FINDING CHANGE POINTS IN THE POLLS

Albert D. Shieh
Harvard University
Department of Statistics
Cambridge, MA 02138, USA

Lynette C. Lee
Harvard University
Department of Government
Cambridge, MA 02138, USA

ABSTRACT
After an election campaign, it is important to identify events that marked change points in voter support. Pre-election polls provide a measure of the state of voter support at points in time during the election campaign. However, polling data is difficult to analyze because it is sparse and comes from multiple sources, which can be individually biased. We propose a change point model for polling data that increases confidence by combining polls and identifying change points simultaneously. We demonstrate the utility of our model on polling data from the 2008 U.S. presidential election.

Index Terms— Change points, pooling, Gibbs sampler

1. INTRODUCTION
After an election campaign, it is important to retrospectively identify events that influenced the outcome of the election in order to better understand the concerns of voters. For example, the revelation of damaging news about a candidate would be a potential change point in the election. Election campaigns have been studied extensively by political scientists for change points [1]. However, approaches to identifying change points have been primarily qualitative so far, resulting in much debate.

Pre-election polls are the most common quantitative measure of the state of voter support at points in time during the election campaign [2] and offer a more objective alternative to identifying change points. Naively, polling data can be treated as a time series, for which change point models are well studied [3]. However, polling data presents problems that make the direct application of standard change point models difficult [4]. First, polling data is not available on many days of the election campaign. Second, individual polls are imprecise due to sampling error and cannot detect small changes in voter support with statistical significance. Finally, different polling houses may have systematic biases due to their survey methodology and candidate preferences.

Pooling models [5, 6, 7] have been proposed to combine polls from different polling houses in order to estimate a latent state of voter support for every day of the election campaign. Pooling decreases sampling error by aggregating data and can incorporate biases for polling houses. A simple strategy for change point analysis is to apply a pooling model and change point model separately. However, this approach ignores uncertainty in the latent states of voter support, which is usually large, especially for periods of time in which few polls were available.

We propose a unified model for combining polls and identifying change points simultaneously. We extend pooling models by introducing partitions over the latent states of voter support and present a Gibbs sampler for efficient inference in our change point model.

We apply our change point model to polling data from the 2008 U.S. presidential election and show that it produces results consistent with political knowledge.

2. CHANGE POINT MODEL
2.1. Data description
Consider an election campaign lasting over \( t = 1, \ldots, T \) days where \( t = 1 \) is the first day of the election campaign and \( t = T \) is the day of the election. We observe polling data \( \{y_i, n_i, t_i, j_i\}_{i=1}^{N} \) consisting of \( N \) polls from \( J \) polling houses. Each poll is a measurement \( y_i \in [0, 1] \) of the latent vote share for the candidate being tracked with an associated sample size \( n_i \), day \( t_i \), and polling house \( j_i \). On some days, we may not observe any polls, while on other days, we may observe multiple polls.

2.2. Combining polls
Our first goal is to estimate the latent vote share \( x_t \in [0, 1] \) for each day \( t \) of the election campaign from the polls. We use a dynamic linear model [6, 7], treating the latent vote shares as states and the polls as biased measurements of the states with normally distributed errors. The model is similar to a Kalman filter [8], but we allow for systematic bias \( \delta_j \) by each polling house \( j \) in their polls.

We assume that each poll is generated from a normal distribution

\[
  y_i \sim N\left(x_{t_i} + \delta_{j_i}, \frac{1 - y_i}{n_i}\right)
\]

with mean determined by the latent vote share \( x_{t_i} \) on the day of the poll plus the bias \( \delta_{j_i} \) of the polling house and variance determined by the standard error of the poll on a simple random sample of voters.

In order to model changes in the latent vote shares, we assume that the latent vote shares follow a random walk

\[
x_t \sim N(x_{t-1}, \omega^2)
\]

initialized by

\[
x_1 \sim \text{Uniform}(0.25, 0.75).
\]

We assume that latent vote shares differ from day to day by random noise from a normal distribution with mean zero and variance \( \omega^2 \). We initialize the random walk by bracketing latent vote shares within a feasible range of \([0.25, 0.75]\) before we observe any polls.

