Election Forecasting Models: Mathematical Foundations

State-Level Presidential Election Forecasting

November 2024

Abstract

We present four probabilistic models for forecasting U.S. presidential elections using state-level polling data: (1) Poll Average, a simple weighted baseline; (2) Kalman Diffusion, treating vote margins as Brownian motion with drift; (3) Improved Kalman, adding stronger regularization; and (4) Hierarchical Bayes, combining fundamentals priors with systematic bias correction. We derive the mathematical foundations for each approach and evaluate their performance on the 2016 election.

Contents

1 Problem Setup and Notation

1.1 Data Structure

For each state $s \in \{1, \ldots, S\}$ (typically $S = 51$ including DC), we observe:

- N_s polls indexed by $i = 1, \ldots, N_s$
- Poll i has:
  - Midpoint date $t_i \in [t_0, T_{elec}]$
  - Sample size n_i
  - Democratic votes d_i, Republican votes r_i
  - Two-party margin $y_i = \frac{d_i - r_i}{d_i + r_i} \in [-1, 1]$
  - Pollster $p(i) \in \mathcal{P}$

1.2 Target Quantity

On election day T_{elec}, the actual two-party margin is:

$$m_s^{true} = \frac{D_s - R_s}{D_s + R_s} \tag{1}$$

where D_s, R_s are total votes received.

1.3 Forecasting Task

Given polls up to forecast date $t_{forecast} < T_{elec}$, predict:

1. **Win probability**: $P(m_s^{true} > 0 \mid \text{data up to } t_{forecast})$
2. **Expected margin**: $\mathbb{E}[m_s^{true} \mid \text{data}]$
3. **Uncertainty**: Standard deviation σ_s

1.4 Evaluation Metrics

For a set of predictions $\{(p_s, \hat{m}_s)\}_{s=1}^{S}$ and outcomes $\{(o_s, m_s^{true})\}_{s=1}^{S}$ where $o_s = \mathbb{1}\{m_s^{true} > 0\}$:

Brier Score (calibration of probabilities):

$$BS = \frac{1}{S} \sum_{s=1}^{S} (p_s - o_s)^2 \in [0, 1] \tag{2}$$

Log Loss (penalizes overconfidence):

$$LL = -\frac{1}{S} \sum_{s=1}^{S} \left[o_s \log(p_s + \epsilon) + (1 - o_s) \log(1 - p_s + \epsilon) \right] \tag{3}$$

where $\epsilon = 10^{-10}$ prevents numerical issues.

Mean Absolute Error (point prediction accuracy):

$$MAE = \frac{1}{S} \sum_{s=1}^{S} |\hat{m}_s - m_s^{true}| \tag{4}$$

Lower is better for all three metrics.

2 Model 1: Poll Average (Baseline)

2.1 Motivation

The simplest approach: average recent polls, weighted by sample size. This represents what a journalist would do manually without statistical modeling.

Algorithm 1 Poll Average Forecasting

Input: Polls $\{(t_i, y_i, n_i)\}$ up to $t_{forecast}$, days to election D
Parameters: Window $W = 14$ days

Filter polls: $\mathcal{I} = \{i : t_i \geq t_{forecast} - W\}$

if $|\mathcal{I}| < 3$ **then**
    Use last 5 polls instead
end if

Compute weights: $w_i = n_i$ for $i \in \mathcal{I}$, normalize: $w_i \leftarrow w_i / \sum_j w_j$

Predicted margin: $\hat{m} = \sum_{i \in \mathcal{I}} w_i y_i$

Empirical std: $s = \sqrt{\sum_{i \in \mathcal{I}} (y_i - \hat{m})^2 / (|\mathcal{I}| - 1)}$

Average sample size: $\bar{n} = \sum_{i \in \mathcal{I}} w_i n_i$

Sampling std: $\sigma_{samp} = 1/\sqrt{\bar{n}}$

Base uncertainty: $\sigma_{base} = \max(s, \sigma_{samp}, 0.02)$

Horizon uncertainty: $\sigma_{horizon} = 0.001 \cdot D$

Total uncertainty: $\sigma = \sqrt{\sigma_{base}^2 + \sigma_{horizon}^2}$

Win probability: $p = \Phi(\hat{m}/\sigma)$, clipped to $[0.05, 0.95]$

return $(p, \hat{m}, \sigma)$

2.2 Algorithm

2.3 Mathematical Justification

Weighted average as MLE:

Assume each poll y_i is an independent draw:

$$y_i \sim \mathcal{N}(\mu, \sigma_i^2) \tag{5}$$

where $\sigma_i^2 = \frac{1}{4n_i}$ (approximate sampling variance for a proportion).

