A HIDDEN MARKOV MODEL FOR LATENT TEMPORAL CLUSTERING

HARRY CRANE

ABSTRACT. We present a new partition-valued Markov chain for modeling latent temporal clustering. The family of Markov chains we consider satisfies notable statistical properties, including exchangeability, consistency under subsampling, and reversibility with respect to a tractable class of two-parameter partition models. These properties endow our model with a robustness to missing data, choice of labels, and changes in the sample over time. When combined with an appropriate model for response data, exchangeability and consistency give rise to the stronger model properties of label equivariance and non-interference, respectively. All of these properties are desirable for statistical applications, as they make model inferences easily interpretable and straightforward. We demonstrate these and other aspects with a detailed analysis of voting data in the Supreme Court over the period 1946–2012.

1. INTRODUCTION

The hidden Markov model framework has been tailored to a wide range of statistical applications involving latent states, including speech recognition [20], computational biology [23], computer vision [6], statistical inference [8, 29], and many others [7]. Here we present a new hidden Markov chain model for latent temporal clustering. The proposed method draws

Date: October 4, 2015.

1991 *Mathematics Subject Classification.* Primary 62F10, 62M99; Secondary 62P10, 62P25.

Key words and phrases. cluster analysis; combinatorial stochastic process; hidden Markov model; Supreme Court; voting data.

on significant progress and healthy crossover between the fields of combinatorial stochastic processes and Bayesian analysis in the foregoing decade and a half.

We have three main objectives. First and foremost, we introduce a new Markov chain model for latent temporal clustering. Relevant mathematical theory for these Markov chains has been developed in recent years within the probability literature [11, 12], but it seems this model has not yet made its way into the mainstream statistical literature. We review its many fundamental statistical properties which make it amenable to a wide range of temporal clustering applications. Second, we demonstrate our model's potential for applications by analyzing a voting dataset for the U.S. Supreme Court (U.S.S.C. or 'the Court'). As ideological allegiances within the Supreme Court have been widely studied in both qualitative legal research [17, 32] and quantitative political science [24, 30, 31], we have ample information on which to verify model performance. Third, the latter Supreme Court application is of interest in its own right, and we expect some readers will enjoy our novel statistical approach to the problem. For this purpose, we have provided substantial historical context for the reader's benefit and enjoyment. Readers most interested in the statistical model will find the relevant details throughout Sections 2–4.

1.1. **Combinatorial stochastic processes.** Combinatorial stochastic processes find their origins in mathematical population genetics and the seminal work of Ewens [14] and Kingman [21, 22] on exchangeable random partitions and the coalescent theory. In the intervening years, the field has grown into separate industries within several applied and mathematical fields. Pitman [26] and others [3] developed Ewens's work into the deep mathematical field

of combinatorial stochastic processes, which lays much of the groundwork for the active area of Bayesian nonparametrics.

The initial contributions within Bayesian nonparametrics go back to Ferguson [15] and Antoniak [2], who derived Ewens's sampling formula concurrently with Ewens, but the field has grown into a juggernaut since the early work of Ishwaran and James [18, 19]. Though within the theme of combinatorial stochastic process modeling in Bayesian statistics, the present paper does not adopt the nonparametric approach that has become common. Instead, we present a parametric method, bringing to bear recent progress in the field of partition-valued Markov processes [12] which has not yet been featured in the general statistics literature. Our discussion below highlights the salient features of our proposed model, further underscoring the intimate connection between combinatorial stochastic processes and modern Bayesian statistics.

More specifically, we describe a new hidden Markov model for tracking temporal clustering over time. Though our model applies much more generally, we test it through in depth analysis of a dataset from Supreme Court voting data over the period 1946–2012. We validate various aspects of our model empirically by comparing to known qualitative and quantitative analyses of the Court over this period.

1.2. **Statistical analysis of the Supreme Court.** There have been many prior applications of statistical methods to Supreme Court data. Thurstone & Degan [31] and Sirovich [30] applied classical statistical techniques, i.e., factor analysis and principal components analysis, for the purpose of inferring the Court's ideological composition; but these methods are only applicable to short periods during which the composition of the Court remains unchanged. In an effort to identify authorship of Supreme Court

decisions, Rosenthal & Yoon [27] analyzed the frequency of certain function words in Supreme Court opinions.

For tracking temporal shifts in justices' ideological alignment, Martin & Quinn [24] estimate attitudinal changes in judges by combining an item response model with dynamic linear models for the trajectory of the ideologies. Clark & Lauderdale [9] use a spatial model to identify Supreme Court decision "locations" and predict future Supreme Court outcomes. Clinton, et al [10] also propose a spatial model for more general roll call data in legislative and judicial data sets.

In the context of the Supreme Court application, we combine our proposed hidden Markov chain with a combinatorial stochastic process model for voting alignment. Each term, justices partition into ideological classes and, given the ideological partition, the case outcomes during that term are conditionally independent and identically distributed in such a way that judges in the same class are more likely to agree on the case outcome. At the start of each term, the ideological clustering changes according to the transition probability of a reversible Markov chain. Altogether, the model affords a clear interpretation of Supreme Court data in terms of a small number of parameters, that is, a partition that represents ideological clustering of justices and scalar parameters that control deviations from this clustering. For model validation, we compare our method to what is known about the Court through prior analysis in [17, 24, 32]; see Section 5.

1.3. **Outline.** We organize the paper as follows. In Section 2, we describe the general framework for our model, present the hidden Markov model, and explain the data. In Section 3, we explain key assumptions for our Supreme Court application. In Section 4, we discuss the model within the specific context of the Supreme Court dataset. In Section 5, we discuss

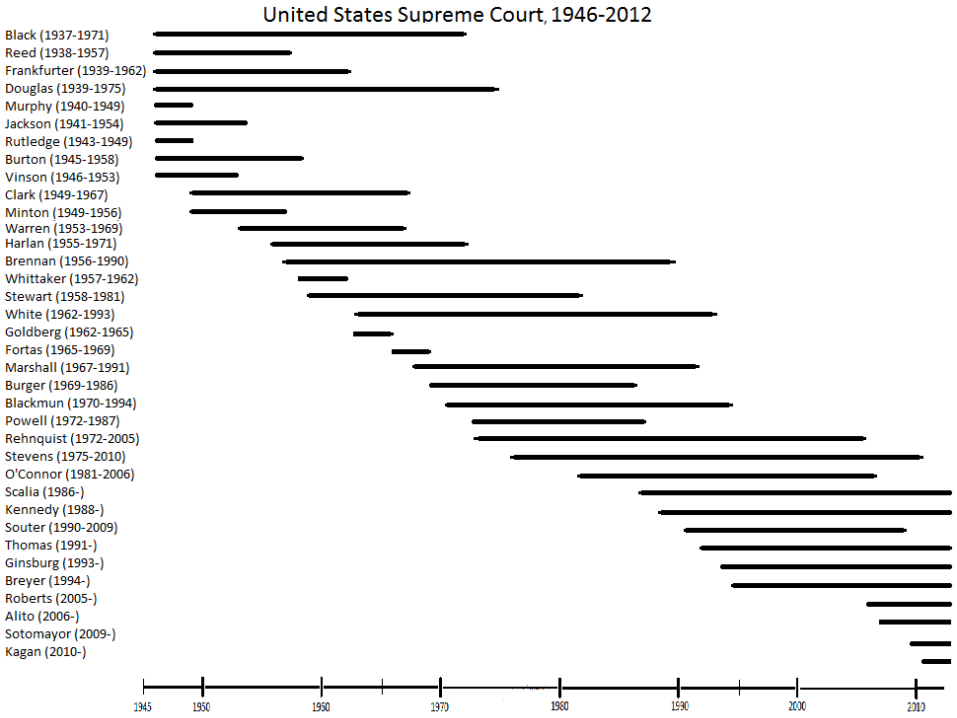

FIGURE 1. Timeline of Supreme Court justices, 1946–2012. The line next to each justice indicates his or her duration on the Court.

conclusions based on our model and compare to other analyses by Martin & Quinn [24]. In Section 6, we summarize these findings and discuss the merits of our model for future applications.

2. CONTEXT AND MOTIVATING EXAMPLE

Below we introduce a general partition-valued Markov chain for modeling latent clustering structures that commonly arise in statistical applications with data of the following form. We assume a population $\mathcal{U}$ of statistical units from which we observe a finite sample $\mathcal{S}_t \subset \mathcal{U}$ at an observed sequence of times $t = 0, 1, \ldots$. (Note that we allow the sample to vary over time, as it does in the Supreme Court application below.)

