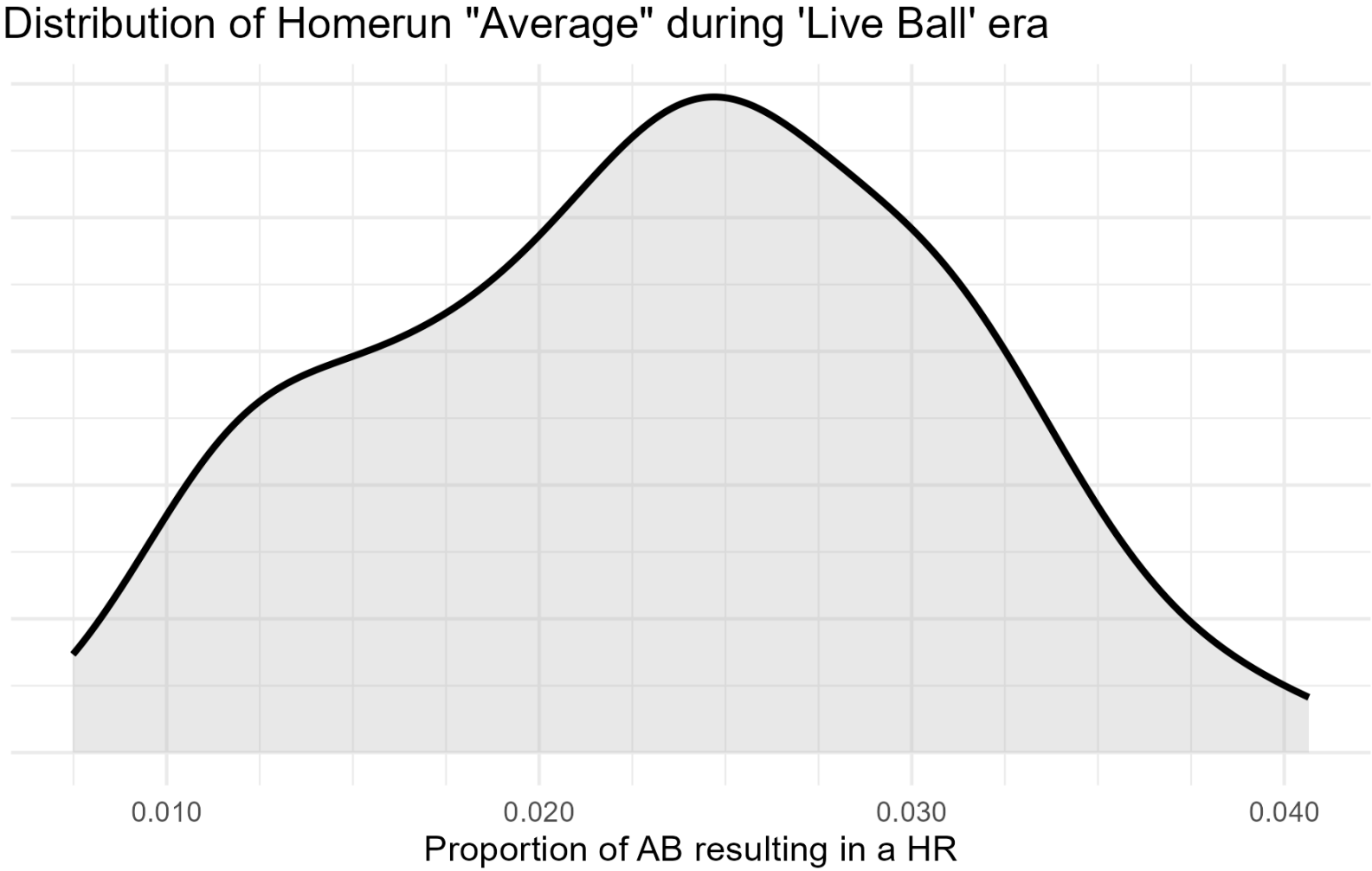
Homeruns are fairly rare events

Player	HR	HR 'Avg'	BA
Barry Bonds	762	0.0774	0.298
Hank Aaron	755	0.0611	0.305
Babe Ruth	714	0.0850	0.342
Albert Pujols	703	0.0616	0.296
Alex Rodriguez	696	0.0659	0.295

