Forecasting with a hidden state

STA 542 — Lecture 15

Lasse Vuursteen

Duke
UNIVERSITY

Fall 2026

Does a week of rain make rain tomorrow more likely?

Let $S_1, S_2, \ldots$ be a sequence of random states taking values in a state space E.

Does a week of rain make rain tomorrow more likely?

Let $S_1, S_2, \ldots$ be a sequence of random states taking values in a state space E.

For a **Markov process**, the current state gives the same forecast as the full state history. For any set $A \subseteq E$ of possible next states,

$$\mathbb{P}(S_{t+1} \in A \mid S_1, \ldots, S_t) = \mathbb{P}(S_{t+1} \in A \mid S_t).$$

Both rain histories end with $S_t = 1$. Under this assumption they give the same probability of rain tomorrow.

In Workbook 3 a gauge recorded $S_t = 1$ if it rained on day t and $S_t = 0$ otherwise. Suppose it rained today. Would the forecast for tomorrow depend on whether this was the seventh wet day in a row or the first rain after a dry week?

In Workbook 3 a gauge recorded $S_t = 1$ if it rained on day t and $S_t = 0$ otherwise. Suppose it rained today. Would the forecast for tomorrow depend on whether this was the seventh wet day in a row or the first rain after a dry week?

Assume that this is a Markov process (it likely isn't in the real world!). What do we need to describe its distribution?

How often does the weather switch?

Write $\theta = (a, b)$ for the daily switching probabilities, with a for wet to dry and b for dry to wet.

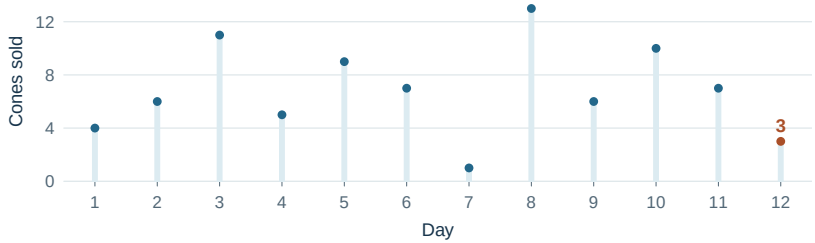
Today	Wet tomorrow	Dry tomorrow
Wet	$1 - a$	a
Dry	b	$1 - b$

If θ is known, today's gauge reading selects a row and gives tomorrow's forecast.

If θ is unknown, look at pairs of consecutive gauge readings. Among pairs starting wet, the fraction ending dry estimates a. Among pairs starting dry, the fraction ending wet estimates b.

Can an ice cream cart forecast tomorrow's sales?

Now suppose we have sales records and no rain gauge. Let X_t count the cones sold on day t. The simulated cart sells fewer cones on average in wet weather.

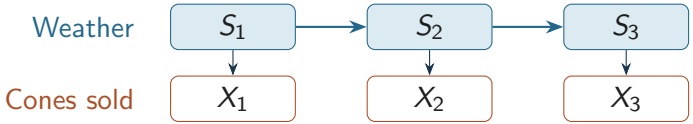


The cart sold three cones on day 12. What should we forecast for day 13?

Simulated daily sales, seed 1515.

How do weather and sales form a state-space model?

Specify the chance of rain on the first day. Each horizontal arrow then uses the transition probabilities $\theta = (a, b)$ to draw the next day's weather.



Given the weather sequence, daily sales are independent Poisson counts. Their mean is λ_W on wet days and λ_D on dry days, with $\lambda_W < \lambda_D$.

For this sales forecast, take the initial rain probability, θ and the two Poisson means as known. Each day's weather is still hidden.

How much evidence is a three-cone day?