For every $t = 0, 1, \ldots$, the sample $\mathcal{S}_t$ clusters into non-overlapping subsets $B_t = \{B_{t,1}, \ldots, B_{t,k}\}$, where $B_{t,1}, \ldots, B_{t,k}$ are non-empty, disjoint and satisfy $\bigcup_{1 \le i \le k} B_{t,i} = \mathcal{S}_t$. Given B_t, the response $Y_t = (Y_{t,1}, Y_{t,2}, \ldots) \in \mathcal{R}^{\mathcal{S}_t}$ is generally an array of observations that are dependent by virtue of the underlying partition. The most generic dataset of this form is structured as

$$\begin{array}{ll} \text{latent clustering} & \begin{array}{ccccc} B_0 & B_1 & B_2 & \cdots & B_T \end{array} \\ \text{response} & \left(\begin{array}{ccccc} Y_0 & Y_1 & Y_2 & \cdots & Y_T \end{array} \right), \end{array} \tag{1}$$

where $(B_t)_{t=0,\ldots,T}$ represent a temporally-varying clustering of the samples $(\mathcal{S}_t)_{t=0,\ldots,T}$, respectively, and $(Y_t)_{t=0,1\ldots,T}$ is the sequence of response data. A typical special case has the form

$$\begin{array}{c} \\ Y^1 \\ Y^2 \\ \vdots \\ Y_n \end{array} \begin{array}{c} \begin{array}{ccccc} B_1 & B_2 & B_3 & \cdots & R_T \end{array} \\ \left(\begin{array}{ccccc} Y_1^1 & Y_2^1 & Y_3^1 & \cdots & Y_T^1 \\ Y_1^2 & Y_2^2 & Y_3^2 & \cdots & Y_T^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_1^n & Y_2^n & Y_3^n & \cdots & Y_T^n \end{array} \right), \end{array} \tag{2}$$

with a separate measurement Y_t^i for each unit $i = 1, \ldots, n$ at each time $t = 1, \ldots, T$. In (2), we write $Y^i = (Y_1^i, \ldots, Y_T^i)$ to denote the vector of responses for unit $i = 1, \ldots, n$.

The distinction between (1) and (2) is that the response data generally need not separate in terms of an observation $Y^u = (Y_t^u)_{t=0,1,\ldots,T}$ for each unit u. A common example of this is network data, where the response Y_t records binary relationships among all individuals in the sample. We encounter another example below in the form of partition data derived from Supreme Court voting data over an extended period of U.S. history. From here on we shall discuss the model within the concrete context of the illustrative application to the U.S. Supreme Court. The Supreme Court data set takes

on a special structure in the form of categorical observations for each case outcome; however, we emphasize that our hidden Markov model is not confined to this setup provided a viable class of conditional distributions for the response given the clustering is given.

2.1. **Hidden Markov model for latent partitions.** Our main objective is to introduce the following Markov chain for modeling latent partition structures that change over time. Partition models arise throughout the statistics literature on clustering, e.g., [5, 16, 25], and the model we present here is closely related to the so-called Ewens–Pitman two-parameter family [14, 26], as we explain below.

Formally, we define a *partition* of $A \subset \mathbb{N}$ as a collection $\pi = B_1/\cdots/B_r$ of non-empty, disjoint subsets, called *blocks*, *clusters*, or *classes*, for which $\bigcup_{i=1}^{r} B_i = A$. The blocks of π are unlabeled and unordered, but we adopt the convention of listing them in increasing order of their smallest element. For example, there are five partitions of $\{1,2,3\}$:

$$\{1,2,3\}, \quad \{1\}/\{2,3\}, \quad \{1,2\}/\{3\}, \quad \{1,3\}/\{2\}, \quad \text{and} \quad \{1\}/\{2\}/\{3\}.$$

We write $\mathcal{P}_A$ to denote the set of partitions of A, so that $\mathcal{P}_{[n]}$ is the set of partitions of $[n] = \{1,\ldots,n\}$.

For $A \subseteq [n]$, the *restriction* of $\pi = B_1/\cdots/B_r$ in $\mathcal{P}_{[n]}$ to a partition of A is

$$\pi_{|A} = B_1 \cap A/\cdots/B_r \cap A, \tag{3}$$

with empty sets removed. For example, $\pi = \{1,3,5,6\}/\{2,4\}/\{7,8\}$ restricted to $\{1,2,3,4\}$ is $\pi_{|[4]} = \{1,3\}/\{2,4\}$. Statistically, the restriction operation relates to subsampling. We discuss its importance throughout Section 4.

Fix $k^* \geq 2$ and $\beta > 0$. We define a Markov chain $(B_t)_{t=0,1,\ldots}$ on partitions of $[n]$ to have transition probabilities

$$\mathbb{P}\{B_{t+1} = \pi' \mid B_t = \pi\} = k^{*\downarrow \#\pi'} \prod_{b \in \pi} \frac{\prod_{b' \in \pi'} (\beta/k^*)^{\uparrow \#(b \cap b')}}{\beta^{\uparrow \#b}}, \tag{4}$$

where $k^{*\downarrow j} := k(k-1)\cdots(k-j+1)$, $\beta^{\uparrow j} := \beta(\beta+1)\cdots(\beta+j-1)$, $\#\pi'$ is the number of blocks of π', and $\#b$ is the cardinality of a block $b \in \pi$. These transition probabilities are reversible with respect to

$$\mathbb{P}\{B_t = \pi\} = k^{*\downarrow \#\pi} \frac{\prod_{b \in \pi} \beta^{\uparrow \#b}}{(k^*\beta)^{\uparrow n}}. \tag{5}$$

Reversibility is easily seen by verifying that (4) and (5) satisfy the detailed balance condition:

$$\mathbb{P}\{B_{t+1} = \pi' \mid B_t = \pi\}\mathbb{P}\{B_t = \pi\} = \mathbb{P}\{B_{t+1} = \pi \mid B_t = \pi'\}\mathbb{P}\{B_t = \pi'\}.)$$

The distribution in (5) is exactly the Ewens–Pitman two-parameter distribution with parameter $(-\beta, k^*\beta)$; see [26]. The transition probabilities in (4) were discovered in [11], but the model has not yet been applied to the problem of temporal clustering we consider here.

The distribution in (5) is also sometimes called the Dirichlet-Multinomial distribution on partitions because of its clear interpretation in terms of multinomial sampling as follows. To generate B_t as in (5), we first generate P from the k^*-parameter symmetric Dirichlet distribution with parameter $(\beta, \ldots, \beta)$. Given $P = (p_1, \ldots, p_{k^*})$, we let $X_1, X_2, \ldots$ be conditionally independent, identically distributed (i.i.d.) with distribution $\mathbb{P}\{X_i = j \mid P\} = p_j$ for each $j = 1, \ldots, k^*$. We then define B_t by putting i and j in the same block provided $X_i = X_j$.

The transition probability in (4) has a similar, though more involved, interpretation. Let $B_t = \pi$ be given and now let $(P_b)_{b \in \pi}$ be i.i.d. Dirichlet

random variables with parameter $(\beta/k^*, \ldots, \beta/k^*)$. Within each block $b \in \pi$, we let $(X_i)_{i\in b}$ be independent from $\mathbb{P}\{X_i = j \mid P_b\} = P_{b,j}$ just as before. (The difference is that the P_b are different for each block.) We obtained B_{t+1} by putting i and j in the same block if $X_i = X_j$.

This class of Markov chains has many remarkable statistical properties. Most important for its application in hidden Markov modeling are its exchangeability and consistency properties. Under exchangeability, the distribution of the Markov sequence $(B_t)_{t=0,1,\ldots}$ is unchanged by arbitrary relabeling of elements according to any permutation of $[n]$. Consistency implies that the restriction of $(B_t)_{t=0,1,\ldots}$ to a sequence of partitions of $[m] \subset [n]$, $m \leq n$, maintains the Markov property and the transition probabilities of the restricted chain $(B_{t|[m]})_{t=0,1,\ldots}$ are given by (4). This latter property is in general hard to satisfy for exchangeable Markov chains, but it plays an important role in statistical applications involving missing data and/or temporal variation in the observed sample. (The Supreme Court dataset below exhibits both.)

We discuss many more aspects of this model later within the specific context of our motivating example of Supreme Court voting data, which we now describe.

2.2. **Supreme Court data.** The Washington University of St. Louis maintains a database with detailed information for every Supreme Court decision since 1946.[1] We obtained the voting alignment for all 8,561 cases in the database. Over this period, 36 justices served, 9 at a time, and the number of cases per term varied from a high of 198 in 1967 to a low of 73 in 2007.

For every case, each active justice either *concurs* with or *dissents* from the majority opinion of the Court, or is *recused* and does not vote. We represent the voting alignment for each case as a length-9 vector with +1 indicating

[1] http://scdb.wustl.edu

concurrence with the majority, −1 indicating dissent, and NA indicating recusal. For example, the vector corresponding to the 1946 case *United States v. Tillamooks Et Al.* is

$$(+1, -1, +1, +1, +1, \text{NA}, -1, -1, +1), \tag{6}$$

where entries correspond to the justices in order of seniority. Specifically, the vector in (6) signifies a 5–3 decision, with Justice Jackson recused, Justices Black, Frankfurter, Douglas, Murphy, and Vinson in the majority, and Justices Reed, Rutledge, and Burton in the minority. Our encoding neither records which side the majority favored nor distinguishes between justices who agreed with the decision but not its reasoning. (This latter omission is supported by other authors, even former Chief Justice William Rehnquist, who have dismissed the importance of differing opinions and believe that only the outcome of a case matters [32]. It also agrees with prior analyses of the Court [30].) For cases with more than one part, we only use the voting alignment on the first part of the case as it is listed in the database. Otherwise, we analyze the data *as is*.

