Chapter 22
Changepoint Inference for Erdős–Rényi Random Graphs

Elena Yudovina, Moulinath Banerjee, and George Michailidis

Abstract We formulate a model for the off-line estimation of a changepoint in a network setting. The framework naturally allows the parameter space (network size) to grow with the number of observations. We compute the signal-to-noise ratio detectability threshold, and establish the dependence of the rate of convergence and asymptotic distribution on the network size and parameters. In addition, we show that inference can be adaptive, i.e. asymptotically correct confidence intervals can be computed based on the data. We apply the method to the question of whether US Congress has abruptly become more polarized at some point in recent history.

22.1 Introduction

The problem of estimating the location of a jump discontinuity (*changepoint*) has been extensively studied in the statistics literature. There are two versions of the problem. The *on-line* version is concerned with the quickest detection of a changepoint in the parameters of a dynamic stochastic system, and is closely related to classical problems in sequential analysis; for a comprehensive treatment, together with a discussion of important applications, see the books by Siegmund [18], Basseville and Nikiforov [1], and the review article by Lai [12] and references therein. In the *off-line* version, data are available for n covariate-response pairs, and one is interested in estimating the location of the changepoint as accurately as possible

E. Yudovina (✉)
Department of Mathematics, University of Minnesota, 127 Vincent Hall, 206 Church St. S.E., Minneapolis, MN 55455, USA
e-mail: eyudovin@umn.edu

M. Banerjee · G. Michailidis
Department of Statistics, University of Michigan, 439 West Hall, 1085 South University, Ann Arbor, MI 48109, USA

M. Banerjee
e-mail: moulib@umich.edu

G. Michailidis
e-mail: gmichail@umich.edu

A. Steland et al. (eds.), *Stochastic Models, Statistics and Their Applications*,
Springer Proceedings in Mathematics & Statistics 122,
DOI 10.1007/978-3-319-13881-7_22

(see Ritov [17], Müller [15], Loader [13], Gijbels, Hall and Kneip [6], Hall and Molchanov [7], Kosorok and Song [11], and the book by Csörgő and Horváth [4]). The on-line version is also closely related to many developments in statistical process control (Hawkins et al. [8]) and associated control charts (e.g. Cumulative Sums (CUSUM), Exponential Weighted Moving Average (EWMA), etc.). However, both versions of the problem have dealt primarily with low- (usually one-) dimensional problems. Although there have been some extensions to multivariate data, they are usually obtained under an assumption of multivariate normality that gives rise to Hotelling's T^2 test.

In this paper, we consider the off-line version in a high-dimensional network setting. Data are indexed by the edges of a graph; in the simplest case, binary data indicate whether the edge is present. We consider edges which evolve independently, so that at each point in time the network looks like an Erdős–Rényi random graph. This is a fundamental problem in changepoint analysis on networks, and already presents technical challenges. As graph size grows, we acquire more data about the changepoint, but have to deal with a higher-dimensional nuisance parameter space; this interaction is the main technical focus of the paper. We obtain the limiting distributions of the maximum likelihood estimates of both the changepoint and the remaining model parameters; although the asymptotic distribution for the changepoint estimate depends on the (unknown) signal-to-noise ratio, we develop an adaptive inference framework that does not require prior information about the limiting regime. Many of our results generalize those known for finite-dimensional models, although to our knowledge the focus on adaptive inference is new.

As a motivating application, we consider the question of whether the US Congress has abruptly become more polarized at some point in recent history. This question has raised a lot of interest in the political science literature; see for example [14, 16]. These works were primarily exploratory in nature, and no attempt was made to make inferences regarding the polarization process. Within the framework of our network-based approach, we use roll call vote data to generate a sequence of graphs, with vertices corresponding to congressmen and edges corresponding to whether they voted in the same way on a particular issue. We are then able to make inference about any changepoints in voting pattern.

Due to space constraints, we skip most of the details. A more extensive version of the paper is in preparation.