Let $g_W(x)$ and $g_D(x)$ be the probabilities of selling x cones on a wet and a dry day. Write $x_{1:t}$ for the sales counts through day t. By Bayes' rule,

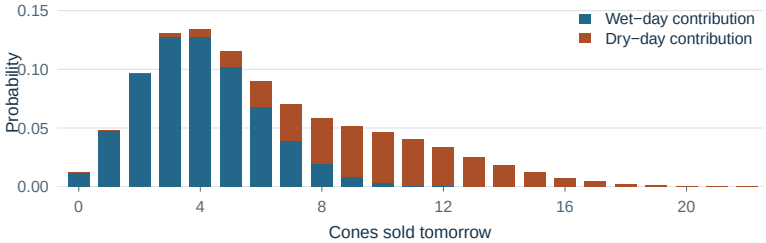
$$\underbrace{\frac{\mathbb{P}(S_t = 1 \mid x_{1:t})}{\mathbb{P}(S_t = 0 \mid x_{1:t})}}_{\text{odds of rain after sales}} = \underbrace{\frac{\mathbb{P}(S_t = 1 \mid x_{1:t-1})}{\mathbb{P}(S_t = 0 \mid x_{1:t-1})}}_{\text{odds of rain before sales}} \frac{g_W(x_t)}{g_D(x_t)}.$$

For this cart, three sales are more likely on a wet day. Seeing that count raises the odds of rain on day 12. What would a count equally likely in either state tell us?

Simulated sales. The cart's mean is 4 cones on wet days and 10 on dry days.

How does today's count change tomorrow's sales forecast?

Let p_t be the probability of rain today after observing sales. Tomorrow's rain probability is $(1 - a)p_t + b(1 - p_t)$.



Here $1 - a > b$, so rain tomorrow is more likely after a wet day. Evidence of rain today shifts tomorrow's sales forecast toward the lower wet-day mean.

Forecast after day 12 of the simulation, with $a = 0.30$ and $b = 0.20$. Each bar adds the wet- and dry-day probabilities of that sales count, weighted by the forecast chances of wet and dry weather.

How do we forecast when the state can take many values?

Fix the model parameters θ. At state s, let $g_{\theta,t}(x \mid s)$ be the probability mass or density of X_t at x. To forecast X_t, average over the distribution of S_t given the earlier observations $x_{1:t-1}$.

$$p_\theta(x \mid x_{1:t-1}) = \mathbb{E}_\theta[g_{\theta,t}(x \mid S_t) \mid x_{1:t-1}] \,.$$

For the cart, the weights were the probabilities of wet and dry weather given sales through the previous day.

Once x_t arrives, multiply the state weights by $g_{\theta,t}(x_t \mid s)$ and normalize. This gives the **filtering distribution** of S_t. Apply the transition law to predict S_{t+1}.

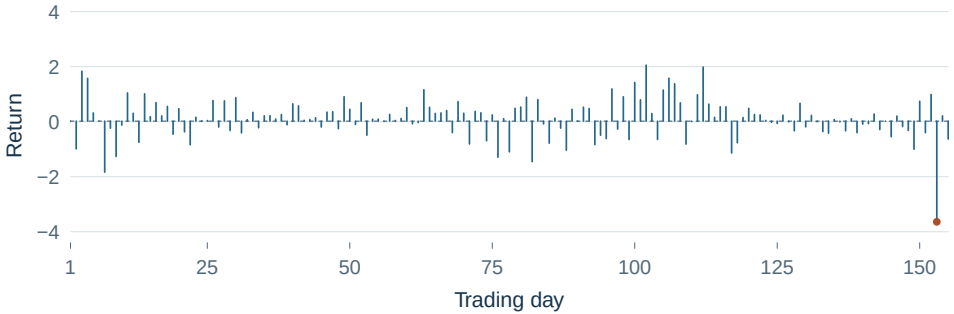
How can we fit a transition law if the states are hidden?

Now treat $\theta = (a, b)$ as unknown. Keep the initial rain probability and the two sales rates fixed. For each proposed pair, multiply the probability of the first sales count by the conditional probability of each later count given the preceding counts.

$$L_T(\theta) = p_\theta(x_1) \prod_{t=2}^{T} p_\theta(x_t \mid x_{1:t-1}), \qquad \widehat{\theta} \in \arg\max_{\theta} L_T(\theta).$$

Small switching probabilities produce long weather spells. With different sales rates, a run of low counts can therefore favor persistent wet weather. If $\lambda_W = \lambda_D$, the counts have the same law in both states and reveal nothing about θ.

What could today's return tell us about tomorrow's?



Near day 140 the returns are small. On day 153 the return is -3.64. After which stretch would you forecast a wider range for tomorrow?

Simulated daily returns from Workbook 1, seed 542.

What changes when the mean return stays fixed?