Formally, we label justices in order of seniority, i.e., we assign label 1 to Hugo Black (1937–1971) and label 36 to Elena Kagan (2010–). We label cases according to term $t = 1946, \ldots, 2012$ and then in chronological order $i = 1, 2, \ldots$ within each term; thus, $x_{t,i} = (x_{t,i}^1, \ldots, x_{t,i}^{36})$ encodes the votes of the justices on the ith case in term t, with

$$x_{t,i}^j = \begin{cases} +1, & \text{Justice } j \text{ concurs with the majority on case } (t,i), \\ -1, & \text{Justice } j \text{ dissents from the majority on case } (t,i), \\ \text{NA}, & \text{Justice } j \text{ is recused from or inactive for case } (t,i). \end{cases} \tag{7}$$

For each $t = 1946, \ldots, 2012$, we write $x_t = (x_{t,i})_{i \geq 1}$ to denote the aggregate collection of vectors for term t. Figure 1 is a timeline of justices who served

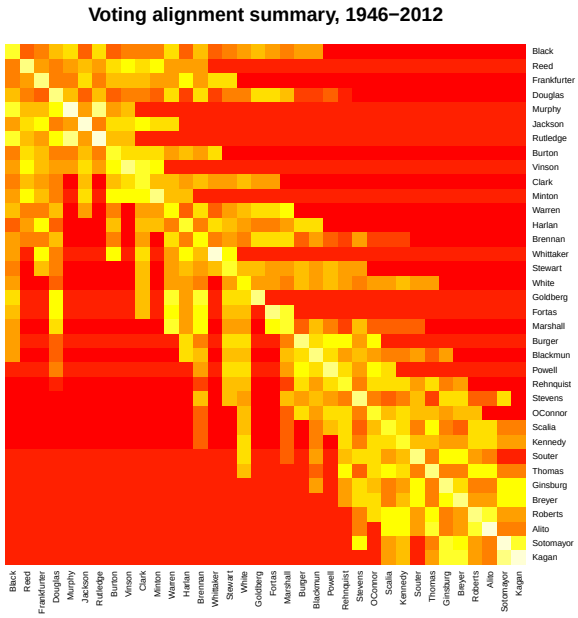

FIGURE 2. A heatmap of the thirty-six justices (1946–2012), with white indicating 100% agreement and red indicating 0% agreement; however, red also denotes justices who were not judicial contemporaries and, therefore, never voted on the same case.

between 1946 and 2012 and Figure 2.2 summarizes the raw voting data for the period under study.

3. INDUCED PARTITION SEQUENCE

We search for structural relationships among justices based on their voting patterns. For a given case, the relevant datum for detecting ideological similarity between two justices is whether they agreed on its outcome, and so we can further simplify the data $(x_{t,i})_{t=1946,\ldots,2012;\, i\geq 1}$ by considering only the *partition sequence* it induces.

For $t = 1946, \ldots, 2012$ and $i = 1, 2, \ldots$, let $C_{t,i} = \{j = 1, \ldots, 36 : x_{t,i}^{j} \neq \text{NA}\}$ be the set of ruling justices on case (t, i). Each vector $x_{t,i}$ determines a

partition $\pi_{t,i}$ of $C_{t,i}$ with

$$j \text{ and } j' \text{ in the same block} \quad \textit{if and only if} \quad x_{t,i}^{j} = x_{t,i}^{j'}. \tag{8}$$

Applying (8) to every case yields a collection of partitions $(\pi_{t,i})_{t=1946,\ldots,2012;\, i\geq 1}$. For example, the *Tillamooks* case in (6) has $C_{t,i} = \{1,2,3,4,5,7,8,9\}$ and $\pi_{t,i} = \{1,3,4,5,9\}/\{2,7,8\}$. Notice that 6, which labels recused Justice Jackson, does not appear in $\pi_{t,i}$. For a given term t, we write $C_t = \bigcup_{i\geq 1} C_{t,i}$ and $\pi_t = (\pi_{t,i})_{i\geq 1}$ as the aggregate set of active judges and induced partitions, respectively, for that term.

3.1. **Unanimous cases.** The plurality of Supreme Court cases are decided unanimously.[2] The prevalence of unanimous decisions dampens the signal of any ideological separation and challenges the efficacy of statistical methods. To better distinguish between justices, previous authors [30, 31] remove all unanimous cases prior to analysis. Other models, e.g., [24], remove unanimous cases because they make no contribution to the likelihood. In either case, removal of these cases may have unintended consequences for analyses that span long periods of time, as exclusion of these cases can invalidate comparisons between estimates for different judicial terms. Intuitively, the proportion of unanimously decided cases varies over time and, therefore, removing these cases affects estimates from different years in different ways. We eliminate this issue by retaining unanimous decisions in the data set.

3.2. **Missing-at-random.** In our analysis, we treat missing observations as *missing-at-random*, i.e., missing votes occur independently of the data generating process. Naturally, justices who were inactive during a given term could not have voted during that term and we reasonably assume

[2]For the period we studied, 3,505 out of 8,561 cases were decided unanimously.

their ideologies have no effect on the voting behavior of active justices. On the other hand, votes for justices who were active for a term but recused from a specific case could be correlated with their underlying relationship to other justices. Recusals typically indicate that a justices' disposition favors one side of the case over another. For a given recused justice, there is no prior reason to believe he or she is more likely to side with the majority or the minority, but two justices recused from the same case may be more likely to share the same ideology. For a case with only one justice recused, we do not know which side he or she favored and, therefore, can safely assume the missing vote is independent of the data generating process. This missing-at-random assumption is likely violated when two or more justices are recused; but such cases are so rare that they should not affect the inference.

4. Hidden Markov binary cluster model

For each term t, we assume active justices C_t cluster according to a partition B_t and, given B_t, the partition sequence $\Pi_{t,1}, \Pi_{t,2}, \ldots$ induced by case outcomes is conditionally independent and identically distributed. Between terms t and $t+1$, the ideological partition B_t changes to B_{t+1} according to the dynamics of a stationary, reversible Markov chain on partitions. Because the sets of active justices C_t and C_{t+1} can be different, B_t and B_{t+1} may not lie in the same state space. To handle this, we model $(B_t)_{t=1946,\ldots,2012}$ as the restriction of a Markov sequence $(\tilde{B}_t)_{t=1946,\ldots,2012}$ on the space of partitions of all 36 justices, active and inactive. Specifically, we model $(\tilde{B}_t)_{t=1946,\ldots,2012}$ as the Markov chain with initial distribution (5) and transition probabilities (4) for some choice of parameters β and k^*, discussed below.

Because the set of active justices sometimes varies between terms, we specify our model for $(B_t)_{t=1946,\ldots,2012}$ by first defining a process for the sequence $(\tilde{B}_t)_{t=1946,\ldots,2012}$ of partitions for all 36 justices. For example, $\tilde{B}_{1946}$ includes justices active in 1946 as well as Justices Kagan, Sotomayor, Alito, etc., who were not yet active. This technicality avoids the unattractive prospect of defining a Markov chain for $(B_t)_{t=1946,\ldots,2012}$ directly, which would require transition probabilities between different state spaces. At any rate, the non-interference property for $(\tilde{B}_t)_{t=1946,\ldots,2012}$ induces a well-defined model for $(B_t)_{t=1946,\ldots,2012}$ that allows us to ignore unobserved aspects of $(\tilde{B}_t)_{t=1946,\ldots,2012}$ in our analysis. In other words, we identify all justices with an ideological cluster, even for terms they are inactive, but we only observe information for active justices.

While assuming $\tilde{B}_t$ partitions both sampled and unsampled individuals may be perceived as a purely philosophical, rather than practical, matter of statistical modeling for any given term t, this property is critical in the case of temporally-varying partitions for which the observed sample varies with time. As a concrete example, suppose $C_t = \{1, 2, 3\}$, $C_{t+1} = \{2, 3, 4\}$, $B_t = \{1, 2\}/\{3\}$, and we want to model B_{t+1}, which partitions C_{t+1}. An intuitive approach would be impute the membership of element 4 in B_t according to an appropriate conditional distribution, use the imputed partition B_t^* to generate B_{t+1}^*, and then obtain B_{t+1} by removing element 1 from B_{t+1}^*. Without non-interference, the distribution of B_{t+1} obtained in this way will depend on the imputed partition B_t^*; with non-interference, it depends only on B_t, as desired. We discuss this in further detail below.

Importantly, the sampling consistency property of $(\tilde{B}_t)_{t=1946,\ldots,2012}$ endows our model with a special *non-interference property*, by which the conditional distribution of $\Pi_{t,i}$, given $\tilde{B}_t$, depends only on B_t. Consequently, the posterior distribution of $(B_t)_{t=1946,\ldots,2012}$, given data $(\Pi_{t,i})_{t=1946,\ldots,2012;\, i\geq 1}$, does not

depend on the hypothetical voting data for inactive justices. Ultimately, these non-interference properties protect against spurious conclusions. The hidden Markov binary cluster model below satisfies non-interference and several other key properties. We highlight several of these properties and point the reader to the appropriate literature for further details.