22.2 Network Changepoint Model and Estimators

Consider a sequence of random graphs indexed by n. Each graph has $m = m(n)$ potential edges; we allow $m(n)$ to grow with n. Each edge has a *state* $\alpha \in \mathscr{S}$; for simplicity, in this note we take $\mathscr{S} = \{0, 1\}$, but the model readily extends to arbitrary common finite state space. We assume that the underlying graphs are embedded into each other, so that it makes sense to speak of "edge 1 of system n". The edges evolve in discrete time; each edge evolves as a Markov chain with its own transition kernel,

independently of all the other edges. Consequently, at each time point the state of the system is an Erdős–Rényi random graph (with different, time-varying, probabilities for each edge). We assume that edges transition according to one set of transition kernels $\{P_k^*,\ 1 \le k \le m(n)\}$ before a time t^*, the *changepoint*, and according to another set of transition kernels $\{Q_k^*,\ 1 \le k \le m(n)\}$ after t^*. The changepoint t^*, as well as the matrices P_k^* and Q_k^*, may depend on n; but note that t^* is the same for all the edges. We may also have $P_k^* = Q_k^*$ for some edges, i.e. the changepoint may only affect a subset of the edges in the graph. For convenience, we will rescale time so that $t^* \in [0, 1]$.

We make n observations of the graph indexed by n, at times $\{\frac{i}{n},\ i = 1, \ldots, n\}$. This means that in the nth experiment, $t^* = t^*(n) \in \{\frac{i}{n}\},\ i = 1, \ldots, n$. We will assume $t^*(n) \to t^0$ as $n \to \infty$, as well as $P_k^* \to P_k^0$ and $Q_k^* \to Q_k^0$ for each k. Below, we will frequently omit the dependence on n.

Let $\mathbf{1}_{k,\alpha\to\beta}(s)$ be the indicator of the event that edge k was in state α at time s and in state β at time $s+1$. The log-likelihood function for this model is

$$l_n^M(P, Q, t) = n^{-1}\left(\sum_{k=1}^{m}\sum_{\alpha,\beta\in\mathscr{S}}\left(\sum_{s=0}^{nt-1}\left(\mathbf{1}_{k,\alpha\to\beta}(s)\log(P_k)_{\alpha\beta}\right)\right.\right.$$
$$\left.\left.+\sum_{s=nt}^{n-1}\left(\mathbf{1}_{k,\alpha\to\beta}(s)\log(Q_k)_{\alpha\beta}\right)\right)\right). \tag{22.1}$$

If the changepoint were at t, we could write down the MLEs $\hat{P} = \hat{P}(t)$ and $\hat{Q} = \hat{Q}(t)$:

$$\left(\hat{P}_k(t)\right)_{\alpha\beta} = \frac{\sum_{s=0}^{nt-1}\mathbf{1}_{k,\alpha\to\beta}(s)}{\sum_{s=0}^{nt-1}\sum_{\gamma\in\mathscr{S}}\mathbf{1}_{k,\alpha\to\gamma}(s)},$$
$$\left(\hat{Q}_k(t)\right)_{\alpha\beta} = \frac{\sum_{s=nt}^{n-1}\mathbf{1}_{k,\alpha\to\beta}(s)}{\sum_{s=nt}^{n-1}\sum_{\gamma\in\mathscr{S}}\mathbf{1}_{k,\alpha\to\gamma}(s)}. \tag{22.2}$$

The MLE $\hat{t}$ can be obtained by iterating over $t \in [0, 1]$ (on the grid of discrete observation times), using the above form for $\hat{P}$ and $\hat{Q}$; in case of ties, we take the smallest maximizer.

Our main results will concern the asymptotic behavior of $\hat{P}$, $\hat{Q}$, and $\hat{t}$ as $n \to \infty$. Below, we describe the necessary assumptions on the behavior of the dimension $m(n)$, the "signal" $\sum_k \|P^* - Q^*\|_F$, and the values of true parameters. Here, $\|A\|_F = (\sum_{i,j} A_{ij}^2)^{1/2}$ is the Frobenius, or Hilbert–Schmidt, norm of the matrix A; and we write $\|P^* - Q^*\|_F^2 = \sum_k \|P_k^* - Q_k^*\|_F^2$.

Assumption 22.1

1. *The underlying parameters converge as follows.*
   a. *$m(n)$ is either constant $m(n) = m^0$ or else monotonically increasing to infinity.*