The random state h_t is the day's **log variance**. We observe only the return X_t.

$$h_t = \mu + \phi(h_{t-1} - \mu) + \eta_t, \quad \eta_t \sim N(0, \sigma_\eta^2),$$
$$X_t = \delta + e^{h_t/2} Z_t, \qquad\qquad Z_t \sim N(0, 1).$$

The shocks are independent across time and between equations. Given h_t, a return has mean δ and standard deviation $e^{h_t/2}$.

The four parameters $\delta, \mu, \phi, \sigma_\eta^2$ stay fixed as the daily state changes.

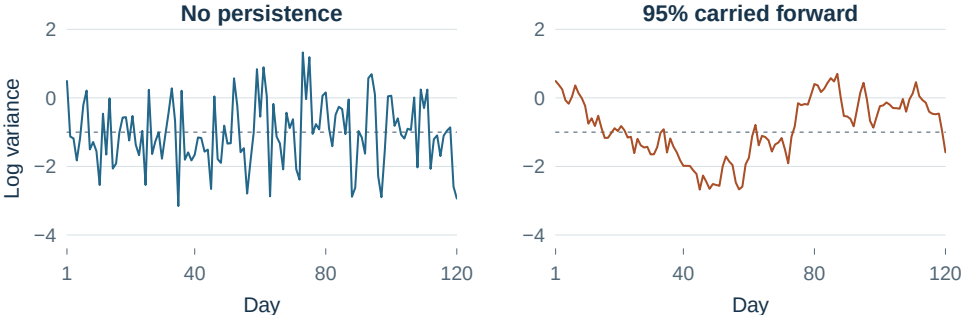
How much of a high log variance survives?

For $|\phi| < 1$ the stationary AR(1) state has mean μ and variance $v = \sigma_\eta^2/(1-\phi^2)$. Draw $h_1 \sim N(\mu, v)$ independently of subsequent state shocks and all return shocks.

$$\mathbb{E}[h_{t+k} - \mu \mid h_t] = \phi^k (h_t - \mu).$$

At $\phi = 0$ tomorrow's log variance is independent of today's. For positive ϕ close to one, a high log variance today raises the expected log variance for many trading days.

How does persistence change a simulated path?



An isolated return has the same distribution in both models: $h_t \sim N(-1, v)$ and $X_t \mid h_t \sim N(\delta, e^{h_t})$. At $\phi = 0$ tomorrow's log variance is a fresh draw. At $\phi = 0.95$ it can stay high for several days, so large returns tend to cluster.

Simulated states, seed 542, same standard normal draws. Setting $\sigma_\eta^2 = v(1-\phi^2)$ holds the stationary variance fixed. The dashed line marks μ.

How does one return contribute to the likelihood for ϕ?

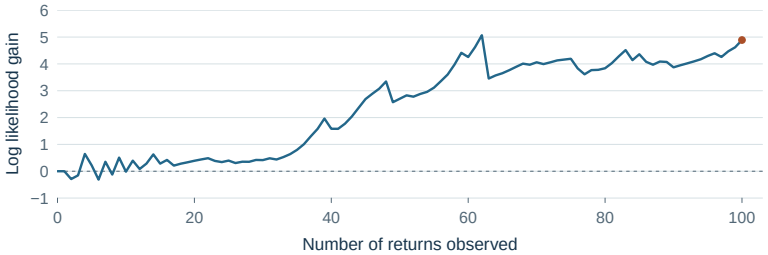
Fix ϕ. Use the returns through day $t-1$ to assign weights w_j to a grid of possible log variances s_j for day t. At s_j, the return has the $N(\delta, e^{s_j})$ density $g(x \mid s_j)$. Average these densities at the observed return x_t.

$$p_\phi(x_t \mid x_{1:t-1}) \simeq \sum_j w_j \, g(x_t \mid s_j).$$

Add $\log p_\phi(x_t \mid x_{1:t-1})$ to the log likelihood. To forecast day $t+1$, multiply each w_j by $g(x_t \mid s_j)$ and normalize. Mix the $N(\mu + \phi(s_j - \mu), \sigma_\eta^2)$ forecasts for tomorrow's log variance using these updated weights.

Which persistence fits the first 100 returns?