4.1. **Model for induced partition sequence.** Given a clustering B_t of active justices C_t in term t, we assume $\Pi_{t,1}, \Pi_{t,2}, \ldots$ are conditionally independent and identically distributed according to

$$\mathbb{P}\{\Pi_{t,i} = \pi \mid B_t,\, \alpha_t\} \propto 2 \prod_{b \in B_t} \frac{\prod_{b' \in \pi} (\alpha_t/2)^{\uparrow \#(b \cap b')}}{\alpha_t^{\uparrow \#b}}, \tag{9}$$

where $\alpha_t > 0$ is a real-valued parameter, $\alpha_t^{\uparrow j} = \alpha_t(\alpha_t + 1) \cdots (\alpha_t + j - 1)$ with convention $\alpha_t^{\uparrow 0} \equiv 1$, $\#b$ is the cardinality of b, and π is a partition of $C_{t,i}$ with at most two blocks. (Remember the blocks of π divide justices into majority and minority classes, of which there are at most two for each case.)

The partition parameter B_t has a clear interpretation as a clustering of the sample, which is the primary object of interest and is discussed in much more depth below. The scalar parameters $\alpha_t > 0$ also have a clear interpretation as variance parameters, in a sense we make precise below.

Note that the distribution in (9) is identical in spirit to the transition probabilities in (4) in the special case $k^* = 2$, with the difference that we will model α_t as random in our hidden Markov framework and we view B_t as a parameter in (9) rather than a state of a Markov chain in (4). Conditional on α_t and B_t, the interpretation of (9) in terms of the Dirichlet-Multinomial sampling from Section 2.1 remains valid. From a statistical viewpoint, (9) affords the further interpretation to observed partitions π as *noisy observations* of B_t, in the sense made precise by the Dirichlet-Multinomial construction. That description also lays bare the within- and between-cluster behavior

of elements, namely, behavior within clusters is positively correlated and behavior across clusters is independent.

Equation (9) satisfies several other important statistical properties. First, as we are interested in detecting a meaningful clustering, (9) gives a reasonable, in fact widely used, null distribution for the outcomes $\Pi_{t,i}$. In the event that $B_t = \mathbf{1}_{C_t}$ places all justices in the same block, (9) becomes

$$\mathbb{P}\{\Pi_{t,i} = \pi \mid \alpha_t,\, B_t = \mathbf{1}_{C_t}\} = 2\frac{\prod_{b\in\pi}(\alpha_t/2)^{\uparrow \#b}}{\alpha_t^{\uparrow \#C_{t,i}}}.$$

This distribution is commonly referred to as the Beta-Binomial process with parameter $(\alpha_t/2, \alpha_t/2)$. It is a special case of the Dirichlet-Multinomial process which acts as a precursor to the Ewens's sampling formula; see [13, 26] for a more thorough account. Also notice that the conditional distribution of $\Pi_{t,i}$ given B_t depends only on the block sizes of B_t and the sizes of the intersections of blocks of $\Pi_{t,i}$ with B_t; thus, the labeling of justices plays no role, a property known as *label equivariance*. Under label equivariance, the conditional probability of the event $\{\Pi_{t,i} = \pi\}$ given B_t equals the conditional probability of $\{\Pi_{t,i} = \pi^\sigma\}$ given B_t^σ for all permutations σ, where π^σ (similarly, B_t^σ) is the partition obtained by relabeling the elements according to σ:

$$i \text{ and } j \text{in same block of } \pi^\sigma \quad \text{if and only if} \quad \sigma^{-1}(i) \text{ and } \sigma^{-1}(j) \text{ in same block of } \pi.$$

Note the difference with exchangeability, under which the conditional probabilities of $\{\Pi_{t,i} = \pi\}$ and $\{\Pi_{t,i} = \pi^\sigma\}$ are equal without any corresponding change to the parameter.

Second, by the convention $\alpha_t^{\uparrow 0} = 1$, $\Pi_{t,i}$ need only partition a subset of C_t; the block assignment in B_t of elements not in $C_{t,i}$ plays no role in the probability assignment. As statistical applications almost always entail

some degree of missing observations, such a property is crucial. This comes into play in the Supreme Court application as it permits us to ignore fit the model to cases with recused justices without issue. Note also that when $C_t = C_{t,i}$, the proportionality relation in (9) becomes equality.

Third, $\alpha_t > 0$ plays the role of a variance parameter: justices in the same block of B_t appear in the same block of $\Pi_{t,i}$ with probability $(\alpha_t+2)/(2\alpha_t+2) > 1/2$ and justices in different blocks of B_t appear in the same block of $\Pi_{t,i}$ with probability $1/2$. To see this, we appeal to the label equivariance and non-interference properties of (9): for any given elements j and j', label equivariance allows us to treat them as $j = 1$ and $j' = 2$ as long as we relabel B_t accordingly. Denoting the relabeling B_t as B'_t, the non-interference property implies that the restriction of $\Pi_{t,i}$ in (9) to $\{1, 2\}$ depends only on the restriction of B'_t to $\{1, 2\}$, namely, whether or not 1 and 2 are in the same block of B_t. The above probability calculations proceed directly from (9).

The lower bound marginal probability $1/2$ when elements are in different blocks has a clear and intuitive interpretation in terms of the explanation following (9). Since elements in different blocks behave independently and the blocks of each $\Pi_{t,i}$ have no additional labels, it follows that the probability that the probability of agreement is up to random chance. It is mathematically impossible to achieve anything lower than this in any model, as it is well known that the correlation of infinitely exchangeable sequences cannot go below 0 (as is the case when the sequence is i.i.d.) [1].

Returning to the pairwise agreement probability $(\alpha_t + 2)/(2\alpha_t + 2)$ for elements in the same block of B_t, we obtain a clear interpretation of the α_t parameters. Specifically, $\alpha_t \approx 0$ corresponds to strong alliance within blocks and $\alpha_t \approx \infty$ corresponds to approximate independence of all justices. Moreover, by allowing α_t to vary across different terms, we account for different behavior of the Court over time: if the justices vote consistently

according to their ideology, α_t is small; if the ideological alignment of justices is weak, α_t is large. Furthermore, notice that the probability in (9) is maximized at $\pi = \mathbf{1}_{C_{t,i}} = \{C_{t,i}\}$, the partition corresponding to a unanimous decision, and $\pi = B_{t|C_{t,i}}$, the outcome if all justices vote strictly along ideological lines. This property fits the typical outcome of most Supreme Court cases—justices either vote unanimously or divide according to ideology.

In the full model for $(\tilde{B}_t, \Pi_t)_{t=1946,\ldots,2012}$, we assume the term-specific variance parameters $(\alpha_t)_{t=1946,\ldots,2012}$ in (9) are independent and identically distributed according to a distribution μ on $(0, \infty)$.

Definition 4.1 (Hidden Markov binary cluster model)**.** *With $\beta > 0$, $k^* \geq 2$, and μ a probability distribution on $(0, \infty)$, we define the* hidden Markov binary cluster model *with parameter (μ, β, k^*) as the distribution of $(\alpha_t, \tilde{B}_t, \Pi_t)$ generated by*

$$\begin{aligned} (\alpha_t) &\overset{\text{i.i.d.}}{\sim} \mu \\ (B_t) &\sim \textit{ Markov chain from (5) and (4)} \\ \Pi_t \mid \alpha_t, B_t &\overset{\text{i.i.d.}}{\sim} \textit{ distribution in (9), for each } t. \end{aligned}$$

By reversibility, $(\tilde{B}_t)_{t=1946,\ldots,2012}$ is stationary and, thus, the marginal distribution of each $\tilde{B}_t$ is exactly (5). Thus, the marginal distribution of $(\alpha_t, \tilde{B}_t, \Pi_t)$ is the same for all t. Since a primary objective of our analysis is to detect changes in the clustering over time, we feel stationarity is a critical property of our model. We assume stationarity so that we can confidently interpret inferred change points as physically meaningful. In the absence of this property, we are open to the possibility that our model implicitly assumes a very different marginal prior distribution for different terms, which could affect inference. Without stationarity, the prior marginal distribution of $(\tilde{B}_t, \Pi_t)$ in year, say, 1992 would vary depending on our arbitrary decision

to begin analysis from 1946, 1947, 1948, etc. While it is arguable that the clustering of the Court evolves in a non-stationary way, we incorporated no prior information about this into our model. Furthermore, reversibility reflects the usual assumption that, while time invariably moves forward, the data contains the same information and determines the same posterior distribution for $(\alpha_t, B_t)_{t=1946,\ldots,2012}$ whether viewed forwards or backwards in time.

Moreover, (5) exhibits non-interference so that the clustering of active justices B_t also satisfies (5) with $n = \#C_t$, the number of active justices for term t, and the conditional distribution of B_{t+1}, given $\tilde{B}_t$, depends only on the restriction of $\tilde{B}_t$ to C_{t+1}. See [11] for a formal description of non-interference, label equivariance, and reversibility for $(\tilde{B}_t)_{t=1946,\ldots,2012}$.

4.2. **Posterior distribution.** The hidden Markov binary cluster model determines a joint distribution for $(\alpha_t, \tilde{B}_t, \Pi_t)_{t=1946,\ldots,2012}$, which depends on the distribution μ of $(\alpha_t)_{t=1946,\ldots,2012}$ and fixed parameters $\beta > 0$ and $k^* \geq 2$ as in (4) and (5). We use Bayes's rule to compute the posterior distribution of $(\alpha_t, B_t)_{t=1946,\ldots,2012}$ by combining the probabilities of the underlying Markov chain with the prior distribution for $(\alpha_t)_{t=1946,\ldots,2012}$ and the conditional distribution of the observed data. We base our inference on the posterior model of these parameters by optimizing the objective function which is proportional to the posterior distribution. The optimization is somewhat involved as we can only compute the posterior up to a normalizing constant and finding the optimal clustering entails a search over all partitions of the justices. We discuss this procedure in further detail in Section 4.3.