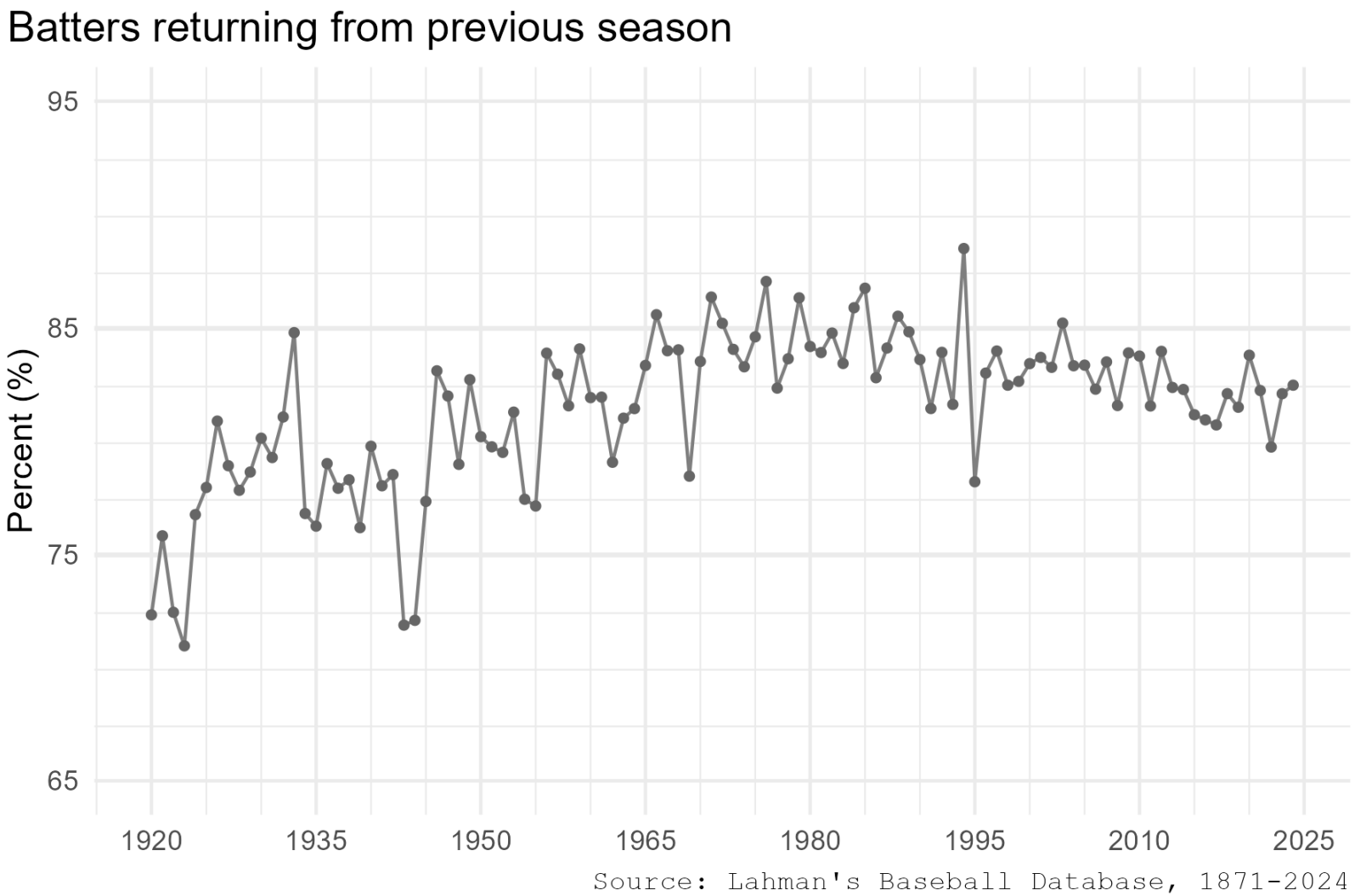
Homeruns in History



Homeruns in History



Players in History



Data Features

Homerun 'averages' in time

- Proportions!
  - Small values (near 0)
  - Mean proportion: 0.0234; standard deviation: 0.00742
- Trend? or *regimes* present
- Almost certainly autocorrelated
  - On average 81.4% of players return year-to-year (SD: 3.4%).
  - Median number of seasons played: 3

Model Form (Beta distribution)

Let X_t be the homerun proportion observed in year t.

We assume $X_t \sim \text{Beta}(\alpha, \beta)$ where the Beta *p.d.f.* is

$$f_{X_t}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \quad 0 \leq x \leq 1,$$

with $B(\alpha, \beta)$ the beta function for $\alpha > 0$ and $\beta > 0$. Moments are

$$E[X_t] = \frac{\alpha}{\alpha+\beta}, \quad \text{and} \quad \text{Var}(X_t) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Can be reparametrized with a precision κ and mean μ defined as

$$\kappa = \alpha + \beta \quad \text{and} \quad \mu = \frac{\alpha}{\kappa}, \text{ respectively.}$$

Can transform back via $\alpha = \mu\kappa$ and $\beta = (1-\mu)\kappa$.

The variability of X_t is inversely related to the precision parameter κ.

Handling Autocorrelation

$\{X_t\}$ has the marginal Beta cumulative distribution function (CDF)

$$F_{\boldsymbol{\theta}_t}(x) = P[X_t \leq x], \quad \text{where } \boldsymbol{\theta}_t = (\mu_t, \kappa_t)'.$$

We *convert* this into a Gaussian series $\{Z_t\}$ via

$$Z_t = \Phi^{-1}(F_{\boldsymbol{\theta}_t}(X_t)). \tag{1}$$

where Φ^{-1} is the inverse of the standard normal CDF:

$$\Phi(z) = \int_{-\infty}^{z} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt$$

In this sense, X_t can be considered a function of Z_t

$$X_t = F^{-1}_{\boldsymbol{\theta}_t}(\Phi(Z_t)), \tag{2}$$

The probability transformation theorem justifies the relationship.

Structure of $\{Z_t\}$

The series $\{Z_t\}$ can be autocorrelated, let

$$\rho_Z(h) = \text{Corr}(Z_t, Z_{t+h}).$$

Assume Z_t is from the class of ARMA models

$$Z_t - \varphi_1 Z_{t-1} - \cdots - \varphi_p Z_{t-p} = \epsilon_t + \beta_1 \epsilon_{t-1} + \cdots + \beta_q \epsilon_{t-q}.$$

Error process $\{\epsilon_t\}$ is $N(0, \sigma_\epsilon^2)$ with σ_ϵ^2 chosen to make $\text{Var}(Z_t) \equiv 1$.
σ_ϵ^2 depends on the ARMA parameters.

The Gaussian likelihood w.r.t. $\{Z_t\}_{t=1}^N$ is

$$L(\boldsymbol{\theta}_{\text{ARMA}} | \{Z_t\}_{t=1}^N) = (2\pi)^{-N/2} \det(\boldsymbol{\Sigma}_Z)^{-1/2} \exp\left(\mathbf{Z}' \boldsymbol{\Sigma}_Z^{-1} \mathbf{Z}\right),$$

with variance-covariance matrix Σ_Z.

Likelihood Function for Z_t

We work with autoregressions (AR) and let $\boldsymbol{\theta}_{\text{AR}} = (\varphi_1, \ldots, \varphi_p)'$ denote all AR model parameters. When $p = 1$ the likelihood can be calculated via

$$-2\ln(L(\varphi_1|\{Z_t\}_{t=1}^N)) =$$

$$N\ln(2\pi) + (N-1)\ln(1-\varphi_1^2) + Z_1^2 + \sum_{t=2}^{N} \frac{(Z_t - \varphi_1 Z_{t-1})^2}{1-\varphi_1^2}.$$

Likelihood Function for X_t

With regard to our observed series $\{X_t\}$, we have the parameters $\boldsymbol{\theta}_F = (\mu, \kappa)$ and $\boldsymbol{\theta}_{\text{AR}} = \varphi$. The likelihood is

$$L(\boldsymbol{\theta}_F, \boldsymbol{\theta}_{\text{AR}} | \{X_t\}_{t=1}^N) = L(\boldsymbol{\theta}_{\text{AR}} | \{Z_t\}_{t=1}^N)|J|,$$

where

$$|J| = \prod_{t=1}^{N} \left| \frac{\partial Z_t}{\partial X_t} \right|.$$

and

$$|J| = (2\pi)^{N/2} \exp\left\{ \frac{1}{2} \sum_{t=1}^{N} \Phi^{-1}(F_{\boldsymbol{\theta}_t}(X_t))^2 \right\} \prod_{t=1}^{N} f_{\boldsymbol{\theta}_t}(X_t).$$

Here, $f_{\boldsymbol{\theta}_t}(X_t)$ is the Beta probability density.