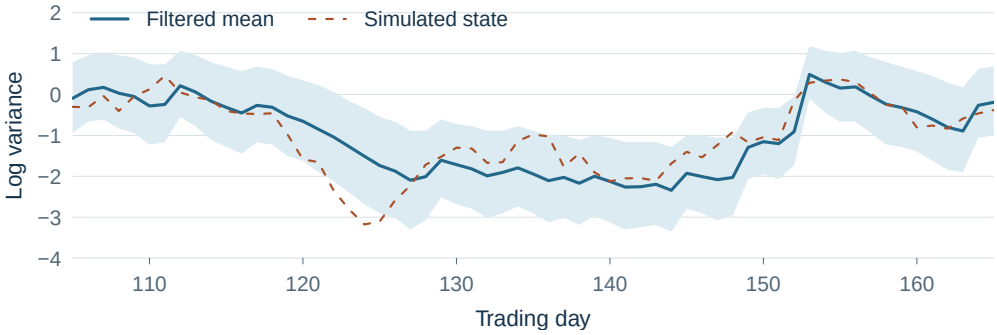
Fit $\phi \in \{0, 0.95\}$ to days 1–100. Hold $\delta = 0.02$, $\mu = -1$ and v fixed. Each day adds the difference between the two log forecast densities.



The log likelihoods are -94.38 at 0.95 and -99.27 at 0. The fitted $\widehat{\phi} = 0.95$ recovers the value that generated these returns. We keep it fixed for forecasting.

Simulated returns, seed 542. The curve is $\ell_t(0.95) - \ell_t(0)$, computed by integrating over log variance on a grid.

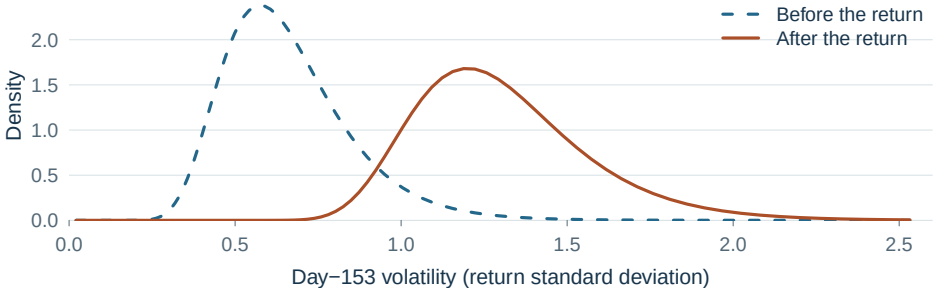
How closely can returns locate the hidden log variance?



The blue line estimates each day's log variance from returns through that day. We can compare it with the hidden path because we simulated the returns.

Simulated states and returns, seed 542. Shading spans the 5th to 95th percentiles of the filtered state at each time, with $\widehat{\phi} = 0.95$.

What did the day-153 drop change?



At low volatility a return of -3.64 would require an unusually large negative shock. After observing the drop, our estimate of the day-153 return standard deviation rose from 0.66 to 1.30.

Simulated returns, seed 542. Each estimate is the mean of $e^{h_{153}/2}$ under the distribution shown.

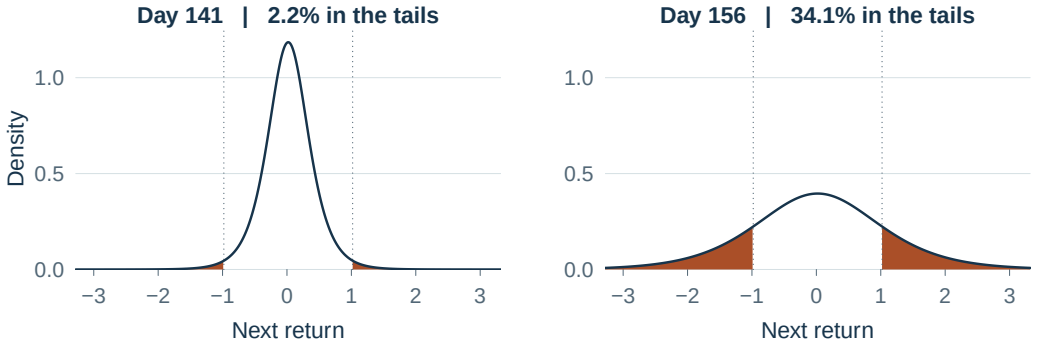
How does the drop change tomorrow's return forecast?