In our analysis below, we take μ to be the standard Exponential distribution $\mathbb{P}\{\alpha_t \in da\} = e^{-a} I_{\{a>0\}} da$, $k^* = 2$, and $\beta = 1/2$. We stress that our inference

is not sensitive to minor changes to these prior assumptions. Empirically, moderate choices of β in the range $(1/10, 5)$ do not alter inferences.

4.3. **Finding the optimal clustering.** We estimate the optimal ideological clustering sequence and its term-specific variance parameters by the posterior mode of $(\alpha_t, B_t)_{t=1946,\ldots,2012}$. For each term, B_t is a partition with at most k^* blocks and, thus, lies in a state space with roughly k^{*9} elements. We find that the estimates of $(B_t)_{t=1946,\ldots,2012}$ are robust to the choice of $k^* \geq 2$. In particular, choosing $k^* = 3$ instead of $k^* = 2$ makes no substantial difference in the inferred ideological partition sequence. The space of partitions of nine elements with at most $k^* = 2$ and $k^* = 3$ blocks has 256 and 3,281 partitions, respectively, and can be searched exhaustively with R on a laptop computer.

The inferred parameters for $(\alpha_t, B_t)_{t=1946,\ldots,2012}$ shown below were obtained by an initial sequential optimization step followed by a random search algorithm. Here is where the reversibility of the model for $(\alpha_t, B_t)_{t=1946,\ldots,2012}$ plays a role in the optimization procedure. We begin with just the voting data from 1946, and we optimize the posterior distribution of $(\alpha_{1946}, B_{1946})$ with respect to the data from 1946 based on the model in Definition 4.1. With $k^* = 2$ or $k^* = 3$, we can do this optimization without problem since the space of partitions of nine justices is manageable for exhaustive search (as mentioned above), where we initialize by $\alpha_{1946} = 1$. Given the estimated $\hat{B}_{1946}$, we then find the optimal $\hat{\alpha}_{1946}$. which proceeds easily since this distribution is uni-model in α. We then iterate back to find update $\hat{B}_{1946}$ based on this choice of $\hat{\alpha}_{1946}$, and so on until convergence. Convergence required only a few steps for this dataset.

We continue by estimating (α_t, B_t) on year at a time, at each step incorporating the information of the previous year's estimates $(\hat{\alpha}_s, \hat{B}_s)_{s=1946,\ldots,t-1}$.

The only key difference in subsequent years is that the prior distribution of B_t is taken as the distribution $\mathbb{P}\{\tilde{B}_t = \cdot \mid \tilde{B}_{t-1} = \hat{B}_{t-1}\}$. Given this, the optimization goes as above by iterating back and forth between B_t and α_t until convergence.

After one run through this procedure, we have estimated $(\hat{\alpha}_t, \hat{B}_t)_{t=1946,\ldots,2012}$; however, we must ensure that it does not depend on our choice to only account for "past" information about the Court. For verification, we run the same optimization scheme backwards, starting at the 2012 term and continuing until the 1946 term. The reversibility of the model makes the estimates robust to this choice. Finally, we go back and randomly select years and consider optimizing the parameters α_t, B_t for a fixed term t in the presence of all other estimates $(\hat{\alpha}_s, \hat{B}_s)_{s\neq t}$. The estimates obtained from this procedure are given throughout Section 5.

For each term, the posterior distribution is strongly peaked at the mode. To quantify uncertainty in the posterior of $(B_t)_{t=1946,\ldots,2012}$, we compute the *co-clustering matrix* $K_t = (K_t(j, j'))_{j,j'=1,\ldots,36}$, where $K_t(j, j')$ is the posterior probability that justices j and j' are in the same cluster of B_t for each t. Moderate values of $K_t(j, j')$ indicate some degree of ambivalence in the ideology of one or both of j or j', typically indicating a moderate viewpoint. For most pairs of justices j, j' and most terms t, $K_t(j, j')$ is very close to 0 or 1, indicating that justices do have well delineated ideologies. In some instances we observe moderate values for $K_t(j, j')$. For example, during her final term, Sandra Day O'Connor's co-clustering probabilities were approximately 0.34 with the liberal justices and 0.66 with the conservative justices, justifying her classification as a swing vote. It is not possible to present all the information in these co-clustering matrices for the sixty-seven years under study. Instead, the co-clustering matrices in Figures 4 and 5 show the minimum and maximum co-clustering probabilities for all justices

who served together on at least one case. Uncertainty of the overall posterior distribution with respect to these relationships is seen by comparing the maximum and minimum for any given pair of justices. For pairs with a maximum of 0 or a minimum of 1, the model is practically certain of their cluster membership relative to one another: the two are surely in different clusters in the former case and surely in the same cluster in the latter case. Non-extreme values, however, provide a measure of uncertainty in the inferred clustering, which is quite informative and meaningful with respect to the specific application we consider.

5. Analysis

5.1. **Fixed clustering model.** As an alternative to temporal dependence, we could assume that a single partition describes judicial voting behavior over the entire sixty-seven year period. Thus, instead of modeling a sequence of ideological partitions $(B_t)_{t=1946,\ldots,2012}$, we assume a single partition B comes from the distribution in (5) with a suitable choice of k^*. Given B, the data sequence $(\Pi_{t,i})_{t=1946,\ldots,2012;\, i\geq 1}$ over all terms is conditionally independent and identically distributed as in (9). For this we take $k^* = 5$; we discuss the prospect of different choices below.

Non-interference plays a key role. Many judges, e.g., Hugo Black (1937–1971) and Elena Kagan (2010–), did not rule on a single case together and, thus, Black's ideology should not affect Kagan's voting behavior and vice versa. Also, since a single partition describes judicial voting behavior over all times, we regard all NA entries as data that is missing at random. This assumption, though likely violated, is implicit in previous such analyses, e.g., Martin & Quinn [24, Section 5.1]. Together, non-interference and the missing-at-random assumption imply that the outcome of case (t, i) depends on B only through its restriction to the active justices $C_{t,i}$ for that case. Fitting

Cluster	Justices
Activist	Black, Douglas, Murphy, Rutledge, Warren, Brennan, Goldberg, Fortas, Marshall
Liberal	Stevens, Souter, Ginsburg, Breyer, Sotomayor, Kagan
Blackmun	Blackmun
Conservative	Reed, Burton, Vinson, Clark, Minton
Judicial Restraint	Frankfurter, Jackson, Harlan, Whittaker, Stewart, White, Burger, Powell, Rehnquist, O'Connor, Scalia, Kennedy, Thomas, Roberts, Alito

TABLE 1. Five ideological clusters obtained by the posterior mode in the fixed clustering model of Section 5.1. We append names to each of the clusters for ease of reference in our discussion.

the above model with $\beta = 1/2$ and $k^* = 5$ to the aggregate data for 1946–2012, we estimate B by the posterior mode in Table 1.[3] Our choice of $k^* = 5$ reflects a balance between computational tractability and informativeness of the estimate. Choosing $k^* = 2$, as we do when considering term-by-term clusterings, would result in a very coarse picture of the Supreme Court, similarly for $k^* = 3$. The choice $k^* = 4$ results in the clustering of Table 1 with Blackmun grouped in the Liberal block. On the higher end, we allowed for $k^* = 6$ and the clustering in Table 1 with 5 blocks remained optimal.

With the notable exception of Harry Blackmun, the partition in Table 1 aligns with our expectations. The *Activist* and *Liberal* clusters contain well-known left-wing judges and the *Conservative* and *Judicial Restraint* clusters contain right-wing judges. But while this separation is plausible along purely ideological lines, i.e., liberal and conservative, we might not have expected the left- and right-wing factions to divide as they have. In particular, *Judicial Restraint* consists of traditional conservatives, such as Harlan and Rehnquist, moderates, such as Powell, Stewart, White, O'Connor,

[3]As mentioned above, our analysis is not sensitive to this choice of β. We infer the same ideological clustering sequence for all moderate values of $\beta \in [0.10, 2.00]$ and expect the same inference to persist outside this range.

and Kennedy, and originalists, such as Frankfurter, Scalia, and Thomas. Among these, only White was nominated by a Democratic president (John F. Kennedy).

The liberal clusters, on the other hand, exhibit strong temporal clustering: every *Activist* judge joined the Court before every *Liberal* judge. John Paul Stevens (1975–2010) is the only *Liberal* judge to overlap with anyone in the *Activist* group and, therefore, acts as the only bridge between the *Liberal* and *Activist* clusters. As implausible as this clustering may seem, it is consistent with the Martin–Quinn scores from the Constant Ideal Point Model [24, Section 5.1]. In Martin & Quinn's analysis, all *Activist* judges have a score to the left of -1.024 (Goldberg) and all *Liberal* judges have a score to the right of -0.555 (Stevens).[4]

While the *Conservative* and *Judicial Restraint* classes do mingle justices from different eras—notably Frankfurter (1939–1962) and Jackson (1941–1954) inhabit the same *Judicial Restraint* cluster as modern-day conservatives Scalia, Thomas, Roberts, and Alito—the *Conservative* class only contains the four Truman appointees and Stanley Forman Reed (1938–1957). Aside from Clark, Truman's appointees have been widely criticized for their jurisprudence [28], while Reed is the last Supreme Court justice to have not graduated from law school. These observations hint at the possibility of a significant difference between the *Conservative* and *Judicial Restraint* clusters that cannot be captured by the one-dimensional Martin–Quinn scores.