We choose a normal prior distribution for the bias of each polling house to be

\[
  \delta_j \sim N(0, (0.25/2)^2).
\]

We assume that the 95% confidence interval for the biases of polling houses is \([-0.25, 0.25]\) such that no polling house is likely to be
biased by more than 25%. Reputable polling houses should easily satisfy this requirement.

We choose a uniform prior distribution for the standard deviation of the random walk $\omega$ to be

$$f(\omega) = 100, \ \omega \in (0, 0.01). \quad (5)$$

We assume that 95% of daily changes in the latent vote shares are less than 2%. We allow for larger daily changes in the latent vote shares in response to events, but we assume that they are rare.

Since the biases of the polling houses and the latent vote shares can offset each other arbitrarily, they must be constrained in order to identify the model. We do not want to constrain the biases of the polling houses because they may be large when there are partisan polling houses. Since we assume that the election is over, we can instead fix the final latent vote share $x_T$ to the actual outcome of the election.

2.3. Identifying change points

Our second goal is to estimate the change points in the latent vote shares. Each latent vote share is approximately generated from a normal distribution

$$x_t \sim N(\mu_t, \sigma_t^2) \quad (6)$$

with an independent mean $\mu_t$ but pooled variance $\sigma_t^2$. We want to find a partition $\rho = (t_0, t_1, \ldots, t_B-1, t_B = T)$ of the latent vote shares into a random number of $B \in \{1, \ldots, T \}$ blocks such that the means are constant $\mu_i = \mu_t$, within each block $t_{r-1} < t \leq t_r$ for $r = 1, \ldots, B$. The transitions between the blocks $t_1 + 1, \ldots, t_{B-1} + 1$ are the change points.

We use a product partition model [9] to form a distribution over partitions. We assume that the probability of a change point at day $t$ is $p$, independently for each day $t = 1, \ldots, T$ in the election campaign. We choose a normal prior distribution for the mean $\mu_{ij}$ of the block from day $i + 1$ to day $j$ to be

$$\mu_{ij} \sim N(\mu_0, \frac{\sigma_0^2}{j-i}) \quad (7)$$

with variance scaled by length of the block. We assume that we can detect small change points as long as there are enough days to estimate them.

We choose prior distributions for the parameters $\mu_0$ and $\sigma^2$ to be

$$f(\mu_0) = 1, \ \mu_0 \in (-\infty, \infty) \quad (8)$$

$$f(\sigma^2) = \sigma^{-2}, \ \sigma^2 \in [0, \infty) \quad (9)$$

in order to make the blocks invariant under location and scale changes and the prior distributions for the parameters $p$ and $w = \sigma^2/\sigma_0^2$, the signal to error variance ratio, to be

$$f(p) = p_0^{-1}, p \in [0, p_0] \quad (10)$$

$$f(w) = w_0^{-1}, w \in [0, w_0] \quad (11)$$

such that

$$P(p) = p_0^{-1} \int_0^{p_0} p^{B-1}(1-p)^{T-B} \, dp \quad (12)$$

where $p_0 \in [0, 1]$ and $w_0 \in [0, 1]$ are hyperparameters to be set. The hyperparameter $p_0$ controls the number of change points, while the hyperparameter $w_0$ controls the lengths of the blocks. Since the latent vote shares are likely to be noisy, we set $p_0 = 0.001$ and $w_0 = 0.1$ to be small so that change points are only identified when the data is overwhelming.

3. GIBBS SAMPLER

There are a total of $2T + J - 1$ parameters in our change point model. We use Markov Chain Monte Carlo (MCMC) methods to simulate the posterior distribution of the parameters. Since all of the conditional posterior distributions have parametric forms [10, 11], we use a Gibbs sampler as follows.

3.1. Latent vote shares

For each day $t = 1, \ldots, T - 1$, we draw the latent vote share $x_t$ as follows.

- For one poll on day $t$, we draw the latent vote share $x_t$ from the normal distribution

$$N\left(\frac{y_t - \delta_{j_t}}{s_{ij_t}^2} + \frac{x_{t+1} + x_{t-1}}{\omega^2}, \frac{1}{s_{ij_t}^2 + \frac{\omega^2}{2}}\right)$$

where $i$ is the index of the poll on day $t$.