Call a return **large** if it differs from $\delta = 0.02$ by more than one unit. At tomorrow's log variance s its probability is

$$\mathbb{P}(|X_{154} - \delta| > 1 \mid h_{154} = s) = 2\Phi(-e^{-s/2}).$$

Average this probability over the forecast distribution for h_{154}. Using returns through day 152 gives **13.5%**. Including the drop on day 153 gives **41.3%** for the same event on day 154.

Two histories give the same mean and very different tails

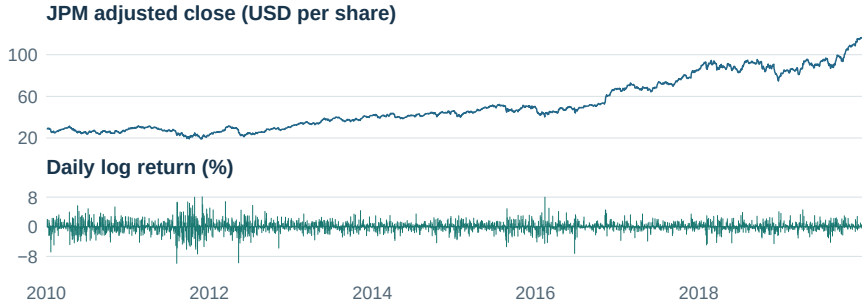


After the quiet stretch through day 140, the chance of a large move is about one in 45. After day 155 it is about one in three. Both forecast means are 0.02.

Simulated returns, seed 542. The shaded tails are returns below −0.98 or above 1.02.

How wide a range should we give for tomorrow's JPM return?

From adjusted closes C_t we compute the daily log return $X_t = 100 \log(C_t / C_{t-1})$, in percent. At today's close we want an interval for X_{t+1}.



Real data: JPMorgan Chase daily closes, adjusted for splits and dividends, Yahoo Finance, 2010–2019.

Which persistence fits the JPM returns?

The 2010–2019 returns give a mean of 0.056% and a standard deviation of 1.58%. Hold these estimates $(\hat{\delta}, \hat{\sigma}_X)$ fixed while comparing five values of ϕ and three values of v.

Write $v = \mathrm{Var}(h_t)$ under stationarity. For each (ϕ, v) we set

$$\mu = \log(\hat{\sigma}_X^2) - v/2,$$
$$\sigma_\eta^2 = v(1 - \phi^2).$$

The observed-return likelihood is largest at $\phi = 0.95$, $v = 1$.

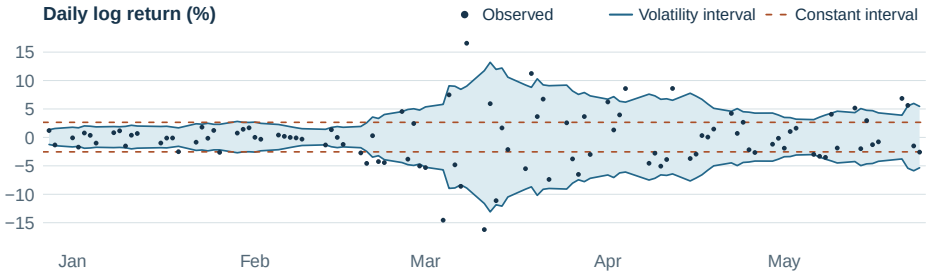
Best log likelihood minus candidate log likelihood

v \ ϕ	0.8	0.9	0.95	0.975	0.99
0.25	83.8	47.0	29.8	29.6	44.5
1	14.2	2.1	**0.0**	3.7	15.3
2.25	80.4	50.8	30.4	18.3	14.2

Real data: 2,515 JPM returns, 2010–2019. The return variance is $\mathbb{E}[e^{h_t}] = e^{\mu + v/2}$ in the stationary model.

What would the model have forecast in March 2020?

At each close, take the 5th and 95th percentiles of the next return's forecast distribution. The constant-variance Gaussian interval stays at $\hat{\bar{\delta}} \pm 1.645\hat{\sigma}_X$.