Blackmun's isolation can easily be misconstrued as moderation. Overall, Blackmun's ideology averages out to the middle, as supported by the

[4]Note that Martin & Quinn's original analysis only considered cases in the years 1953–1999, which do not include Justices Murphy, Rutledge, Kagan, or Sotomayor. Moreover, based on the sign of his score, Souter would be classified as conservative before 1999. We focus on Souter more closely in Section 5.2.5.

inferred clustering in Table 1 as well as Martin & Quinn [24], who estimate Blackmun's ideal point at −0.039 with a 95% credible interval of (−0.122, 0.044). We dissect Blackmun's trajectory further in Section 5.2.6.

5.2. **Dynamic analysis of U.S.S.C., 1946–2012.** Table 1 mostly captures the Court's ideological alignment, but Blackmun's erroneous classification cautions against assuming a fixed ideological clustering over time. On the other hand, while Blackmun's and Souter's ideologies shifted, they did so gradually. The hidden Markov binary cluster model from Section 4 accommodates gradual shifts in judicial ideology by incorporating Markovian dependence among the ideological partitioning of justices for each term. Similarly, Martin & Quinn [24] adopt a random walk model for term-to-term changes to ideal points.

We apply the hidden Markov binary cluster model with $\beta = 1/2$ and $k^* = 2$ to all 8,561 cases in the database. Table 5.2 shows the sequence of partitions estimated by the posterior mode. We choose $k^* = 2$ as it affords the most straightforward comparisons to other analyses of the Court based on ideal points; see Section 5.3. We also fit the model to $k^* = 3$ and found only minor deviations from Table 5.2.

5.2.1. *Comparison to fixed clustering model.* While there are several drawbacks to the fixed clustering model of Section 5.1, many of its conclusions agree with those from the hidden Markov binary cluster model. Of course, there are important differences that could not possibly be detected based on the results of Table 1 alone. Notably, the dynamic cluster model detects a striking difference between the Court's behavior under Chief Justices Earl Warren (1953–1969) and William H. Rehnquist (1986–2005). The dynamic model also picks out the well-documented ideological shifts of Harry Blackmun and David Souter.

Justice	1946	47	48	49	50	51	52	53	54	55	56	57	58	59	60
Black	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Reed	●	●	●	●	●	●	●	●	○	●	●				
Frankfurter	●	●	●	●	●	●	●	●	○	●	●	●	●	●	●
Douglas	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Murphy	○	○	○												
Jackson	●	●	●	●	●	●	●	●							
Rutledge	○	○	○												
Burton	●	●	●	●	●	●	●	●	○	●	●	●	●		
Vinson	●	●	●	●	●	●	●								
Clark				●	●	●	●	●	○	●	●	●	●	●	●
Minton				●	●	●	●	●	○	●					
Warren								●	○	○	○	○	○	○	○
Harlan										●	●	●	●	●	●
Brennan											○	○	○	○	○
Whittaker												●	●	●	●
Stewart													●	●	●

Justice	1961	62	63	64	65	66	67	68	69	70	71	72	73	74	75
Black	○	○	○	○	○	○	○	○	○	○					
Frankfurter	●														
Douglas	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○
Clark	●	●	○	○	○	●									
Warren	○	○	○	○	○	○	●	○							
Harlan	●	●	●	○	○	●	○	○	○	●					
Brennan	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○
Whittaker	●														
Stewart	●	●	○	○	○	●	●	○	○	●	○	●	●	●	●
White		○	○	○	○	●	○	○	○	●	●	●	●	●	●
Goldberg		○	○	○											
Fortas					○	○	●	○							
Marshall							●	○	○	○	○	○	○	○	○
Burger									○	●	●	●	●	●	●
Blackmun										●	●	●	●	●	●
Powell												●	●	●	●
Rehnquist												●	●	●	●
Stevens															●

Justice	1976	77	78	79	80	81	82	83	84	85	86	87	88	89	90
Brennan	○	○	○	○	○	○	○	○	○	○	○	○	○	○	
Stewart	●	●	●	●	●										
White	●	●	●	●	●	○	●	●	●	●	●	●	●	●	○
Marshall	○	○	○	○	○	○	○	○	○	○	○	○	○	○	●
Burger	●	●	●	●	●	●	●	●	●	●					
Blackmun	●	●	●	●	●	○	○	●	●	○	○	○	○	○	●
Powell	●	●	●	●	●	●	●	●	●	●					
Rehnquist	●	●	●	●	●	●	●	●	●	●	●	●	●	●	○
Stevens	●	●	○	●	●	○	○	○	○	○	○	○	○	○	●
O'Connor						●	●	●	●	●	●	●	●	●	○
Scalia											●	●	●	●	○
Kennedy													●	●	○
Souter															○

Justice	1991	92	93	94	95	96	97	98	99	00	01	02	03	04	05
White	○	○													
Blackmun	●	●	○												
Rehnquist	○	○	●	○	○	○	○	○	○	○	○	○	○	○	
Stevens	●	●	○	●	●	●	●	●	●	●	●	●	●	●	○
O'Connor	○	○	●	○	○	○	○	○	○	○	○	○	○	○	●
Scalia	○	○	●	○	○	○	○	○	○	○	○	○	○	○	●
Kennedy	○	○	●	○	○	○	○	○	○	○	○	○	○	○	●
Souter	○	○	○	●	●	●	●	●	●	●	●	●	●	●	○
Thomas	○	○	●	○	○	○	○	○	○	○	○	○	○	○	●
Ginsburg			○	●	●	●	●	●	●	●	●	●	●	●	○
Breyer				●	●	●	●	●	●	●	●	●	●	●	○
Roberts															●

Justice	2006	07	08	09	10	11	12
Stevens	○	○	○	○			
Scalia	●	●	●	●	○	○	○
Kennedy	●	●	●	●	○	○	○
Souter	○	○	○				
Thomas	●	●	●	●	○	○	○
Ginsburg	○	○	○	○	●	●	●
Breyer	○	○	○	○	●	●	●
Roberts	●	●	●	●	○	○	○
Alito	●	●	●	●	○	○	○
Sotomayor				○	●	●	●
Kagan					●	●	●

TABLE 2. Estimated ideological cluster sequence $(B_t)_{t=1946,\ldots,2012}$ based on posterior mode in the hidden Markov binary cluster model with $\beta = 1/2$ and $k^* = 2$. Within each term, active justices partition into one or two blocks. Any justice not listed or without a symbol for a given term was inactive for that term. We indicate block membership by open (○) and closed (●) circles. There is no meaning, e.g., liberal/conservative, assigned to these symbols.

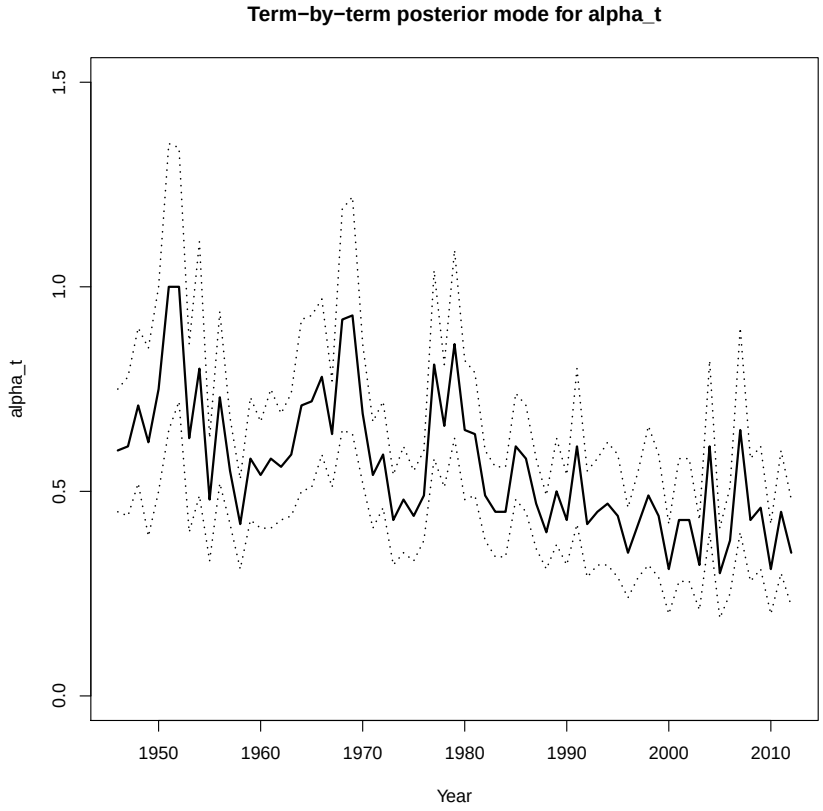

FIGURE 3. Posterior mode for $(\alpha_t)_{t=1946,\ldots,2012}$ based on Exponential(1) prior distribution. Dotted lines show 95% credible intervals.