Model Selection

We use penalized likelihood methods to find changepoints in our data.

We set κ and φ to be constant in time, but μ_t can vary by regime.

Bayesian Information Criterion (BIC)

- Well known criteria
- Similar to AIC but prefers simpler models

Minimum Description Length (MDL)

- Based on coding and information theory
- Better models = models with minimal storage
- Each parameter can uniquely be penalized

See Shi et al. [2022] for comparison of these approaches.

Model Selection Details

Two fixed parameters κ and φ, while μ_t can vary by regime.

For m changepoints at times τ_i, there are $m+1$ regimes.

BIC Objective Function

$$-2\ln(L) + \log(N)(2m+2),$$

MDL Objective Function

$$-2\ln(L) + \frac{\ln(N)}{2} + \sum_{j=1}^{m+1} \frac{\ln(\tau_j - \tau_{j-1})}{2} + \ln(m) + \sum_{j=1}^{m} \ln(\tau_j),$$

where the boundaries $\tau_0 = 0$ and $\tau_{m+1} = N$ are enforced.

Optimization of Penalized Likelihood

A Genetic Algorithm is used to find the optimal BIC and MDL objective functions

- Binary search on $N - 1$ potential change point locations
  - Max *iterations* 5000, a *run* of 500 for convergence
  - Mutation probability: 0.1; Crossover probability: 0.8
- *Suggestions* for initial populations
  - No change point configuration
  - All single change point configurations
  - Random selection of 2 change point configurations
- Additional regime penalty (minimum regime length 8 years)

See Lund et al. [2025] for more on genetic algorithms and MDL.

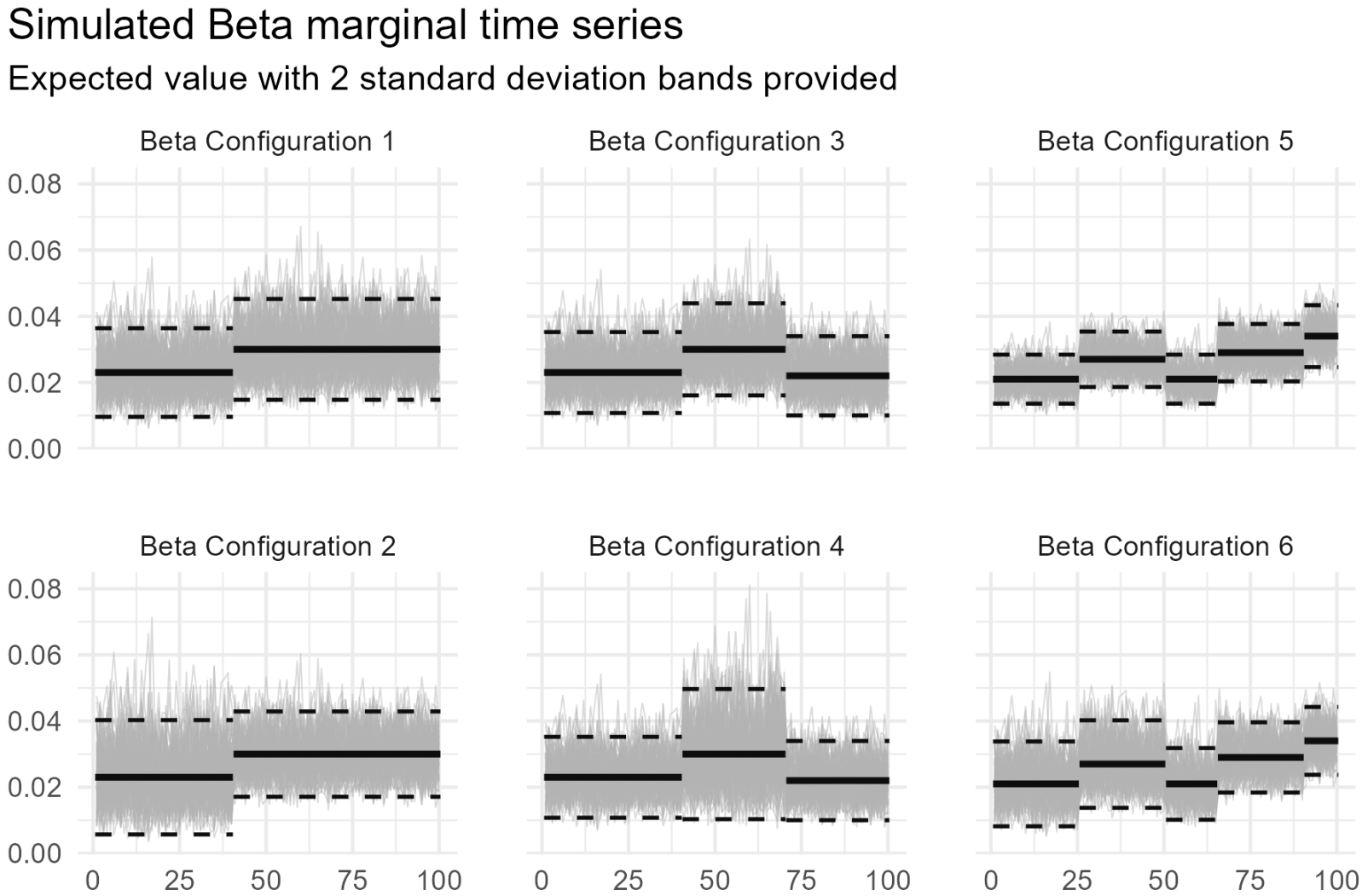
Simulation Setup

Six changepoint configurations.

100 Gaussian AR(1) series with $\varphi = 0.3$ are generated, each is used to build Beta series of length $N = 100$ with the following parameterizations (see (2)).