The maximum likelihood estimator for μ with known variances is:

$$\hat{\mu}_{MLE} = \frac{\sum_i w_i y_i}{\sum_i w_i}, \quad w_i = \frac{1}{\sigma_i^2} = 4n_i \tag{6}$$

For simplicity, we use $w_i = n_i$ (proportional weighting).

Forecast horizon uncertainty:

The term $\sigma_{horizon} = 0.001 \cdot D$ adds uncertainty that grows linearly with days until election. This captures:

- Possible shifts in voter sentiment
- Systematic polling errors
- Late-deciding voters

This linear scaling is implemented in the code as:

```
horizon_uncertainty = 0.001 * days_to_election
total_std = np.sqrt(total_std**2 + horizon_uncertainty**2)
```

2.4 Strengths and Weaknesses

Strengths:

- Simple and transparent
- No parameters to tune (besides window size)
- Robust to model misspecification
- Fast to compute

Weaknesses:

- Ignores temporal structure (treats all polls in window equally)
- Doesn't model pollster house effects
- Arbitrary window size choice
- Wastes older polling data

3 Model 2: Kalman Diffusion

3.1 Motivation

Model the latent "true" vote margin X_t as evolving continuously over time, observed through noisy polls. This allows us to:

- Use all available polls (not just recent ones)
- Smooth out polling noise
- Account for pollster-specific biases
- Make principled forecasts with uncertainty quantification

3.2 Model Specification

Latent state process (Brownian motion with drift):

$$dX_t = \mu\, dt + \sigma\, dW_t \tag{7}$$

where W_t is standard Brownian motion.

In discrete time ($\Delta t = 1$ day):

$$X_{t+1} | X_t \sim \mathcal{N}(X_t + \mu, \sigma^2) \tag{8}$$

Parameters:

- μ: daily drift (trend toward one candidate)
- σ^2: daily diffusion variance (day-to-day volatility)

Observation model:

Poll i at time t_i by pollster $p(i)$ observes:

$$y_i = X_{t_i} + b_{p(i)} + \eta_i \tag{9}$$

where:

- b_p: systematic bias (house effect) of pollster p
- $\eta_i \sim \mathcal{N}(0, \tau_i^2)$: measurement error

Observation variance:

$$\tau_i^2 = \frac{1}{n_i} + \tau_{extra}^2 \tag{10}$$

where $\tau_{extra}^2 = (0.015)^2$ accounts for methodology differences beyond sampling error.

3.3 Parameter Estimation: EM Algorithm

We estimate $\theta = \{\mu, \sigma^2, \{b_p\}_{p \in \mathcal{P}}\}$ using the Expectation-Maximization (EM) algorithm.

3.3.1 E-step: Kalman Filter + RTS Smoother

Given current parameters $\theta^{(k)}$, compute the posterior distribution of the latent path $\{X_t\}$ given observations $\{y_i\}$.

State-space form:

$$X_t = X_{t-1} + \mu\Delta t_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2 \Delta t_t) \tag{11}$$

$$y_t = X_t + \nu_t, \quad \nu_t \sim \mathcal{N}(0, \tau_t^2) \tag{12}$$

Kalman filter (forward pass):
Initialize: $X_{0|0} = y_0$, $P_{0|0} = \tau_0^2$ (observe first poll)
For $t = 1, \ldots, T$:

$$\text{Predict:} \quad X_{t|t-1} = X_{t-1|t-1} + \mu\Delta t_t \tag{13}$$

$$P_{t|t-1} = P_{t-1|t-1} + \sigma^2 \Delta t_t \tag{14}$$

$$\text{Update:} \quad K_t = \frac{P_{t|t-1}}{P_{t|t-1} + \tau_t^2} \quad \text{(Kalman gain)} \tag{15}$$