5.2.2. *Comparison to i.i.d. clustering model.* In the context of the temporal clustering, it is natural to compare the performance of the hidden Markov model to the model for which the latent partition sequence $(\tilde{B}_t)_{t=1946,\ldots,2012}$ are chosen as independently and identically distributed. For this comparison, we assumed $(\tilde{B}_t)_{t=1946,\ldots,2012}$ as i.i.d. from (5) with parameter $\beta = 1$ and $k^* = 2$. (Note our choice $\beta = 1$ here reflects that this choice corresponds to the stationary distribution for the Markov model with $\beta = 1/2$.) For many years, including the stretch since the late 1980s, the i.i.d. model agrees with Table 5.2 in each year. This is not surprising since the Court's behavior has been

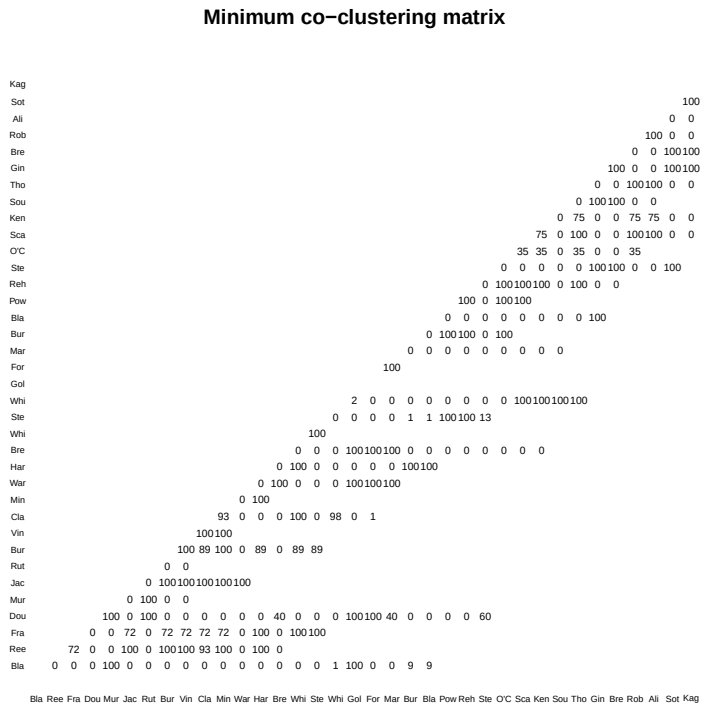

FIGURE 4. Matrix $K = (K_{jj'})$ with entry $K_{jj'}$ the minimum of the posterior co-clustering probabilities (multiplied by 100) for justices j and j' who served together for at least one term. Justices are listed in order of seniority by the first three letters of their surname, e.g., Hugo Black (Bla), Stanley Reed (Ree), Felix Frankfurter (Fra), etc.

rather predictable and quite steady during this time. The hidden Markov model does serve its purpose, however, in smoothing out irregularities during the particularly tumultuous years of the Warren Court, as we discuss next.

5.2.3. *The Warren Court, 1953–1969.* Landmark cases were commonplace during Earl Warren's time as Chief Justice. In Warren's second year as Chief, the Court decided perhaps the watershed case of the past century,

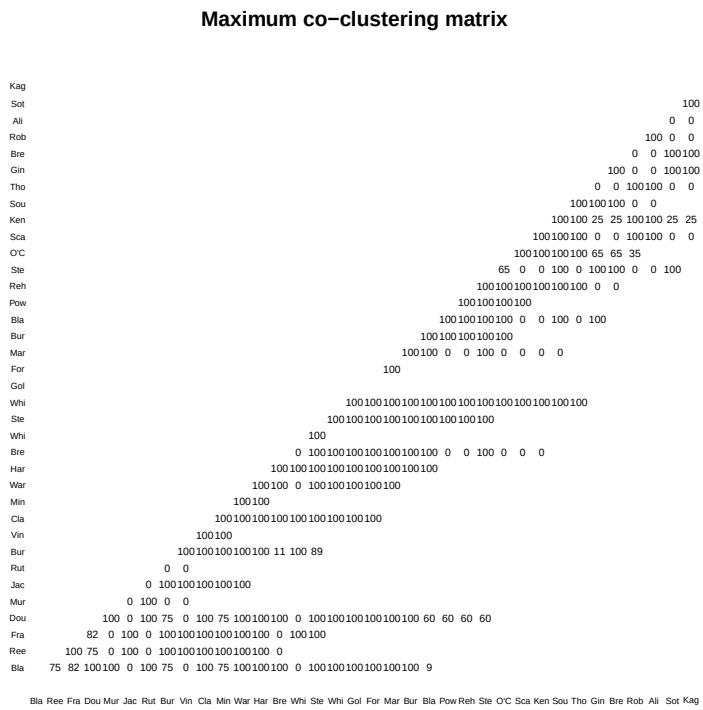

FIGURE 5. Matrix $K = (K_{jj'})$ with entry $K_{jj'}$ the maximum of the posterior co-clustering probabilities (multiplied by 100) for justices j and j' who served together for at least one term. Justices are listed in order of seniority by the first three letters of their surname, e.g., Hugo Black (Bla), Stanley Reed (Ree), Felix Frankfurter (Fra), etc.

Brown v. Board of Education of Topeka, 347 U.S. 483 (1954), which outlawed racial segregation in schools. Before Warren's retirement in 1969, he and his associate justices reached several key decisions on voting rights (e.g., *Baker v. Carr*, 369 U.S. 186 (1962), *Reynolds v. Sims*, 377 U.S. 533 (1964)) and due process (e.g., *Gideon v. Wainwright*, 372 U.S. 335 (1963), *Miranda v. Arizona*, 384 U.S. 436 (1966)). These cases, though controversial, were handled decisively: the justices voted 9–0 in *Brown* and *Gideon*, 8–1 in

Reynolds (Harlan dissenting), and 6–2 in *Baker* (Harlan and Frankfurter dissenting). Many of his contemporaries credit Warren's leadership for the justices' solidarity.

Unlike other quantitative analyses of the Court, Table 5.2 detects this unity in the Warren Court. We estimate only a single ideological cluster in the years 1954, 1964, 1965, and 1968 and a cluster with eight judges, excluding Harlan, in 1963, the year Harlan was the lone dissent in *Reynolds*. We attribute our model's ability to uncover this phenomenon to two factors, our choice to include unanimous decisions in our analysis and the clear meaning of the relationship between ideological clustering and case outcome in the binary cluster model. Other analyses that exclude unanimous decisions, e.g., [24], do not spot this aspect.

The Warren Court's willingness to vote across ideological lines is also borne out by the relatively large estimates of α_t in the mid-1950s and mid-late 1960s; see Figure 3. In the coming section, we contrast these estimates with those for the Rehnquist Court.

5.2.4. *The Rehnquist Court, 1986–2005.* After fourteen years as associate justice, William Rehnquist was promoted to Chief in 1986. Both Rehnquist and Warren were appointed by Republican presidents, Eisenhower and Nixon, respectively, but their similarities end there. Where Warren's liberal pragmatism united his associate justices, Rehnquist's staunch conservatism may have divided his. In Section 5.2.3, we noted several unanimous clusterings as well as the relatively higher estimates of α_t during Warren's term. In contrast, our inferred clusterings for Rehnquist's court are remarkably rigid and the estimated values of α_t in Figure 3 are relatively smaller.

Fourteen justices served between 1986 and 2005 but only once did a justice change his/her ideological association, David Souter between 1992 and 1993.

Our estimates of α_t throughout the 1990s and early 2000s further support the division within the Rehnquist Court. From Section 4.1, the probability that two justices in the same ideological cluster vote together on a case is $(\alpha_t + 2)/(2\alpha_t + 2)$; thus, small values of α_t indicate a stricter obedience to the ideological clustering B_t. Overall, the trajectory of $(\alpha_t)_{t=1946,\ldots,2012}$ in Figure 3 supports the conclusion that the Court has been relatively easier to predict since Rehnquist's ascension to Chief Justice. Since 2005, four new justices have joined the Court, two Republican appointees and two Democratic appointees, and have fallen into the same rigid ideological demarcation of the Rehnquist Court: the Republican appointees, Roberts and Alito, align with the conservatives, while the Democratic appointees, Sotomayor and Kagan, align with the liberals.

5.2.5. *David Souter, 1990–2009.* Perhaps the most enigmatic of the justices we study, David Souter's nomination by Republican George H.W. Bush was met with glee by conservatives and scorn by liberals [32], but this sentiment changed shortly thereafter: Table 5.2 shows that Souter clustered with conservatives only in his first three terms, 1990, 1991, and 1992. Alone, this might signify that Souter was moderate, first leaning slightly right and then leaning slightly left, but the co-clustering matrices in Figures 4 and 5 support the conclusion that Souter began as a full-fledged conservative and finished as a devout liberal.

Figure 4 contains the minimum co-clustering probability for justices who served together for at least one term; Figure 5 contains the maximum. Thus, justices who were strong allies throughout their tenure have a minimum probability near 1, whereas sworn adversaries have a maximum probability near 0. Comparing Souter's co-clustering probabilities for select justices

with whom he only served before 1992 and after 1992 tells a clear story. Byron White, a conservative, and Thurgood Marshall, a liberal, retired before 1992. Souter has a minimum co-clustering probability of 1.00 with White and a maximum of 0.00 with Marshall. Conversely, Justices Ginsburg and Breyer, liberals, and Roberts and Alito, conservatives, were appointed after 1992. Souter has a minimum co-clustering probability 1.00 with Ginsburg and Breyer and maximum 0.00 with Roberts and Alito.

