Election Forecasting Model

Henry Linder
mhlinder@gmail.com
November 5, 2018

Notation derives from (Cargnoni, Müller, & West, 1997), itself citing (West & Harrison, 1997).

1 Setup

Denote by $\mathbf{p}_i = (p_{i1}, \ldots, p_{i(r+1)})'$ a vector of r observed proportions on day t_i, $i = 1, \ldots, n$, $t_i \in \{0, \ldots, T\}$. $\mathbf{p}_i$ gives voter responses to an opinion poll about their preference among r candidates, as well as an additional category for "Other". Accompanying $\mathbf{p}_t$ is a sample size, N_i, so that an estimate of the number of respondents who prefer the j^{th} candidate may be calculated as $y_{ij} = p_{ij} N_i$, $j = 1, \ldots, r$. The data is reported in proportions, so to ensure the proper sample size after calculating y_{ij}, we define $y_{i(r+1)} = N_i - \sum_{j=1}^r y_{ij}$.

It is possible that multiple polls concluded on the same day, in which case we have replicates within the day. A basic way to incorporate this data into the model without modification is to calculate the counts y_{ij} within each poll, then sum within each of the categories j. WLOG, we assume that when there is more than one observation on a single day, we reduce the dimension in this fashion.

We consider the vector of counts $\mathbf{y}_i = (y_{i1}, \ldots, y_{i(r+1)}$, and we consider a multinomial likelihood:

$$[\mathbf{y}_i | \boldsymbol{\pi}_{t_i}, N_i] \sim \text{Mult.}(\boldsymbol{\pi}_{t_i}, N_i) \tag{1}$$

$$\boldsymbol{\eta}_i = h(\boldsymbol{\pi}_{t_i}) = (h_1(\pi_{t_i 1}), \ldots, h_r(\pi_{t_i r}))' \in \mathbb{R}^r, \quad i = 1, \ldots, N. \tag{2}$$

$$h_j(\pi_{tj}) = \log \frac{\pi_{tj}}{\pi_{t(r+1)}}, \quad j = 1, \ldots, r \tag{3}$$

$$\boldsymbol{\pi}_{t_i} = h^{-1}(\boldsymbol{\eta}_i) = (h_1^{-1}(\boldsymbol{\eta}_i), \ldots, h_{r+1}^{-1}(\boldsymbol{\eta}_i))' \in [0,1]^r \tag{4}$$

$$h_j^{-1}(\boldsymbol{\eta}_i) = \frac{e^{\eta_{ij}}}{1 + \sum_{j'=1}^r e^{\eta_{ij'}}}, \quad j = 1, \ldots, r \tag{5}$$

$$h_{r+1}^{-1}(\boldsymbol{\eta}_i) = \frac{1}{1 + \sum_{j'=1}^r e^{\eta_{ij'}}} \tag{6}$$

where $\sum_{j=1}^{r+1} \pi_{tj} = 1$.

We model $\boldsymbol{\eta}_i$ using a dynamic linear model, which provides a structure to relate all polls that occur on the same day, i.e., the sets $\{\mathbf{y}_i | t_i = t\}$, $t = 1, \ldots, T$.

$$\boldsymbol{\eta}_t = \boldsymbol{\beta}_t + \boldsymbol{\nu}_t, \quad \boldsymbol{\nu}_t \sim N_r(\mathbf{0}, v\mathbf{I}_r) \tag{7}$$

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N_r(\mathbf{0}, w_t\mathbf{I}_r) \tag{8}$$

and we assume $\boldsymbol{\beta}_0$ known. This model is a random walk for the daily multinomial proportion vectors.

In the notation and terminology of (West & Harrison, 1997), this is a constant dynamic linear model where

$$\mathbf{F} = \mathbf{I}_r \tag{9}$$

$$\mathbf{G} = \mathbf{I}_r \tag{10}$$

To construct a Bayesian model, we assign constant variance in the observation equation (v), and iid time-varying variances in the system equation (w_t).

By the forward-filtering, backwards-sampling (FFBS) algorithm, the prior distributions for all $\boldsymbol{\beta}_t$ depend recursively only on $\boldsymbol{\beta}_0$, so we complete the model by assigning priors to the parameters $\{\boldsymbol{\beta}_0, v, w_1, \ldots, w_T\}$.

$$\boldsymbol{\beta}_0 \sim N_r(\mathbf{m}_0, C_0\mathbf{I}_r) \tag{11}$$

$$v \sim \text{Inv.-Gamma}(\alpha_0, \beta_0), \quad t = 1, \ldots, T \tag{12}$$

$$w_t \sim \text{Inv.-Gamma}(\alpha_0, \beta_0) \tag{13}$$

We choose diffuse priors for v, w_t, with $\alpha_0 = \beta_0 = 0.001$.

We specify the mean structure of the model with the forward-filtering, backward-sampling procedure.

1.1 Forward-filtering

(West & Harrison, 1997) give the updates for the multivariate dynamic linear model on page 582.

We suppose $\mathbf{m}_0, C_0$ known, so $\boldsymbol{\beta}_0 \sim N_r(\mathbf{m}_0, C_0\mathbf{I}_r)$.