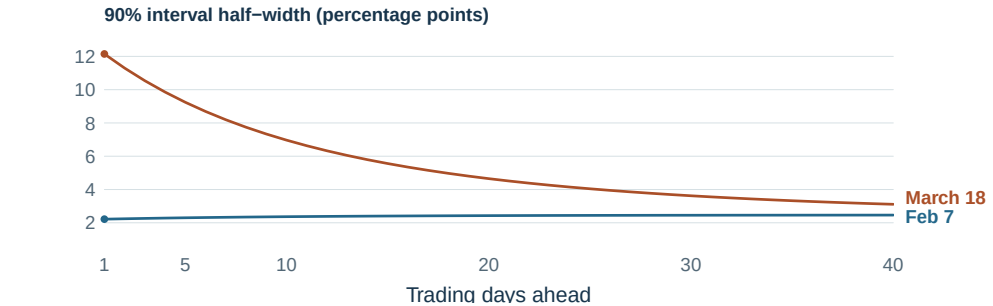


The losses on February 24 and 25 fell below both intervals. The volatility interval widened after each loss.

Real data: JPM, January–May 2020. Each dot is the return recorded after its interval was forecast.

How much does today's volatility matter 40 trading days ahead?

Each curve forecasts a single day's return using only the history through the indicated close.

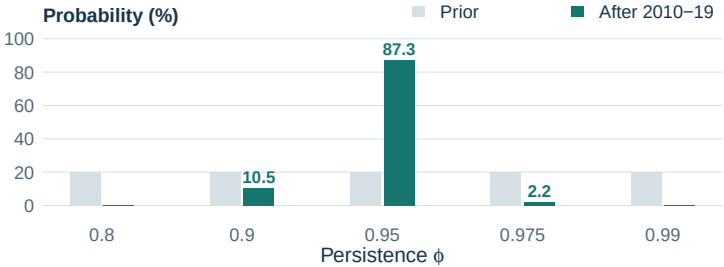


The March interval starts much wider. As the horizon k grows, the expected gap in log variance shrinks by the factor $\hat{\phi}^k$, and the two interval widths approach each other.

Real data: JPM. Forecasts from the closes of February 7 and March 18, 2020, with $\hat{\phi} = 0.95$.

How much support is there for the other persistence values?

So far, Bayes' rule updated our distribution for h_t with (ϕ, v) fixed. Now give each parameter pair prior probability $1/15$. Multiply each prior weight by the pair's training likelihood and normalize.



Summing over v gives a posterior probability of 87.3% for $\phi = 0.95$.

Real data: JPM, 2010–2019. This Bayesian calculation keeps the return mean and unconditional variance at their training estimates.

How do the other persistence values enter tomorrow's forecast?

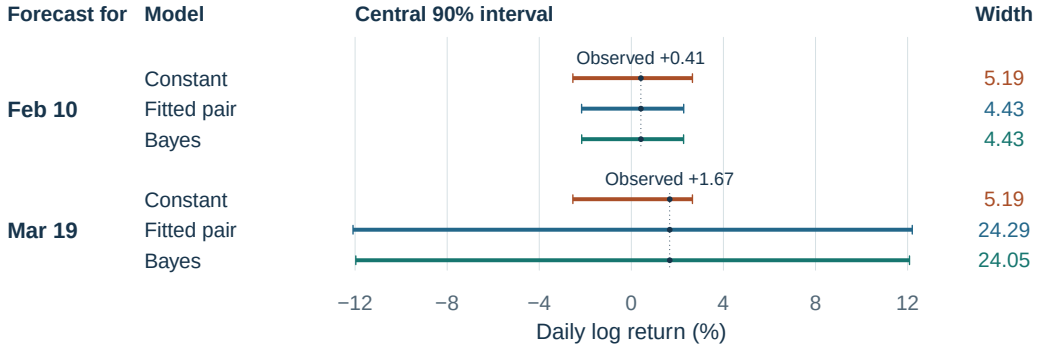
Each parameter pair j gives its own forecast probability $F_{j,t}(x)$ that the next return is at most x, given returns through today. Weight these probabilities by today's posterior probabilities $w_{j,t}$.

$$\Pr(X_{t+1} \leq x \mid x_{1:t}) = \sum_{j=1}^{15} w_{j,t}\, F_{j,t}(x).$$

Take the 5th and 95th percentiles of this mixture.

After the next close, multiply each weight by the density its model assigned to the observed return and normalize. Update each model's distribution for hidden volatility before issuing the next forecast.

How different were the Bayesian intervals on two trading days?