5.2.6. *Harry Blackmun, 1970–1994.* Harry Blackmun is best known for his majority opinion in the 1973 blockbuster *Roe v. Wade* [4], perhaps an early indication of his impending shift to the left. Though Table 1 suggests independence from both liberal and conservative sides of the Court, Table 5.2 tells a different story. Blackmun aligns closely with conservative Chief Justice Warren Burger through the 1980 term. In the early 1980s, Blackmun's loyalty wavers between liberal and conservative, indicating moderation. From 1985 onward, Blackmun clusters in the liberal block.

As for Souter, the co-clustering matrices in Figures 4 and 5 paint a clear picture of Blackmun's evolution. Both John Marshall Harlan II (1955–1971), a conservative, and Hugo Black (1937–1971), a liberal, retired early in Blackmun's service. Blackmun has minimum co-clustering probability 1.00 with Harlan and maximum 0.09 with Black, both evidence of Blackmun's initial conservatism. On the other hand, the slew of conservatives appointed after 1986, e.g., Scalia, Kennedy, and Thomas, all have maximum co-clustering probability 0.00 with Blackmun, while liberal Ruth Bader Ginsburg (1993–) has minimum 1.00.

5.3. **Comparison to latent space analyses.** We compare our results to prior analysis by Martin & Quinn [24][5]. For a detailed description of their ideal

[5]Updated analyses obtained at http://mqscores.berkeley.edu/

point approach, see [24]. In a nutshell, the ideal point model assigns a utility difference $z_{t,i,j}$ for justice j on case (t,i). If $z_{t,i,j} > 0$, justice j votes to reverse the decision of the lower court; otherwise, justice j votes to affirm. The preference for each case is modeled as $z_{t,i,j} = \alpha_{t,i} + \beta_{t,i}\theta_{t,j} + \epsilon_{t,i,j}$, where $\alpha_{t,i}, \beta_{t,i}$ are random effects specific to case (t,i), $\epsilon_{t,i,j}$ is random noise, and $\theta_{t,j}$ is the *ideal point* of justice j for term t, i.e., the *Martin–Quinn score*. Negative values of $\theta_{t,j}$ indicate a left-leaning ideology and positive values a right-leaning ideology. Based on ideal point estimates between 1946–2012, we obtain a sequence of estimated ideological partitions $(B_t^{MQ})_{t=1946,\dots,2012}$ $(B_t^{B})_{t=1946,\dots,2012}$ for Martin–Quinn by putting justices in the same cluster for a term if their ideal points for that term have the same sign.

As noted above, the Martin–Quinn analysis removes unanimous cases and, thus, is unable to detect the high degree of cohesion during the latter part of the Warren Court. Furthermore, the precise relationship of the parameters and the voting alignment is not explicit in the ideal point model. In particular, we see no way to succinctly summarize the level of adherence to the ideological clustering B_t as we can through the term-specific variance parameters in our model, cf. Figure 3.

Overall, our estimates in Table 5.2 and Figures 4 and 5 closely align with the Martin–Quinn analysis. There are some anomalous years in our estimate, e.g., all justices in the same cluster in 1954 and several times in the 1960s, which identify a real characteristic of the Warren Court and are likely caused by our choice not to eliminate unanimous cases from the data set. For fair comparison, we ran our analysis with unanimous cases removed and obtained partitions much closer to Martin & Quinn's for all years under study; however, as we discussed in Section 5.2.3 above, the inferred unanimous partitions for the Warren Court do provide real

insight into the Warren Court's behavior and, thus, highlight a benefit to our approach.

Table 5.2 also agrees with Martin–Quinn on David Souter: in our analysis, Souter clusters with conservatives during his first three terms, 1990–1992, and then switches to the liberal cluster for 1993–2008; similarly, Souter's Martin–Quinn score is positive during 1990–1992 and negative thereafter. On Harry Blackmun, our analysis agrees with Martin–Quinn prior to 1980 and after 1986, with some discrepancy in the intervening years. For the period 1981–1986, the Martin–Quinn scores are slightly negative, but the 95% credible interval contains 0 each time, signifying an inconclusive classification. Similarly, our classification of Blackmun wavers between liberal and conservative for these years. Thus, it is reasonable to conclude that Blackmun behaved as a moderate in those years and did not cluster definitely with liberals or conservatives.

6. Summary

We have presented a new hidden Markov chain model for latent clusters. The Markov chain is a new development in the applied probability literature for modeling DNA sequences, and here we present its first known use in a fully Bayesian context of hidden Markov modeling. The model possesses many nice theoretical properties, all of which have been established rigorously in the appropriate literature [11, 12]; therefore, the model is well suited for much more general applications than the one we have considered here, particularly political science, sociology, and economics, where data sets affected by regime change commonly arise.

Our detailed analysis here for Supreme Court voting provides a thorough empirical demonstration of the model's performance in an extensive dataset for which there is ample understanding of reasonable inferences and prior

quantitative analyses for comparison. Comparisons with common knowledge in the field reveal that our method has performed well in this instance, and we anticipate it will perform similarly in other suitable applications. Aside from being an application of its own interest to many readers, our detailed analysis of the Court augments or supports prior research on the Court. Though not our main goal, our analysis provides a quantitative justification for certain well known aspects of the Court, e.g., the cohesion of the Warren Court, which prior quantitative analyses are unable to detect.

References

[1] D. J. Aldous. Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, volume 1117 of *Lecture Notes in Math.*, pages 1–198. Springer, Berlin, 1985.

[2] C. Antoniak. Mixtures of dirichlet processes with applications to non-parametric problems. *The Annals of Statistics*, 2:1152–1174, 1974.

[3] J. Bertoin. *Random fragmentation and coagulation processes*, volume 102 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2006.

[4] H. Blackmun. Roe v. Wade. *410 U.S. 113*, 1973.

[5] J. Booth, G. Casella, and J. Hobert. Clustering using objective functions and stochastic search. *JRSS B*, 70:119–139, 2008.

[6] H. Bunke and T. E. Caelli. *Hidden Markov Models: Applications in Computer Vision*.

[7] O. Cappé. Ten years of HMMs (An HMM bibliography), year = March 2001,.

[8] O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer, 2007.

[9] T. S. Clark and B. Lauderdale. Locating supreme court opinions in doctrine space. *American Journal of Political Science*, 54(4):871–890, 2010.

[10] J. Clinton, S. Jackman, and D. Rivers. The Statistical Analysis of Roll Call Data. *American Political Science Review*, 98(2):355–370, 2004.

[11] H. Crane. A consistent Markov partition process generated from the paintbox process. *J. Appl. Probab.*, 43(3):778–791, 2011.

[12] H. Crane. The cut-and-paste process. *Annals of Probability*, 42(5):1952–1979, 2014.

[13] H. Crane. The ubiquitous Ewens sampling formula (with discussion and rejoinder by the author). *Statistical Science*, to appear, 2016.

[14] W. J. Ewens. The sampling theory of selectively neutral alleles. *Theoret. Population Biology*, 3:87–112, 1972.

[15] T. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[16] J. A. Hartigan. Partition models. *Comm. Statist. Theory Methods*, 19(8):2745–2756, 1990.

[17] P. Irons. *A People's History of the Supreme Court: The Men and Women Whose Cases and Decisions Have Shaped Our Constitution*. Penguin Books, 2006.

[18] H. Ishwaran and L. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–173, 2001.

[19] H. Ishwaran and L. James. Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, 13:1211–1235, 2003.

[20] B. Juang and L. Rabiner. Hidden Markov Models for Speech Recognition. *Technometrics*, 33(3):251–272, 1991.

[21] J. F. C. Kingman. Random partitions in population genetics. *Proc. Roy. Soc. London Ser. A*, 361(1704):1–20, 1978.

[22] J. F. C. Kingman. The coalescent. *Stochastic Process. Appl.*, 13(3):235–248, 1982.

[23] T. Koski. *Hidden Markov Models for Bioinformatics*. Computational Biology. Springer, 2001.

[24] A. D. Martin and K. M. Quinn. Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999. *Political Analysis*, 10(2):134–153, 2002.

[25] P. McCullagh and J. Yang. How many clusters? *Bayesian Anal.*, 3(1):101–120, 2008.

[26] J. Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006.

[27] J. S. Rosenthal and A. H. Yoon. Detecting Multiple Authorship of United States Supreme Court Legal Decisions Using Function Words. *Annals of Applied Statistics*, 5.

[28] B. Schwartz. *A History of the Supreme Court*. Oxford University Press, USA, 1995.

[29] S. Scott. Bayesian Methods for Hidden Markov Models. *Journal of the American Statistical Association*, 97(457):337–351, 2002.

[30] L. Sirovich. A pattern analysis of the second Rehnquist U.S. Supreme Court. *PNAS*, 100(13):7432–7473, 2003.

[31] L. Thurstone and J. Degan. Factorial study of the Supreme Court. *PNAS*, 37:628–635, 1951.

[32] J. Toobin. *The Nine: Inside the Secret World of the Supreme Court*. Anchor, 2008.

Department of Statistics & Biostatistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854, hcrane@stat.rutgers.edu