For $t = 1, \ldots, T$,

1. **Posterior at** $t-1$

$$\boldsymbol{\beta}_{t-1}|\mathcal{D}_{t-1} \sim N_r(\mathbf{m}_{t-1}, C_{t-1}\mathbf{I}_r) \tag{14}$$

2. **Prior at** t

$$\boldsymbol{\beta}_t|\mathcal{D}_{t-1} \sim N_r(\mathbf{m}_{t-1}, R_t\mathbf{I}_r) \tag{15}$$

3. **One-step forecast**

$$\boldsymbol{\eta}_t|\mathcal{D}_{t-1} \sim N_r(\mathbf{m}_{t-1}, Q_t\mathbf{I}_r) \tag{16}$$

4. **Posterior at t**

$$\boldsymbol{\beta}_t|\mathcal{D}_t \sim N_r(\mathbf{m}_t, C_t) \tag{17}$$

$$\begin{align}
\mathbf{m}_t &= \mathbf{m}_{t-1} + A_t(\boldsymbol{\eta}_t - \mathbf{m}_{t-1}) \tag{18}\\
R_t &= C_{t-1} + w_t \tag{19}\\
Q_t &= (R_t + v) \tag{20}\\
A_t &= R_t/(R_t + v) \tag{21}\\
C_t &= R_t - A_t^2 Q_t \tag{22}
\end{align}$$

When filtering, also calculate B_t given below.

1.1.1 Missing observations

Most series are comprised of unequally spaced time points, that is, there may be "missing observations". In this case, we have no data, but the state vector is not of interest: by properties of a normal distribution we can "solve out" the intervening time periods. Practically speaking, this means the variance increases, but the mean remains the same:

$$\begin{align}
\mathbf{m}_t &= m_{t-1} \tag{23}\\
R_t &= C_{t-1} + w_t \tag{24}
\end{align}$$

1.2 Backward-sampling

As written on page 570 of (West & Harrison, 1997), we can then sample backwards by the procedure procedure is then:

- Sample $\boldsymbol{\beta}_T|\mathcal{D}_T \sim N_r(\mathbf{m}_T, C_T\mathbf{I}_r)$
- For $t = n-1, n-2, \ldots, 1, 0$, sample

$$\begin{align}
\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t+1}, \mathcal{D}_t &\sim N_r(\mathbf{h}_t, H_t\mathbf{I}_r) \tag{25}\\
\mathbf{h}_t &= \mathbf{m}_t + B_t(\boldsymbol{\beta}_{t+1} - \mathbf{m}_{t+1}) \tag{26}\\
H_t &= C_t - B_t^2 R_{t+1} \tag{27}\\
B_t &= \frac{C_t}{R_{t+1}} \tag{28}
\end{align}$$

References

Cargnoni, C., Müller, P., & West, M. (1997). Bayesian forecasting of multinomial time series through conditionally Gaussian dynamic models. *Journal of the American Statistical Association*, *92*(438), 640–647.

West, M., & Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models.* Springer-Verlag New York.

2 Example dataset

This dataset was scraped for informational and research purposes from RealClearPolitics[1].

```
   Date            MoE Poll                          Year     N VoterType Start StartDate  End   EndDate    Abrams.D Kemp.R Other
   <chr>         <dbl> <chr>                         <int> <dbl> <chr>     <chr> <date>     <chr> <date>        <dbl>  <dbl> <dbl>
 1 10/29 - 10/31   3.7 Emerson                        2018   724 LV        10/29 2018-10-29 10/31 2018-10-31     0.47   0.49  0.04
 2 10/21 - 10/30   3   Atlanta Journal-Constitution   2018  1091 LV        10/21 2018-10-21 10/30 2018-10-30     0.47   0.47  0.06
 3 10/28 - 10/29   3.9 FOX 5 Atlanta/Opinion Savvy*   2018   623 LV        10/28 2018-10-28 10/29 2018-10-29     0.48   0.47  0.05
 4 10/21 - 10/22   3.4 FOX 5 Atlanta/Opinion Savvy*   2018   824 LV        10/21 2018-10-21 10/22 2018-10-22     0.48   0.48  0.04
 5 10/14 - 10/18   4.8 NBC News/Marist                2018   554 LV        10/14 2018-10-14 10/18 2018-10-18     0.47   0.49  0.04
 6 9/30 - 10/9     2.8 Atlanta Journal-Constitution*  2018  1232 LV        9/30  2018-09-30 10/9  2018-10-09     0.46   0.48  0.06
 7 10/3 - 10/8     4.9 WXIA-TV/SurveyUSA              2018   655 LV        10/3  2018-10-03 10/8  2018-10-08     0.45   0.47  0.08
 8 10/1 - 10/1     3.2 Landmark Communications*       2018   964 LV        10/1  2018-10-01 10/1  2018-10-01     0.46   0.48  0.06
 9 8/26 - 9/4      3.1 Atlanta Journal-Constitution*  2018  1020 LV        8/26  2018-08-26 9/4   2018-09-04     0.45   0.45  0.1
10 7/27 - 7/29     3.8 Gravis                         2018   650 LV        7/27  2018-07-27 7/29  2018-07-29     0.46   0.44  0.1
11 7/15 - 7/19     4.3 WXIA-TV/SurveyUSA              2018  1199 LV        7/15  2018-07-15 7/19  2018-07-19     0.44   0.46  0.1
```

[1] https://www.realclearpolitics.com/epolls/latest_polls/