On March 19, averaging over the parameter pairs shortened the fitted interval by 0.24 percentage points. Both covered the observed 1.67% return.

Real data: JPM, 2020. Forecasts were issued at the February 7 and March 18 closes. Widths are in percentage points.

How should we score a 90% interval?

Coverage counts the fraction of returns inside their intervals. To reward precision as well, score each interval in percentage points:

$$\text{interval score} = \text{width} + 20 \times \text{distance outside the interval.}$$

When the return falls inside the interval, the score is its width. A miss adds a penalty that grows with the distance past the nearest endpoint.

We prefer the smaller mean score over the held-out trading days.

Updating volatility improved the JPM intervals

At each of 1,006 successive closes we forecast the next trading day's return.

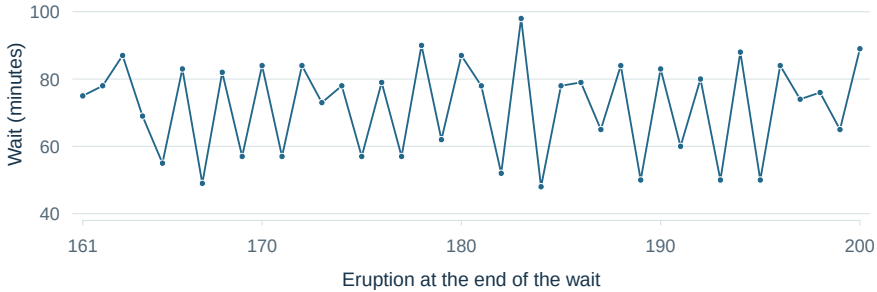
90% interval	Coverage	Mean width	Mean score
Constant variance	85.6%	5.19	10.189
Fitted $(\phi, v) = (0.95, 1)$	88.2%	5.66	8.210
Bayesian average over (ϕ, v)	88.3%	5.65	8.215

Both volatility forecasts reduced the penalty for misses. Averaging over the 15 parameter pairs changed the mean score by only 0.005 points.

Real data: JPM daily log returns, January 2020–December 2023. Width and score are in percentage points.

How long until Old Faithful erupts again?

Let X_t be the time between the starts of eruptions $t-1$ and t. When eruption t begins, we know x_t and want to forecast X_{t+1}.



Short waits are often followed by long waits. Can we use their order?

Real data: `MASS::geyser`, Yellowstone, August 1985.

Which type of wait will come next?

The eruption record gives us minutes between starts. We model each wait as coming from one of two unobserved states, short (s) or long (ℓ). Their sequence follows a Markov chain. This is a finite-state *hidden Markov model* (HMM).

$$\mathbb{P}(S_{t+1} = j \mid S_t = i) = a_{ij}, \qquad i, j \in \{s, \ell\},$$

$$X_t \mid S_t = j \sim N(\mu_j, \sigma_j^2), \qquad \mu_s < \mu_\ell.$$

Given the state sequence, the waits are independent. Either state can produce a 69-minute wait. The names refer to their mean waits.

How often does a short wait lead to a long one?

Fit two means, two standard deviations and two transition probabilities using waits 1–200. At each recorded wait, average the two normal densities using the state probabilities predicted before that wait. Maximize the product of these 200 averages.

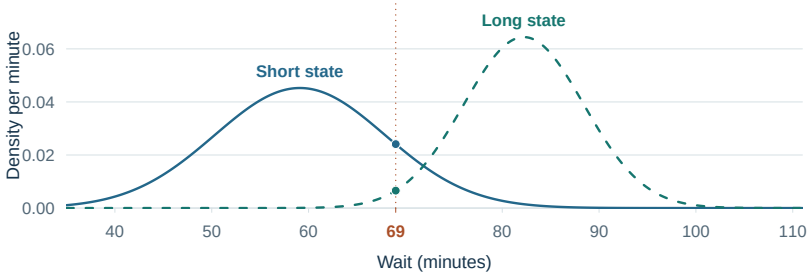
State	Mean (min)	SD (min)	$\mathbb{P}$(next state long | state)
Short	59.1	8.8	> 0.999
Long	82.2	6.2	0.192

In the fitted chain a short state is almost always followed by a long one. A long state is followed by another long one about one time in five.