	Changepoints τ_i	Means μ	Precision κ
1	41	(0.023, 0.030)	500
2	41	(0.023, 0.030)	(400, 700)
3	41, 71	(0.023, 0.030, 0.022)	600
4	41, 71	(0.023, 0.030, 0.022)	(600, 300, 600)
5	26, 51, 66, 91	(0.021, 0.027, 0.021, 0.029, 0.034)	1500
6	26, 51, 66, 91	(0.021, 0.027, 0.021, 0.029, 0.034)	(500, 600, 700, 1000, 1250)

Simulation Setup Visualized



Checking Changepoint Accuracy

Two potential sources of misclassification arise:

- the number of changepoints detected
- the changepoint locations.

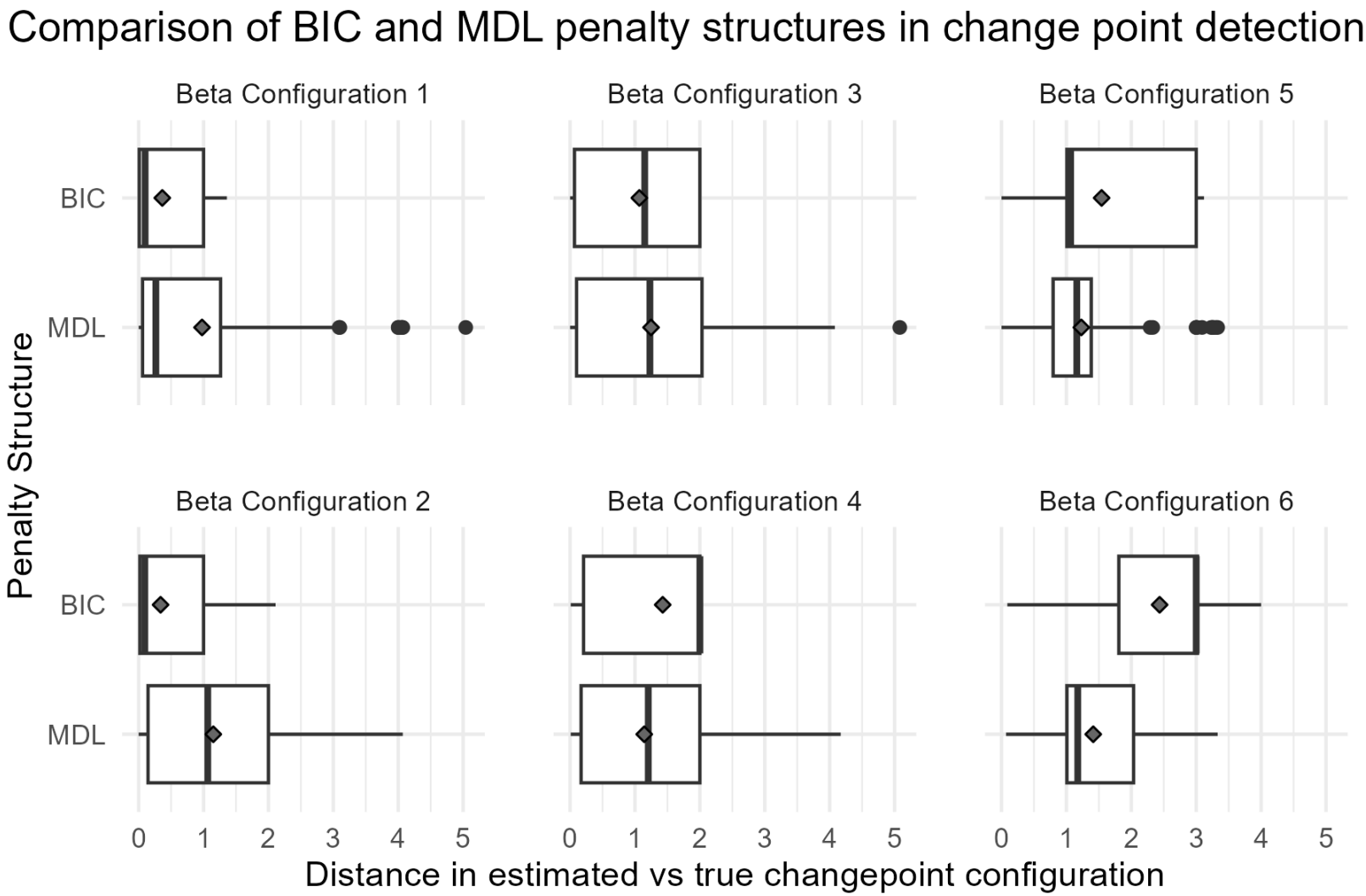
The changepoint configuration distance from Shi et al. [2022] is used, which compares two changepoint configurations, $\mathcal{C}_1$ and $\mathcal{C}_2$, having $m_{\mathcal{C}_1}$ and $m_{\mathcal{C}_2}$ changepoints, respectively.