- For no polls on day $t$, we draw the latent vote share $x_t$ from the normal distribution

$$N\left(\frac{x_{t+1} + x_{t-1}}{2\omega^2}, \frac{1}{2\omega^2}\right).$$

- For multiple polls on day $t$, we draw the latent vote share $x_t$ from the normal distribution

$$N\left(\frac{\sum_{i \in Y_t} y_i - \delta_{j_t}}{s_{ij_t}^2} + \frac{x_{t+1} + x_{t-1}}{\omega^2}, \frac{1}{\sum_{i \in Y_t} s_{ij_t}^2 + \frac{\omega^2}{2}}\right)$$

where $Y_t = \{i : t_i = t\}$ indexes the polls on day $t$.

- For the first day $t = 1$, we draw the latent vote share $x_1$ from the normal distribution

$$N\left(\frac{y_1 - \delta_{j_1}}{s_{i1}^2} + \frac{x_2}{\omega^2}, \frac{1}{s_{i1}^2 + \frac{\omega^2}{2}}\right)$$

truncated such that $x_1 \in [0.25, 0.75]$.

3.2. Biases of polling houses

For each polling house $j = 1, \ldots, J$, we draw the bias $\delta_j$ from the normal distribution

$$N\left(\frac{\sum_{i \in Y_j} y_i - x_i}{s_{ij}^2}, \frac{1}{\sum_{i \in Y_j} s_{ij}^2 + \frac{1}{\omega^2}}\right)$$

where $Y_j = \{i : j_i = j\}$ indexes the polls from polling house $j$.  

1907
Polling House | Polls | Size | Bias
--- | --- | --- | ---
ABC News/Wash Post | 15 | 1136 | +2.30*
CBS News/NY Times | 20 | 877 | +0.18
CNN | 14 | 861 | +2.42*
Cook/RT Strategies | 6 | 824 | −2.68*
Democracy Corps | 6 | 1170 | +1.40
Diageo/Hotline | 6 | 830 | −2.08
FOX News | 13 | 910 | −3.27*
Gallup/USA Today | 38 | 2079 | −0.20
GWU/Battleground | 11 | 891 | −1.08
IBD/TIPP | 11 | 929 | −3.45*
Ipsos/AP/McClatchy | 14 | 862 | −0.40
LA Times/Bloomberg | 7 | 1254 | −1.38
Marist College | 6 | 797 | −2.93*
NBC News/Wall St. Jnrl | 12 | 967 | −0.35
Newsweek | 9 | 1065 | −0.25
NPR | 1 | 1000 | +1.90
Pew Research | 13 | 1625 | +1.12
Quinnipiac | 4 | 1434 | +2.18
Rasmussen Reports | 23 | 2548 | +0.32
Reuters/C-SPAN/Zogby | 6 | 1205 | −0.57
Reuter/Zogby | 8 | 1023 | −1.15
Time | 5 | 951 | +0.45

Table 1. Polling houses and their biases where * denotes statistical significance at the 5% level.

3.3. Variance of random walk

We draw the precision of the random walk \( \tau = \omega^{-2} \) from the gamma distribution

\[
\text{Gamma}\left( \frac{T - 4}{2}, \frac{1}{2} \sum_{t=2}^{T} (x_t - x_{t-1})^2 \right)
\]

truncated such that \( \tau \in (10000, \infty) \).

3.4. Change points

Let \( u_t = 1 \) if there is a change point at day \( t + 1 \) and let \( u_t = 0 \) otherwise such that a partition can be parameterized as \( \rho = (u_1, \ldots, u_{T-1}) \). For each day \( t = 1, \ldots, T - 1 \), we draw a value of \( u_t \) according to the odds ratio

\[
P(u_t = 1) = \frac{\int_{0}^{p_0} p^B (1 - p)^{T - B - 1} dp}{\int_{0}^{p_0} p^{B-1} (1 - p)^{T - B} dp} \int_{0}^{w_0} w^{B/2} \left( W_1 + B_1 w \right)^{(T-1)/2} dw \int_{0}^{w_0} \left( W_0 + B_0 w \right)^{(T-1)/2} dw
\]

where \( W_0, B_0, W_1, \) and \( B_1 \) are the within and between block sums of squares obtained when \( u_t = 0 \) and \( u_t = 1 \) respectively.