$$X_{t|t} = X_{t|t-1} + K_t(y_t - X_{t|t-1}) \tag{16}$$

$$P_{t|t} = (1 - K_t)P_{t|t-1} \tag{17}$$

RTS smoother (backward pass):
Initialize: $X^s_{T|T} = X_{T|T}$, $P^s_{T|T} = P_{T|T}$
For $t = T-1, \ldots, 0$:

$$J_t = \frac{P_{t|t}}{P_{t|t} + \sigma^2 \Delta t_{t+1}} \tag{18}$$

$$X^s_{t|T} = X_{t|t} + J_t(X^s_{t+1|T} - X_{t|t} - \mu\Delta t_{t+1}) \tag{19}$$

$$P^s_{t|T} = P_{t|t} + J_t^2(P^s_{t+1|T} - P_{t|t} - \sigma^2 \Delta t_{t+1}) \tag{20}$$

Output: smoothed estimates $\{X^s_{t|T}, P^s_{t|T}\}$ given all data.

3.3.2 M-step: Parameter Updates

Update pollster biases:
For each pollster p, with shrinkage $\lambda = 0.5$:

$$b_p^{new} = \begin{cases} \lambda \cdot \frac{1}{N_p} \sum_{i:p(i)=p} (y_i - X^s_{t_i|T}) & \text{if } N_p \geq 2 \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

Update drift:

$$\mu^{new} = \frac{1}{\sum_{t=1}^{T-1} \Delta t_t} \sum_{t=1}^{T-1} (X^s_{t+1|T} - X^s_{t|T}) \tag{22}$$

Update diffusion variance:

$$(\sigma^2)^{new} = \max\left(\frac{1}{\sum_{t=1}^{T-1} \Delta t_t} \sum_{t=1}^{T-1} \left(P^s_{t|T} + P^s_{t+1|T}\right), 0.0005 \right) \tag{23}$$

with floor 0.0005 to prevent degeneracy.
Iterate E and M steps for $K = 10$ iterations.

3.4 Forecasting

Given the posterior at the last observation time t_{last}, forecast to election day T_{elec} using Monte Carlo simulation.

Incorporating fundamentals prior:

Apply prior weight $w = 0.1$:

$$X_{last}^{adj} = (1 - w)X_{last|T}^{s} + w \cdot m_{prior} \tag{24}$$

Forecast horizon uncertainty:

Add additional variance growing with forecast horizon:

$$P_{last}^{adj} = P_{last|T}^{s} + (0.001 \cdot D)^2 \tag{25}$$

where $D = T_{elec} - t_{last}$ is days to election.

Euler-Maruyama simulation:

For $j = 1, \dots, N_{sim}$ (e.g., $N_{sim} = 2000$):

1. Initialize: $X_0^{(j)} \sim \mathcal{N}(X_{last}^{adj}, P_{last}^{adj})$
2. For $d = 1, \dots, D$:
$$X_d^{(j)} = X_{d-1}^{(j)} + \mu + \sigma Z_d, \quad Z_d \sim \mathcal{N}(0, 1) \tag{26}$$
3. Record: $M^{(j)} = X_D^{(j)}$ (simulated election margin)

Output:

- Win probability: $\hat{p} = \frac{1}{N_{sim}} \sum_{j=1}^{N_{sim}} \mathbb{K}\{M^{(j)} > 0\}$, clipped to $[0.01, 0.99]$
- Expected margin: $\hat{m} = \frac{1}{N_{sim}} \sum_j M^{(j)}$
- Uncertainty: $\hat{\sigma} = \sqrt{\frac{1}{N_{sim}-1} \sum_j (M^{(j)} - \hat{m})^2}$

3.5 Implementation Details

Using recent polls only:

To reduce computational cost and avoid overfitting to early polls, use only the most recent $\lceil N_s/3 \rceil$ polls (at least 10).