This distance is defined by

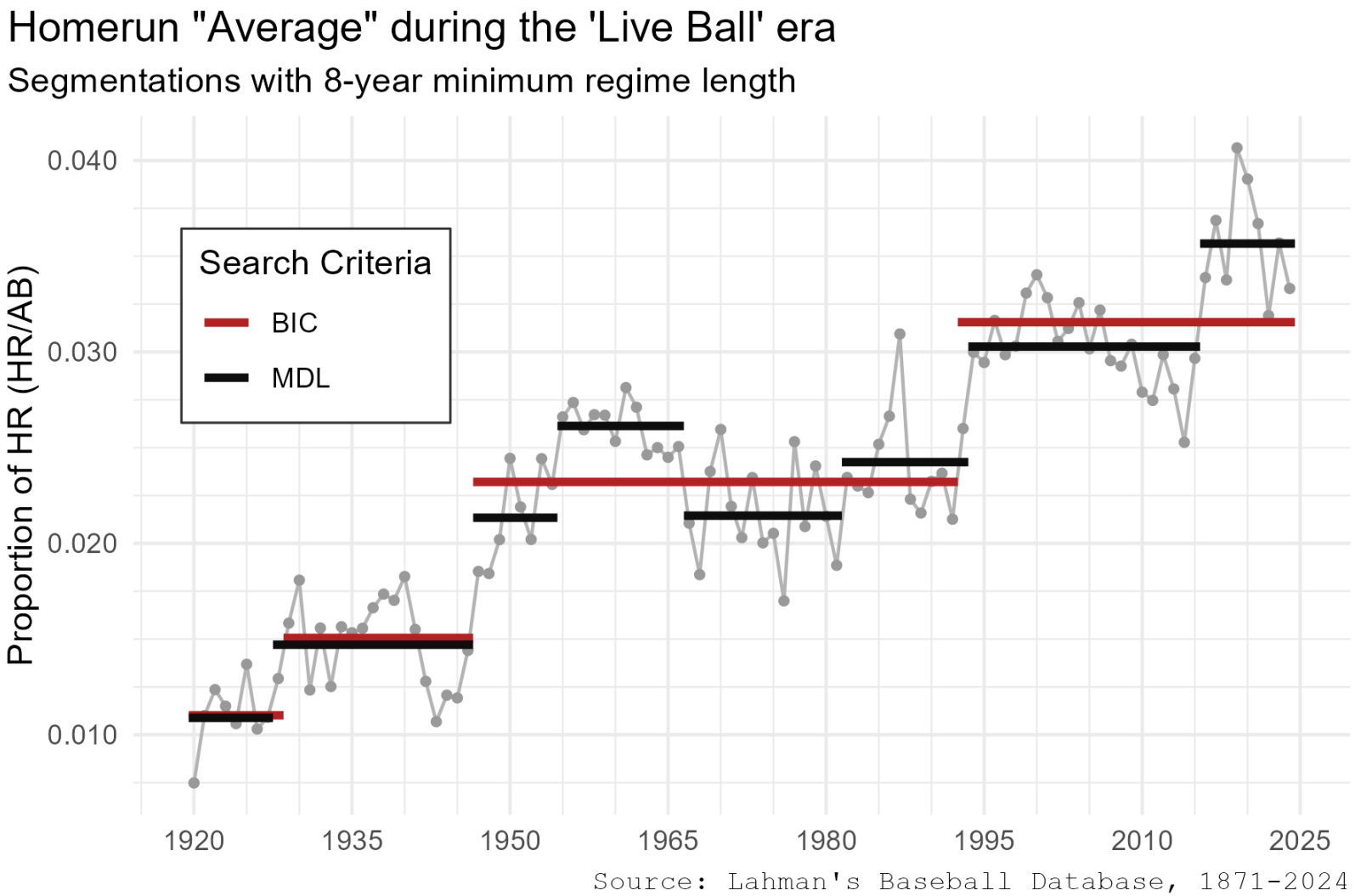
$$d(\mathcal{C}_1, \mathcal{C}_2) = |m_{\mathcal{C}_1} - m_{\mathcal{C}_2}| + \min\left\{\mathcal{A}(\mathcal{C}_1, \mathcal{C}_2)\right\},$$

See Shi et al. [2022] or Lund et al. [2025] for more details.

Simulation Results



Back to Baseball



Fitted Model Comparison

Comparisons of the model fits on baseball homerun 'averages'.

Model	Changepoints τ_i	BIC	MDL	$\hat{\varphi}_1$
Fits assuming iid				
No change points	—	-718.909	-723.563	—
Trend	—	-861.501	-868.482	—
Fits with AR-term				
No changes	—	-911.733	-916.387	0.939
Trend	—	-924.938	-931.919	0.684
BIC Selected	1930, 1948, 1994	-920.537	-938.213	0.474
MDL Selected	1929, 1948, 1956, 1968, 1983, 1995, 2017	-911.491	-945.628	0.202

Assessing Model Adequacy

With a set of $\hat{\mu}_t$, $\hat{\kappa}$ and $\hat{\varphi}$ from our fitted model, we can calculate an estimate for the underlying Normal series via equation (1

$$\hat{Z}_t = \Phi^{-1}(F_{\hat{\boldsymbol{\theta}}_t}(X_t)), \qquad \text{where } \hat{\boldsymbol{\theta}}_t = (\hat{\mu}_t, \hat{\kappa})$$

If the model formulation is adequate, the $\{\hat{Z}_t\}$ series should behave as a Gaussian AR(1) process with $\hat{\varphi}$.

We assess can visually and with goodness-of-fit testing (we use Fisher and Gallagher [2012] and Anderson and Darling [1954] tests below).

Assessing Model Adequacy

With a set of $\hat{\mu}_t$, $\hat{\kappa}$ and $\hat{\varphi}$ from our fitted model, we can calculate an estimate for the underlying Normal series via equation (1

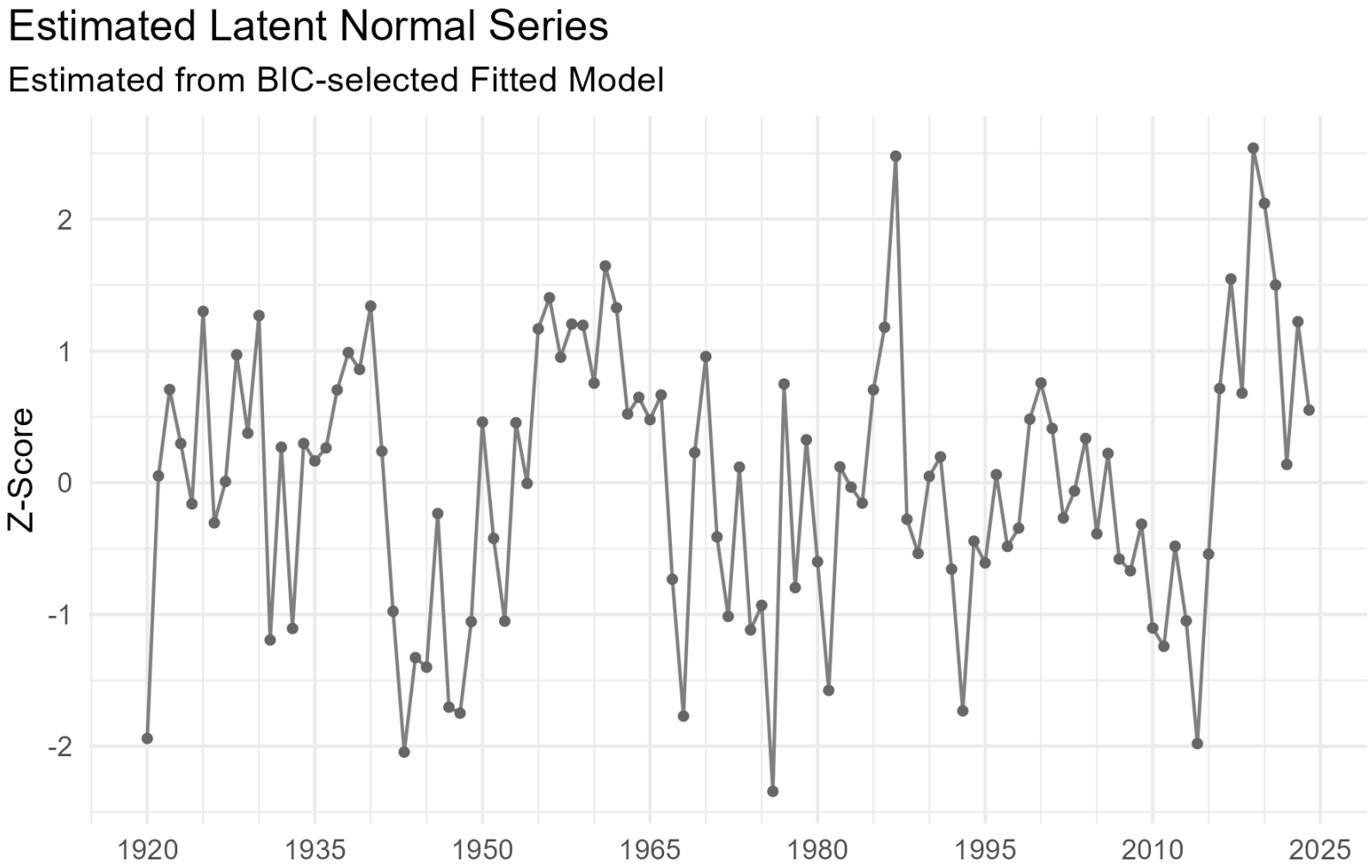
$$\hat{Z}_t = \Phi^{-1}(F_{\hat{\boldsymbol{\theta}}_t}(X_t)), \quad \text{where } \hat{\boldsymbol{\theta}}_t = (\hat{\mu}_t, \hat{\kappa})$$