   b. $t^*(n) \to t^0$ *as* $n \to \infty$. (*For example, we could have* $t^*(n) = n^{-1}\lfloor nt^0 \rfloor$.)
   c. $P_k^*(n) \to P_k^0$ *and* $Q_k^*(n) \to Q_k^0$ *uniformly in* k.
2. *There exists a constant* $\varepsilon > 0$ (*which we need not know*) *such that, for each* k, *one of the following holds*: *either* $\|Q_k^0 - P_k^0\|_F > \varepsilon$, *or else* $Q_k^0 = P_k^0$.
3. *For each* n *and* k, *the transition matrices* $P_k^*(n)$ *and* $Q_k^*(n)$ *correspond to irreducible, aperiodic Markov chains with state space* $\mathscr{S}$. *There exists some known constant* $c > 0$ *such that* $t^* \in (c, 1-c)$, *and all entries of* P_k^* *and* Q_k^* *belong to* $(c, 1-c)$. (*The same is then true of* t^0, P_k^0, *and* Q_k^0.) *We will only consider estimates of the changepoint that fall within* $(c, 1-c)$.
4. *The number of edges* m *satisfies* $n^{-1/2}\log m(n) \to 0$.
5. *The signal-to-noise ratio satisfies* $\frac{n}{m}\sum_{k=1}^m \|P_k^* - Q_k^*\|_F^2 \to \infty$.

Remark 22.1 Assumption 22.1.3 implies that the Markov chains with transition kernels P_k^* and Q_k^* have uniformly bounded mixing times; in particular, observations $\mathbf{1}_{k,\alpha\to\beta}(\cdot)$ form a mixing sequence, with mixing coefficients bounded uniformly in k. For discussion of variants of the changepoint problem where the changepoint is very close to the edge of the interval, see for example [4, Theorem 1.5.3].

Assumption 22.1.4 implies that with high probability, all estimates $\hat{P}_k(t)$ and $\hat{Q}_k(t)$ will satisfy Assumption 22.1.3; and together with Assumption 22.1.2, it means that we will correctly identify which of the edges experienced a change at t^*. The requirement $n^{-1/2}\log m(n) \to 0$ still allows quite large graphs, e.g. we may have $m(n) = \exp(n^{1/4})$.

Assumption 22.1.5 asserts that the "average" per-edge signal $\|P_k^* - Q_k^*\|_F^2 \gg n^{-1}$. With finitely many edges ($m(n) = m^0$), this is necessary for detectability; when $m(n) \to \infty$, the necessary condition is very slightly weaker.

22.3 Results

We now present our main results. Theorem 22.1 addresses the rates of convergence of the estimators and their asymptotic distributions. Finally, Theorem 22.2 addresses the question of adaptive inference, that is, inferring the parameters of the asymptotic distribution from the data.

Because the exact formulae below get somewhat involved, we state only the qualitative form of the limiting processes and distributions. Full expressions for the parameters will be found in our forthcoming longer paper on the subject. The form of the result is qualitatively similar to finite-dimensional models, cf. [4, Chap. 1], although our model is considerably more general.

Theorem 22.1 (Rates of convergence and asymptotic distribution.) *Under Assumptions* 22.1.1 *through* 22.1.5, $n\|Q^* - P^*\|_F^2|\hat{t} - t^*| = O_P(1)$.

For any finite set of edges K *and simultaneously for all* $k \in K$, $n\|\hat{P}_k - P_k^*\|_F^2 = O_P(1)$ *and* $n\|\hat{Q}_k - Q_k^*\|_F^2 = O_P(1)$.

Define the local parameters $h_k^P = \sqrt{n}(P_k - P_k^*)$, $h_k^Q = \sqrt{n}(Q_k - Q_k^*)$. *For each* k, h_k^P *and* h_k^Q *are asymptotically normal*:

$$\big(h_k^P\big) \Longrightarrow N\big(0, \big(t^0\big)^{-1} V_k^P\big), \qquad h_k^Q \Longrightarrow N\big(0, \big(t^0\big)^{-1} V_k^Q\big),$$

where the $S^2 \times S^2$ *covariance matrices* V_k^P, V_k^Q *depend on* P_k^0, *respectively* Q_k^0. *For any fixed finite set* K *of edges, the estimates* $\{\hat{h}_k^P, \hat{h}_k^Q, \hat{t}\colon\ k \in K\}$ *are asymptotically independent.*

For the limiting distribution of $(\hat{t} - t^*)$, *we distinguish three cases, one of which is further subdivided*:

1. *If* $\|P^* - Q^*\|_F^2 \to \infty$, *then* $n(\hat{t} - t^*) \to 0$ *in probability. That is, asymptotically we precisely identify the index of the transition where the transition probability matrix changed.*
2. *If* $\|P^* - Q^*\|_F^2 \to 0$, *then*

$$n \sum_{k=1}^{m} \sum_{\alpha,\beta \in \mathscr{S}} \frac{(\pi_k^0)_\alpha}{(P_k^0)_{\alpha\beta}} \big((P_k^* - Q_k^*)_{\alpha\beta}\big)^2 \big(\hat{t} - t^*\big) \to \sigma^{-1} \arg\max_{h \in \mathbb{R}} \left(B(h) - \frac{1}{2}|h| \right),$$

   where $B(h)$ *is a standard Brownian motion, and* σ^2 *comes from the Markov chain central limit theorem* (*cf.* [10, *Case* 1 *of Theorem* 5]).
3. *If* $\|P^* - Q^*\|^2 \to C \in (0, \infty)$, *then* $n(\hat{t} - t^*)$ *converges to the* (*smallest*) *maximizer of a limiting jump process supported on* $\mathbb{Z}$: $n(\hat{t} - t^*) \to \arg\max_{h \in \mathbb{Z}} [M(h) + G(h) - D(h)]$. *Here,* D *is a deterministic triangular drift,* G *is a random walk with correlated Gaussian step sizes, and* M *is a functional of the Markov chain trajectories of some of the edges. Let* $\mathscr{I}_+ = \{k\colon\ P_k^0 \neq Q_k^0\}$ (*necessarily finite*); $M(\cdot)$ *depends only on the edges in* $\mathscr{I}_+$, *and* $D(\cdot)$ *and* $G(\cdot)$ *depend only on the remaining edges.*

Interestingly, the network size m does not appear in the scaling of $\hat{t} - t^*$; however, Assumption 22.1.5 places a lower bound on $\|Q^* - P^*\|_F^2$ that scales with m.

The proofs follow the approach of [20, Theorem 3.4.1], making extensive use of Doob's martingale maximal inequality (the use for Markov chains is somewhat unusual). The continuity of the argmax functional in Case 22.1.3 is non-standard. The high-dimensional nuisance parameter space makes it hard to apply many classical changepoint techniques, such as those in [4].

Lastly, we present a result which allows adaptive inference of the limiting distribution from the data, irrespective of the limiting regime that applies. This means that we can provide asymptotically correct quantile estimation of the distribution based only on the data, without knowledge of the true parameters. The adaptive process is essentially the one that appears in case 3 of Theorem 22.1 when $|\mathscr{I}_+| = m$.

Theorem 22.2 (Adaptive inference.) *Define the process* $\widetilde{M}(h)$ *as follows. Let* $\tilde{X}_k(h), h \geq 0$ *be the* reversed *Markov chain with initial distribution* $\hat{\pi}_k$ *and transition kernel* $\hat{\mathscr{P}}_k$, $(\hat{\mathscr{P}}_k)_{\alpha\beta} = \frac{(\hat{\pi}_k)_\beta}{(\hat{\pi}_k)_\alpha} (\hat{P}_k)_{\beta\alpha}$. *Here,* $(\hat{\pi}_k)_\alpha := \sum_{s=0}^{n\hat{t}-1} \sum_{\beta \in \mathscr{S}} \mathbf{1}_{k,\alpha\to\beta}(s)$

is the empirical proportion of time that edge k spends in state α up to time $\hat{t}$. Let $\tilde{Y}_k(h), h \geq 0$ be the (ordinary) Markov chain with initial distribution $\hat{\pi}_k$ and transition kernel $\hat{Q}_k$. For different values of k, let the Markov chains be independent; moreover, let $X_k(0) = Y_k(0)$ and let their transitions be independent otherwise. Define

$$\widetilde{M}(h+1) - \widetilde{M}(h) = \begin{cases} \sum_{k=1}^{m} \sum_{\alpha,\beta \in \mathscr{S}} \mathbf{1}_{\tilde{Y}_k, \alpha \to \beta}(h) \log \frac{(\hat{P}_k)_{\alpha\beta}}{(\hat{Q}_k)_{\alpha\beta}}, & h \geq 0, \\ \sum_{k=1}^{m} \sum_{\alpha,\beta \in \mathscr{S}} \mathbf{1}_{\tilde{X}_k, \beta \to \alpha}(|h|-1) \log \frac{(\hat{P}_k)_{\alpha\beta}}{(\hat{Q}_k)_{\alpha\beta}}, & h < 0. \end{cases}$$

Let $\tilde{h}$ be the smallest maximizer of $\widetilde{M}(\cdot)$. Then $\tilde{h}$ has the same asymptotic distribution as $n(\hat{t} - t^*)$, in the following sense:

1. *If $\|Q^* - P^*\|_F^2 \to \infty$, then both $\tilde{h} \to 0$ and $n(\hat{t} - t^*)$ in probability.*
2. *If $\|Q^* - P^*\|_F^2 \to 0$, then we have convergence in distribution for the renormalized estimate*:

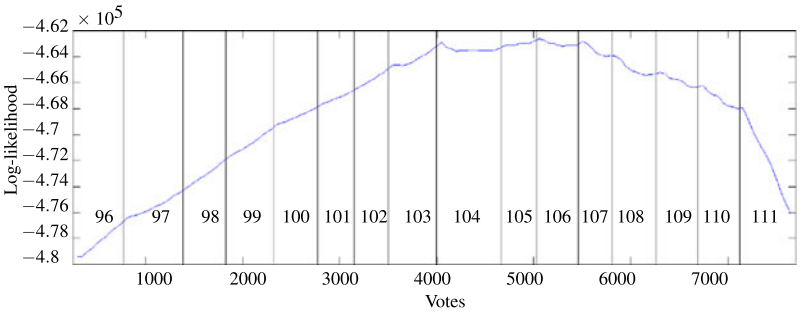
$$\sum_{k=1}^{m} \sum_{\alpha,\beta \in \mathscr{S}} \frac{(\pi_k^0)_\alpha}{(P_k^0)_{\alpha\beta}} \left(\left(P_k^* - Q_k^*\right)_{\alpha\beta} \right)^2 \tilde{h} \to \sigma^{-1} \arg\max_{h \in \mathbb{R}} B(h) - \frac{1}{2}|h|,$$

   where $B(h)$ is a standard Brownian motion, and σ^2 is as in Theorem 22.1.
3. *If $\|Q^* - P^*\|_F^2 \to C \in (0, \infty)$, then $\tilde{h} \to \arg\max_{h \in \mathbb{Z}} [M(h) + G(h) - \frac{1}{2} D(h)]$, where $M(\cdot)$, $G(\cdot)$, and $D(\cdot)$ are as in Theorem 22.1.*

22.4 Application: Polarization in US Congress

We consider the question of whether the dynamics of discussion in the US Senate have experienced a changepoint in recent past. To construct the sequence of graphs as above, we identify the senators with senate seats (two per state, e.g. Michigan 1 and Michigan 2). We then consider 7949 roll call votes on bills during the years 1979–2012. The state of the edges of the (complete) graph on 100 vertices is then 1 if the corresponding senators voted in the same way on the issue, and 0 if they voted differently. The Markovian structure is, of course, an approximation of this data, but represents the fact that a particular pair of senators will tend to either agree or disagree on most issues. We note that while the occupants of a particular seat can change, this does not occur very often in practice, so the assumption that the parameters of the model are time-independent aside from the changepoint is not unreasonable.

In Fig. 22.1, we present the (profile) log-likelihood function for the location of the changepoint. We see broadly that the log-likelihood function reaches its maximum somewhere between the 104th and 107th Congresses, i.e. 1995–2003. (2003 corresponds to the Iraq war.) Within this interval, there are several local maxima; as the table to the right of Fig. 22.1 shows, which changepoint is dominant depends in particular on when data analysis starts. We can also examine the nature of the change



Year	Estimate	CI
1995	4025	(3995, 4152)
1999	5100	(5000, 5225)
2001	5850	(5775, 5875)

Fig. 22.1 Log-likelihood function for the senate roll call data. The *horizontal axis* is labelled with the index of the roll call vote; *vertical bands* identify the Congress, i.e. the two-year inter-election period. The table to the right presents the dominant changepoint as a function of the year when data collection begins

by examining the estimated transition parameters before and after the changepoint (in this case, before the 104th and after the 107th Congress). We do not show the graphs due to space constraints, but the average probability of changing the status of an edge decreases by almost a factor of 2, from approximately 0.2 to approximately 0.1, leading to longer negotiation times until a compromise is reached.

22.5 Discussion and Simulation Issues

We have presented a model which can address questions of changepoint inference in a networked setting. We begin by discussing several extensions of the model assumptions, and then discuss the computational complexity of the estimation.

Vertex Labels and Dependent Edges A natural extension to community structures is to add labels to the vertices (e.g. political party affiliation for the US Congress), and allow dependence among the edges. There are many possibilities for such extensions; some are the subject of future work.