3.6 Strengths and Weaknesses

Strengths:

- Principled probabilistic framework
- Uses all polling data efficiently
- Accounts for pollster house effects

- Quantifies uncertainty naturally

Weaknesses:

- Can overfit with sparse data
- Pollster bias estimates unstable
- Assumes constant drift/diffusion
- No cross-state correlation

4 Model 3: Improved Kalman

4.1 Motivation

The basic Kalman model suffers from overfitting due to:

1. Estimating separate bias for each pollster (data is sparse)
2. Insufficient diffusion variance (overconfidence)
3. Too much forecast horizon uncertainty growth

The improved model adds regularization and better uncertainty calibration.

4.2 Key Modifications

1. Stronger regularized pollster biases:
Shrinkage increased to $\lambda = 0.7$ (was 0.5):

$$b_p^{new} = \begin{cases} 0.7 \cdot \frac{1}{N_p} \sum_{i:p(i)=p} (y_i - X^s_{t_i|T}) & \text{if } N_p \geq 2 \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

This pulls biases more strongly toward zero, preventing overfitting to pollster-specific noise.

2. Increased minimum diffusion:

$$\sigma^2 \geq (0.0008)^2 = 0.00000064 \quad \text{(was 0.0005)} \tag{28}$$

3. Reduced forecast horizon growth:

$$P^{adj}_{last} = P^s_{last|T} + (0.0005 \cdot D)^2 \quad \text{(was 0.001)} \tag{29}$$

4. Same probability clipping:

$$p \in [0.01, 0.99] \tag{30}$$

4.3 Mathematical Justification

Bayesian interpretation of shrinkage:

The MLE bias estimate is:

$$\hat{b}_p^{MLE} = \frac{1}{N_p} \sum_i r_i \tag{31}$$

Under a hierarchical prior $b_p \sim \mathcal{N}(0, \tau_b^2)$, the posterior mean is:

$$\mathbb{E}[b_p|\text{data}] = \frac{N_p}{N_p + \sigma^2/\tau_b^2} \cdot \hat{b}_p^{MLE} \tag{32}$$

Setting $\lambda = 0.7$ approximates this with $\tau_b^2 \approx 0.43\sigma^2/N_p$.

4.4 Empirical Performance

On the 2016 election:

- Improved Kalman achieves better calibration than basic Kalman
- Brier score on final forecast (Nov 7): 0.127 vs 0.145
- Still underperforms simple poll average (0.079)
- Main issue: insufficient data to reliably estimate 30+ pollster biases

5 Model 4: Hierarchical Bayes (HBE-SBA)

5.1 Motivation

The winning model. Combines three key ideas:

1. **Fundamentals prior**: Historical election results provide informative prior
2. **Hierarchical pollster effects**: Pool information across pollsters
3. **Systematic bias adjustment**: Correct for cycle-wide polling errors

5.2 Model Architecture

Three-layer hierarchical model:

Layer 1: Fundamentals (prior)

$$X_0^{(s)} \sim \mathcal{N}(m_{fund}^{(s)}, \sigma_{fund}^2) \tag{33}$$

where $m_{fund}^{(s)}$ is the weighted historical margin from the fundamentals file.

Layer 2: Latent process (Kalman)

$$dX_t^{(s)} = \mu^{(s)} dt + \sigma^{(s)} dW_t \tag{34}$$

$$y_i^{(s)} = X_{t_i}^{(s)} + h_{p(i)} + \eta_i \tag{35}$$

where h_p is pollster p's house effect.

Layer 3: Hierarchical house effects

$$h_p \sim \mathcal{N}(0, \tau_h^2) \tag{36}$$

All pollsters share a common variance τ_h^2, estimated from data.

5.3 House Effect Estimation

Instead of estimating h_p separately per state, we pool across all states:

Algorithm:

1. For each pollster p, collect all their polls across all states
2. For each poll i by p in state s at time t_i:
   - Find other polls in same state within ± 7 days
   - Compute local average $\bar{y}_{s,t_i}^{-p}$ excluding pollster p
   - Residual: $r_i = y_i - \bar{y}_{s,t_i}^{-p}$
3. Average residuals: $\bar{r}_p = \frac{1}{N_p} \sum_i r_i$
4. Apply hierarchical shrinkage:

$$h_p = \frac{N_p}{N_p + \lambda} \bar{r}_p, \quad \lambda = 10 \tag{37}$$

Interpretation:
The shrinkage factor $\frac{N_p}{N_p+\lambda}$ pulls estimates toward zero:

- Pollster with $N_p = 10$ polls: shrinkage factor $= 0.5$
- Pollster with $N_p = 100$ polls: shrinkage factor $= 0.91$

This is equivalent to empirical Bayes estimation under $h_p \sim \mathcal{N}(0, \tau_h^2)$ with $\lambda = \sigma_\eta^2 / \tau_h^2$.