If the model formulation is adequate, the $\{\hat{Z}_t\}$ series should behave as a Gaussian AR(1) process with $\hat{\varphi}$.

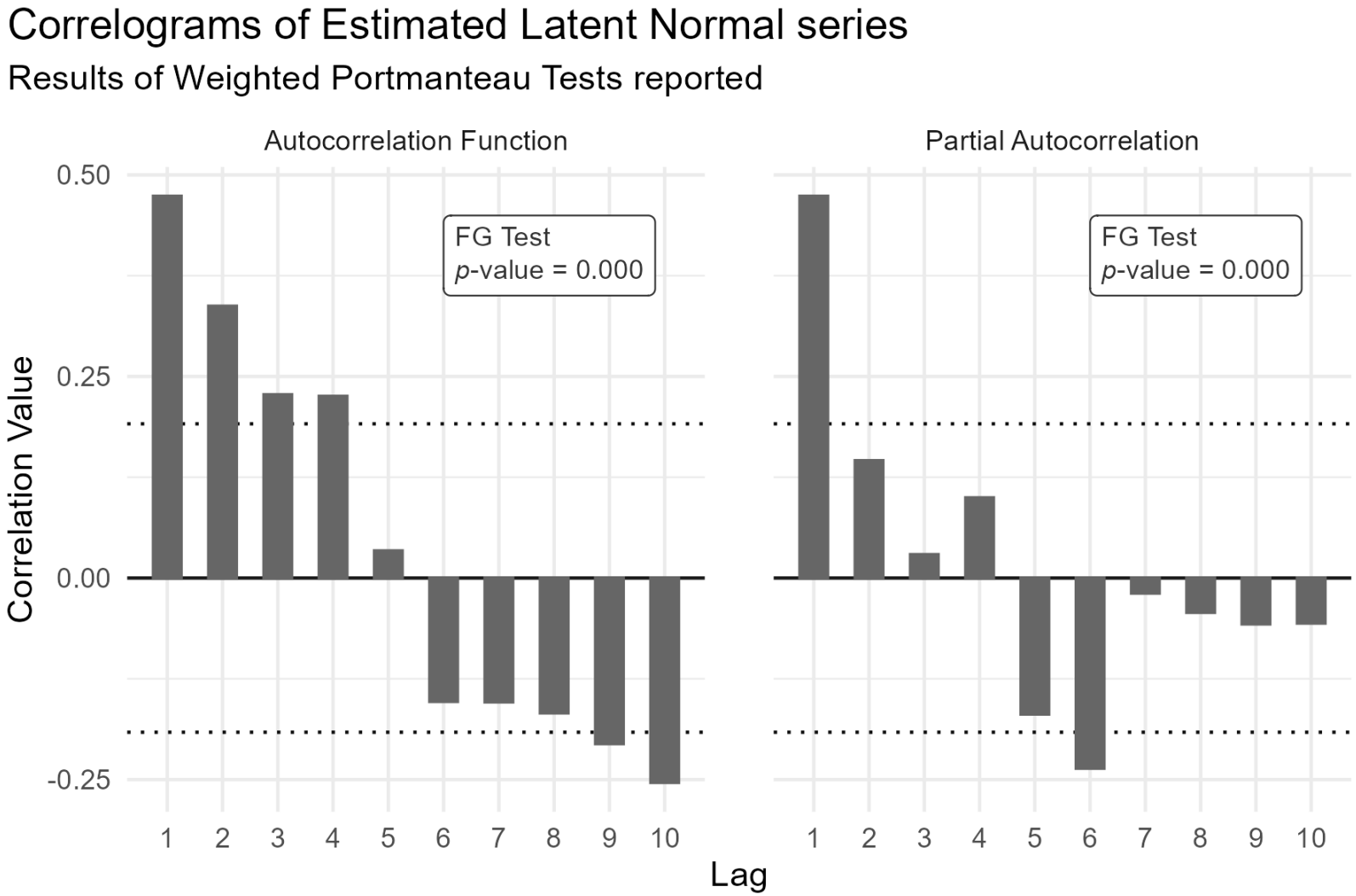
We assess can visually and with goodness-of-fit testing (we use Fisher and Gallagher [2012] and Anderson and Darling [1954] tests below).

Demonstrated on the BIC selected segmented model (four regimes).