Multiple Changepoints Although our research is only directly applicable under the assumption of exactly one changepoint, we may use techniques similar to the binary segmentation method of [3, 21] to find multiple changepoints. The basic idea is to locate the dominant changepoint, and keep looking in the two smaller subintervals around it; an extra elimination step may reduce the probability of finding too many changepoints. In general, estimating multiple changepoints is a challenging issue; we refer to the survey article [9] for a discussion of current approaches.

Computational Complexity When the signal-to-noise ratio is either quite large or quite small (Cases 22.1 and 22.2 of Theorem 22.1), computing $\hat{t}$ is the main computational challenge; the distribution of the maximizer of a Brownian motion with triangular drift, which appears in Case 22.2, can be computed precisely [2, 19]. In Case 22.3, which corresponds to the adaptive regime, the limiting process is easily simulated if $P^0 = Q^0$; see also Fotopoulos et al. [5] for computing the maximizer. However, even in the case of Gaussian jumps, there is not a universal scaling that can relate different examples to each other, in part due to the non-stationarity of the process. For the generalized binomial component of the limiting random process, it seems necessary to simulate the trajectories of all m Markov chains in order to estimate the maximizer; the computation is, however, parallelizable, and can scale up to fairly large networks.

Acknowledgements E.Y.'s research was partially supported by US NSF grant DMS-1204311. M.B.'s research was partially supported by US NSF DMS-1007751, US NSA H98230-11-1-0166, and a Sokol Faculty Award, University of Michigan. G.M.'s research was partially supported by US NSF DMS-1228164 and US NSA H98230-13-1-0241. The authors thank the referees for helpful comments.

References

1. Basseville M, Nikiforov IV (1993) Detection of abrupt changes: theory and application. Prentice Hall, Englewood Cliffs
2. Bhattacharya PK, Brockwell PJ (1976) The minimum of an additive process with applications to signal estimation and storage theory. Probab Theory Relat Fields 37(1):51–75
3. Cho H, Fryzlewicz P (2012) Multiple change-point detection for high-dimensional time series via sparsified binary segmentation. Preprint
4. Csörgő M, Horváth L (1997) Limit theorems in change-point analysis. Wiley, New York
5. Fotopoulos SB, Jandhyala VK, Khapalova E (2010) Exact asymptotic distribution of change-point MLE for change in the mean of Gaussian sequences. Ann Appl Stat 4(2):1081–1104
6. Gijbels I, Hall P, Kneip A (1999) On the estimation of jump points in smooth curves. Ann Inst Stat Math 51(2):231–251
7. Hall P, Molchanov I (2003) Sequential methods for design-adaptive estimation of discontinuities in regression curves and surfaces. Ann Stat 31(3):921–941
8. Hawkins DM, Qiu P, Kang CW (2003) The changepoint model for statistical process control. J Qual Technol 35(4):355–366
9. Jandhyala V, Fotopoulos S, MacNeill I, Liu P (2013) Inference for single and multiple change-points in time series. J Time Ser Anal 34(4):423–446
10. Jones GL (2004) On the Markov chain central limit theorem. Probab Surv 1:299–320
11. Kosorok MR, Song R (2007) Inference under right censoring for transformation models with a change-point based on a covariate threshold. Ann Stat 35(3):957–989
12. Lai TL (2001) Sequential analysis: some classical problems and new challenges. Stat Sin 11(2):303–350
13. Loader CR (1996) Change point estimation using nonparametric regression. Ann Stat 24(4):1667–1678
14. Moody J, Mucha PJ (2013) Portrait of political party polarization. Netw Sci 1(01):119–121
15. Müller HG (1992) Change-points in nonparametric regression analysis. Ann Stat 20(2):737–761

16. Poole KT, Rosenthal H (1997) Congress: a political-economic history of roll call voting. Oxford University Press, Oxford
17. Ritov Y (1990) Asymptotic efficient estimation of the change point with unknown distributions. Ann Stat 18(4):1829–1839
18. Sigmund D (1985) Sequential analysis: tests and confidence intervals. Springer, New York
19. Stryhn H (1996) The location of the maximum of asymmetric two-sided Brownian motion with triangular drift. Stat Probab Lett 29(3):279–284
20. van der Vaart AW, Wellner JA (1996) Weak convergence and empirical processes. Springer, New York
21. Yu VL (1981) Detecting disorder in multidimensional random processes. Sov Math Dokl 23:55–59