5.4 State-Level Forecasting

For each state at forecast date $t_{forecast}$:

1. Fundamentals Prior:

$$m_{fund} = \text{historical margin from fundamentals file} \quad (38)$$

$$\sigma^2_{fund} = (0.08)^2 + (0.0015 \cdot D)^2 \quad (39)$$

where D is days to election.

2. Process Recent Polls:

Use polls from last 45 days (at least 10 polls). Apply house effect correction:

$$y_i^{corrected} = y_i - h_{p(i)} \quad (40)$$

3. Kalman Filter Estimation:

Apply Kalman filter + RTS smoother with:

- Observation variance: $\tau_i^2 = \frac{1}{n_i} + (0.015)^2$
- Daily diffusion: $\sigma^2 = (0.003)^2$
- Drift: $\mu = \frac{\bar{y}_{final} - \bar{y}_{initial}}{\Delta t}$ (simple estimate)

Output: X^s_{poll}, P^s_{poll}

4. Bayesian Combination:

Time-adaptive prior weight (decreases as election approaches):

$$w_{prior} = \frac{0.3}{1 + (days\_elapsed/21)^2} \quad (41)$$

where $days\_elapsed$ is days since September 1.

Precision-weighted combination:

$$\pi_{prior} = \frac{w_{prior}}{\sigma^2_{fund}} \quad (42)$$

$$\pi_{polls} = \frac{1}{P^s_{poll}} \quad (43)$$

$$X_{combined} = \frac{m_{fund} \cdot \pi_{prior} + X^s_{poll} \cdot \pi_{polls}}{\pi_{prior} + \pi_{polls}} \quad (44)$$

$$P_{combined} = \frac{1}{\pi_{prior} + \pi_{polls}} \quad (45)$$

5. Systematic Bias Correction:

Adaptive bias learning (ramps up over time):

$$w_{learning} = \min(1.0, days\_elapsed/30) \quad (46)$$

Simple bias model (calibrated to 2016 patterns):

$$\delta = w_{learning} \cdot (0.02 - 0.03 \cdot pvi) \tag{47}$$

where pvi is the state's partisan lean from fundamentals.

Apply correction:

$$X_{corrected} = X_{combined} - \delta \tag{48}$$

6. Forecast Uncertainty:

Total variance combines multiple sources:

$$\sigma^2_{evolution} = (0.003 \cdot D)^2 \quad \text{(future diffusion)} \tag{49}$$

$$\sigma^2_{bias} = (0.04)^2 \quad \text{(systematic bias uncertainty)} \tag{50}$$

$$\sigma^2_{total} = P_{combined} + \sigma^2_{evolution} + \sigma^2_{bias} \tag{51}$$

7. Win Probability:

Using normal CDF:

$$p = \Phi\left(\frac{X_{corrected}}{\sigma_{total}}\right), \quad \text{clipped to } [0.02, 0.98] \tag{52}$$

5.5 Why This Works

Compared to Poll Average:

- Uses fundamentals to anchor estimates (prevents wild swings)
- Properly accounts for pollster house effects
- Models temporal evolution (not just recent window)

Compared to Basic Kalman:

- Hierarchical house effects prevent overfitting
- Fundamentals prior regularizes estimates
- Systematic bias correction handles 2016-style errors

5.6 Empirical Results

On 2016 election (average across 4 forecast dates):

Model	Brier	Log Loss	MAE	Rank
Hierarchical Bayes	**0.090**	**0.291**	0.071	**1.33**
Poll Average	0.091	0.304	**0.067**	1.67
Improved Kalman	0.145	0.535	0.076	3.0
Kalman Diffusion	0.188	0.680	0.320	4.0

Key finding: Hierarchical Bayes is the first model to outperform the simple baseline, winning on Brier score and Log Loss.