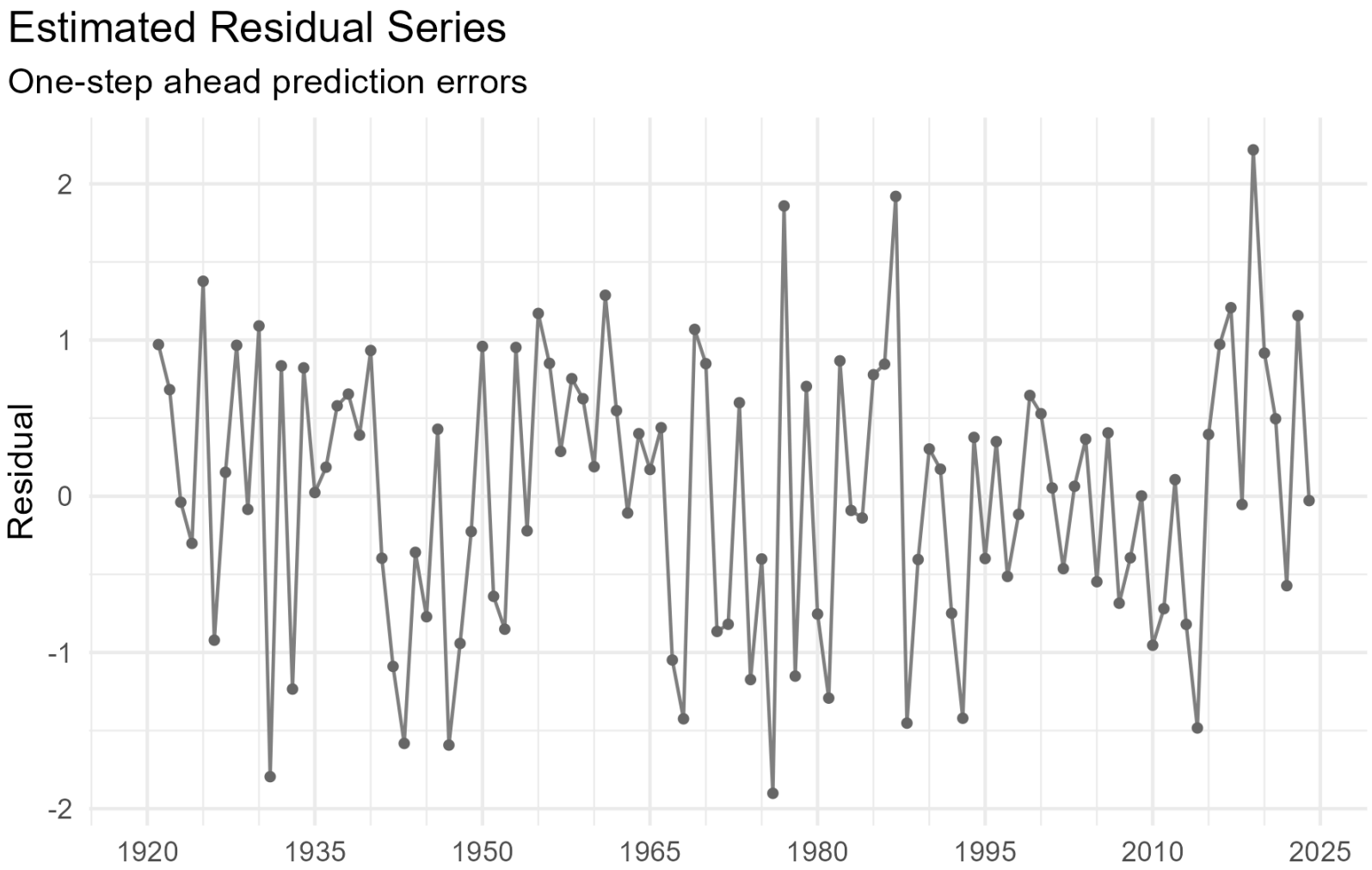
Latent Normal Series



Autocorrelation in Normal Series



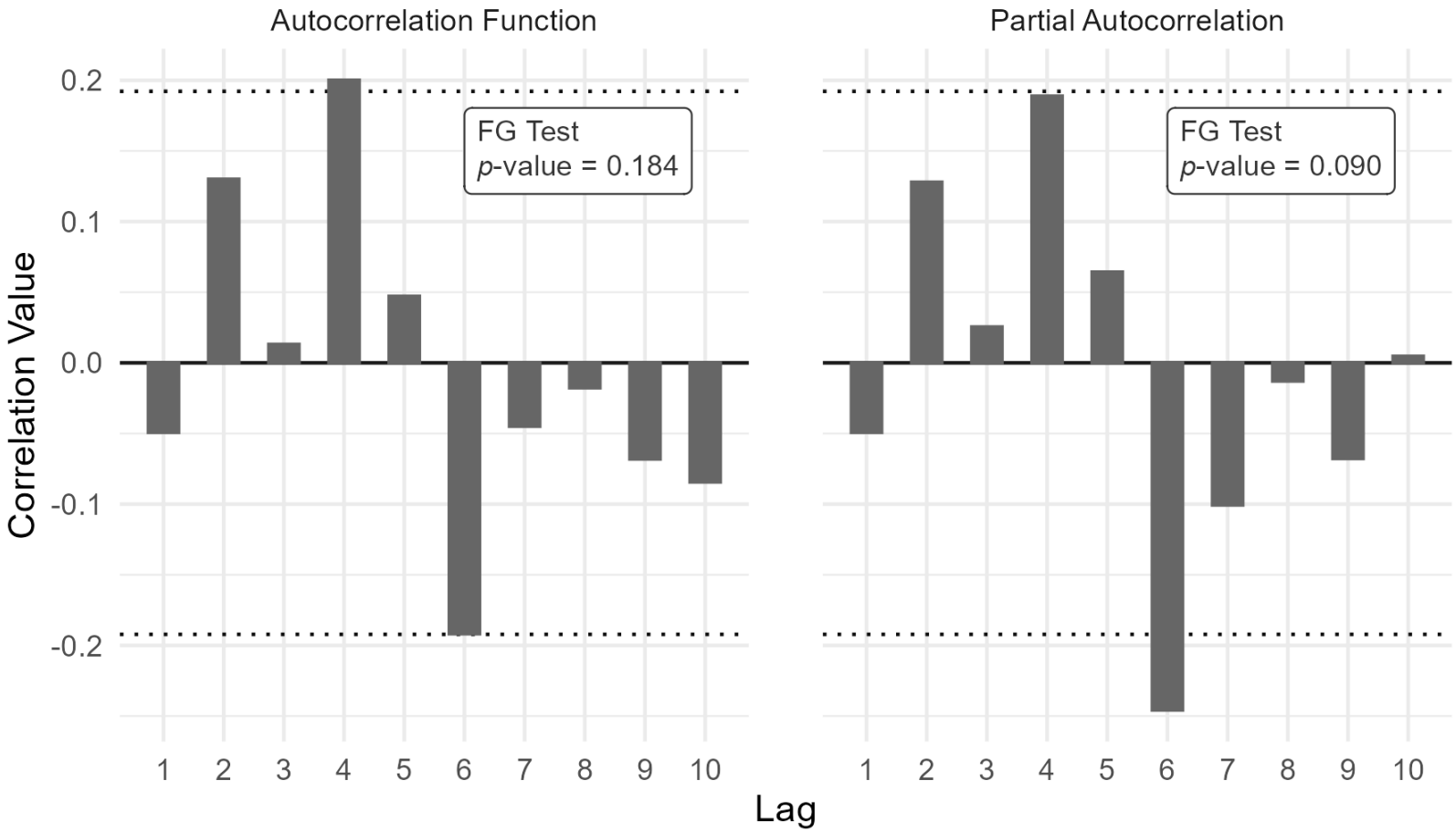
Residual Series



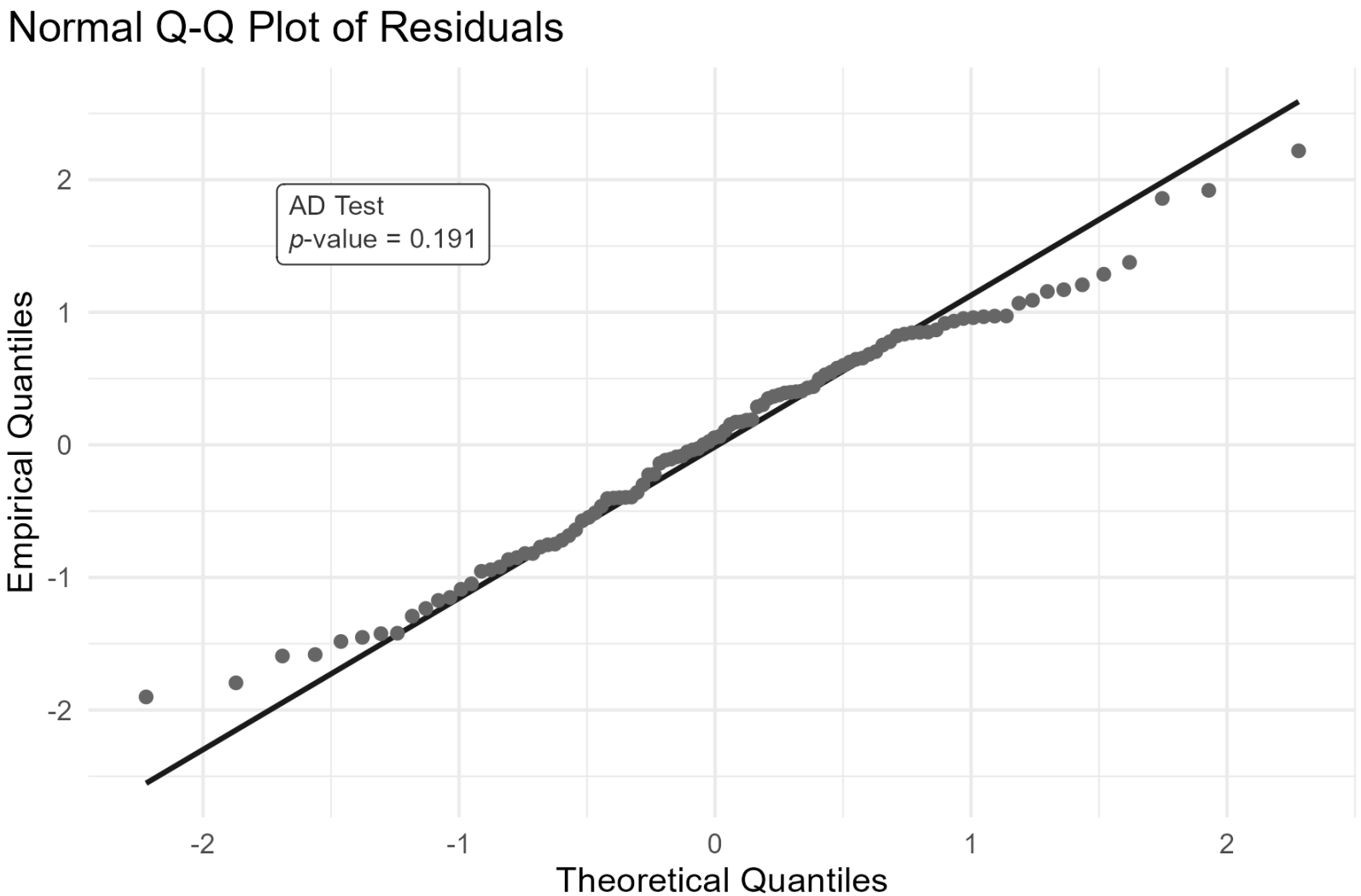
Autocorrelation in Residual Series

Correlograms of Residuals

Results of Weighted Portmanteau Tests reported



Normality of Residual Series



References

T. W. Anderson and D. A. Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954. URL https://doi.org/10.1080/01621459.1954.10501232.

Thomas J. Fisher and Colin M. Gallagher. New weighted portmanteau statistics for time series goodness of fit testing. *J. Amer. Statist. Assoc.*, 107(498):777–787, 2012. ISSN 0162-1459. doi: 10.1080/01621459.2012.688465. URL http://dx.doi.org/10.1080/01621459.2012.688465.

Robert Lund, Thomas J. Fisher, Norou Diawara, and Michael Wehner. Multiple changepoint detection for non-gaussian time series. *Journal of Time Series Analysis*, n/a(n/a), 2025. URL https://doi.org/10.1111/jtsa.12833.

Xuesheng Shi, Colin Gallagher, Robert Lund, and Rebecca Killick. A comparison of single and multiple changepoint techniques for time series data. *Computational Statistics & Data Analysis*, 170:107433, 2022. ISSN 0167-9473. URL https://doi.org/10.1016/j.csda.2022.107433.

Thank You