4. EXPERIMENTS

We evaluated our change point model on polling data from the recent 2008 U.S. presidential election between Barack Obama, the Democratic Party (liberal) candidate and winner, and John McCain, the Republican Party (conservative) candidate. We collected 247 polls from 22 major polling houses spanning a 308 day period from January 1, 2008, before either of the candidates had been nominated, to November 4, 2008, the day of the election. Details of the polling houses, their numbers of polls, and the mean sample sizes of their polls can be found in Table 1.

We drew two million samples, throwing away the first half of the samples as burn-in. Since some of the parameters were highly correlated, we thinned the chain by saving every 1000th sample, resulting in 1000 samples. We used the sample means as point estimates and the 2.5% and 97.5% sample quantiles as 95% confidence intervals for the parameters. Estimated daily Obama latent vote shares and probabilities of a change point are shown in Figure 1. Despite having only 247 polls to cover 308 days, we recovered a smooth trajectory of latent vote shares.

We used a threshold of 0.5 for the daily probabilities of a change point in order to identify 16 major change points. We then searched newspaper articles for events within a one week window of the change points that would explain the shift in voter support. For each change point, we were able to identify a corresponding event that is believed by political pundits to have influenced the outcome of the election. The change points and events are shown in Table 2.

<table>
<thead>
<tr>
<th>Date</th>
<th>Change</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Jan 4</td>
<td>+1.86</td>
<td>Obama wins Iowa</td>
</tr>
<tr>
<td>2 Jan 16</td>
<td>−0.94</td>
<td>Clinton wins New Hampshire</td>
</tr>
<tr>
<td>3 Jan 26</td>
<td>+1.38</td>
<td>Obama wins South Carolina</td>
</tr>
<tr>
<td>4 Jan 30</td>
<td>+3.15</td>
<td>John Edwards drops out</td>
</tr>
<tr>
<td>5 Mar 3</td>
<td>−1.43</td>
<td>McCain wins nomination</td>
</tr>
<tr>
<td>6 Mar 22</td>
<td>−1.20</td>
<td>Jeremiah Wright controversy</td>
</tr>
<tr>
<td>7 Apr 12</td>
<td>+1.51</td>
<td>Obama beats Clinton in debates</td>
</tr>
<tr>
<td>8 Jun 4</td>
<td>+1.52</td>
<td>Obama wins nomination</td>
</tr>
<tr>
<td>9 Jun 30</td>
<td>−1.53</td>
<td>Obama runs towards center</td>
</tr>
<tr>
<td>10 Aug 1</td>
<td>−1.02</td>
<td>Obama is a celebrity ad</td>
</tr>
<tr>
<td>11 Aug 25</td>
<td>+2.26</td>
<td>Democratic National Convention</td>
</tr>
<tr>
<td>12 Aug 31</td>
<td>−1.58</td>
<td>Sarah Palin nomination</td>
</tr>
<tr>
<td>13 Sep 11</td>
<td>+2.12</td>
<td>Financial crisis begins</td>
</tr>
<tr>
<td>14 Oct 5</td>
<td>+1.55</td>
<td>Obama beats McCain in debates</td>
</tr>
<tr>
<td>15 Oct 18</td>
<td>+1.23</td>
<td>Obama garners endorsements</td>
</tr>
<tr>
<td>16 Oct 30</td>
<td>+1.42</td>
<td>Obama prime time ad</td>
</tr>
</tbody>
</table>

Table 2. Change points, their effects, and the corresponding events.
5. CONCLUSION

We proposed a change point model for polling data that identifies change points in latent states of voter support. We presented a Gibbs sampler for efficient inference in our change point model and demonstrated the utility of our change point model on polling data from the 2008 U.S. presidential election. Political scientists have long considered polls to be too noisy to learn complicated structure from [4]. To our knowledge, our study is the first to demonstrate the feasibility of using polls to find change points in election campaigns.

One major limitation of our change point model is that we need the actual outcome of the election in order to constrain the latent vote shares, restricting the use of our change point model to retrospective analysis. A useful extension of our change point model would be an adaptation from batch to online learning, which would allow it to be used during election campaigns.

6. REFERENCES


Fig. 1. (Top) Estimated daily Obama latent vote shares with a 95% confidence interval and individual polls. (Bottom) Estimated daily probabilities of a change point with detected change points marked.