6 Comparative Analysis

6.1 Model Complexity vs Performance

Model	**Parameters**	**Final Brier (Nov 7)**
Poll Average	2	0.079
Kalman Diffusion	$2 + \|\mathcal{P}_s\|$	0.145
Improved Kalman	$2 + \|\mathcal{P}_s\|$	0.127
Hierarchical Bayes	$2 + \|\mathcal{P}\| + 2$	**0.061**

Where $|\mathcal{P}_s|$ = number of pollsters in state s (typically 5-15), $|\mathcal{P}|$ = total unique pollsters (30-40).

Observation: Adding parameters helps only with proper regularization. Improved Kalman and Hierarchical Bayes succeed by:

- Sharing information across states/pollsters
- Using informative priors
- Applying shrinkage

6.2 Temporal Evolution

Performance as election approaches:

Model	**Aug 1**	**Sep 23**	**Oct 15**	**Nov 7**	**Trend**
Poll Average	0.096	0.077	0.111	0.079	Stable
Kalman Diffusion	0.236	0.209	0.159	0.145	Improving
Improved Kalman	0.167	0.146	0.139	0.127	Improving
Hierarchical Bayes	0.118	0.091	0.092	0.061	**Strong improvement**

Insight: Sophisticated models improve more as data accumulates. Simple poll average is consistently good but doesn't improve as much.

6.3 When to Use Each Model

Poll Average:

+ Few polls available (early in cycle)
+ Need transparency/explainability
+ No computational resources

− Can't incorporate fundamentals

Kalman Diffusion:

+ Want to model temporal dynamics

+ Many polls over time

− Requires careful tuning

− Can overfit with sparse data

Hierarchical Bayes:

+ Sufficient data across multiple states/pollsters

+ Have fundamentals/historical data

+ Late in election cycle

− More complex to implement

7 Limitations and Extensions

7.1 Shared Limitations

All models:

1. **Single-cycle data**: Cannot learn systematic polling biases without multi-cycle training
2. **Independence assumption**: States modeled separately (reality: correlated shocks)
3. **Stationary parameters**: μ, σ^2 constant (reality: campaign dynamics change)
4. **No late shift**: Cannot predict October surprises or late momentum

7.2 Possible Extensions

1. Multivariate state-space model
Model all states jointly with correlated dynamics:

$$dX_t = (\mu + BF_t)dt + \Sigma^{1/2} dW_t \tag{53}$$

where F_t are national factors (economy, approval rating) and B is state-specific factor loadings.

2. Time-varying parameters
Allow drift to change around events:

$$\mu_t = \mu_0 + \sum_k \beta_k \mathbb{1}\{t \geq t_{event_k}\} \tag{54}$$

where t_{event_k} are debate dates, convention dates, etc.

3. Non-Gaussian errors

Replace Normal with Student-t for robustness to outlier polls:

$$\eta_i \sim t_\nu(0, \tau_i^2) \tag{55}$$

4. Turnout modeling
Current models predict vote margin among voters. Could extend to predict turnout:

$$\text{Votes}_i = \text{VEP}_i \cdot \text{Turnout}_i \cdot \text{Vote share}_i \tag{56}$$

5. Systematic bias learning
With multiple cycles of data, estimate δ from historical forecast errors:

$$\delta \sim \mathcal{N}(\mu_\delta, \sigma_\delta^2) \tag{57}$$

where μ_δ, σ_δ^2 are learned from past elections.

8 Conclusion

We have presented four increasingly sophisticated models for election forecasting:

1. **Poll Average**: Simple and robust baseline
2. **Kalman Diffusion**: Principled state-space model with EM estimation
3. **Improved Kalman**: Adds regularization for better calibration
4. **Hierarchical Bayes**: Winner, combining fundamentals, hierarchy, and bias correction

Key lessons:

- Sophistication without regularization hurts (Kalman vs Poll Average)
- Proper Bayesian hierarchy helps (HBE vs Improved Kalman)
- Fundamentals matter for anchoring (HBE's secret weapon)
- Systematic bias correction is critical for 2016

The Hierarchical Bayes model achieves state-of-the-art performance while remaining interpretable and principled, demonstrating that careful statistical modeling can beat simple aggregation—but only with appropriate regularization.