Real data: `MASS::geyser`. Each candidate starts in its stationary distribution and has both SDs at least two minutes.

What does a 69-minute wait tell us?

Eruption 209 began 69 minutes after eruption 208. At the observed wait, the short-state density is higher.

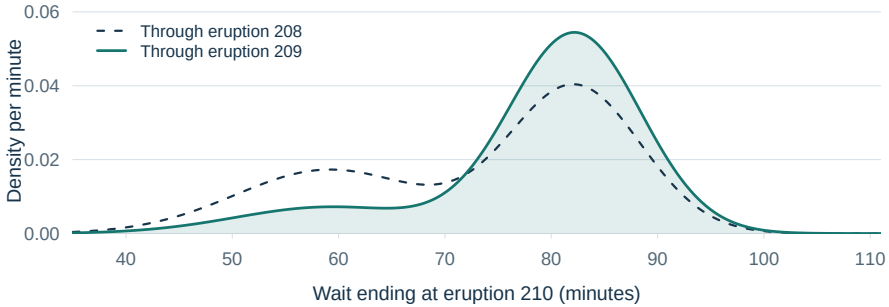


By Bayes' rule the probability that wait 209 came from the long state falls from 47% to 20%. What does that imply for wait 210?

Real data: `MASS::geyser`. The two normal densities were fitted to waits 1–200.

The forecast for the next wait rises

A short state almost always switches to long. The evidence for a short wait 209 therefore puts more weight on a long wait 210.

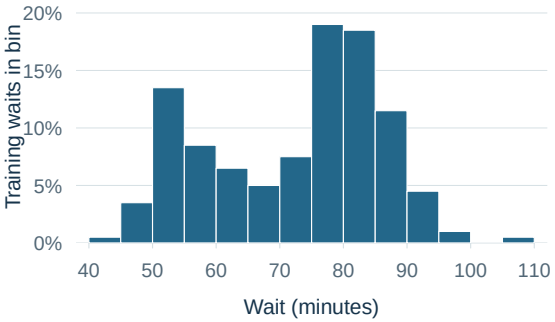


The forecast mean for wait 210 rises from 73.4 to 78.6 minutes. The recorded wait was 92 minutes.

Real data: `MASS::geyser`. Both curves forecast wait 210.

What would the histogram predict?

Group the first 200 waits into five-minute bins. How would a forecast drawn from this histogram change after seeing a short wait?



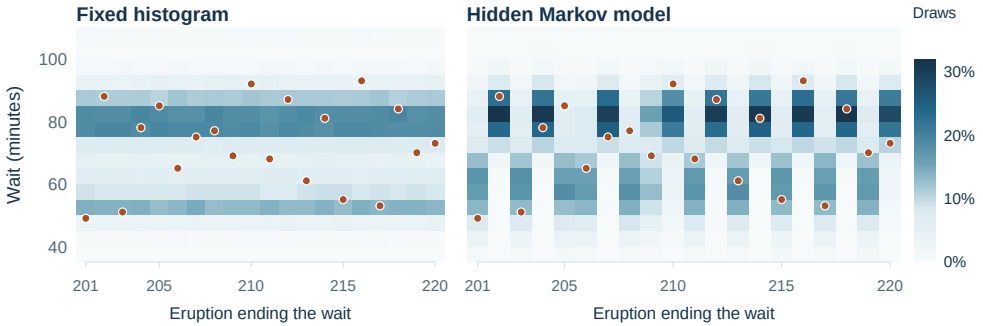
To draw a wait, choose a five-minute bin with probability equal to its training frequency. Draw uniformly within that bin.

The bin probabilities stay fixed after every eruption, regardless of the length of the preceding wait.

Real data: `MASS::geyser`, waits 1–200.

Which waits does each forecast anticipate?

Each column shows 4,000 draws for the next wait, grouped into five-minute bins. The dot marks the wait that followed.



The HMM shifts probability between short and long waits as each eruption occurs.

Real data: `MASS::geyser`, first 20 held-out waits. Simulated forecast draws use only earlier waits, seed 54215.

Did the order of the waits improve the forecasts?

Forecast waits 201–299 in order. For comparison, also use the two fitted normal densities with their stationary state proportions fixed at every eruption.

If a method assigned probability p to the five-minute bin containing the observed wait, its loss is $-\log p$. Smaller is better.

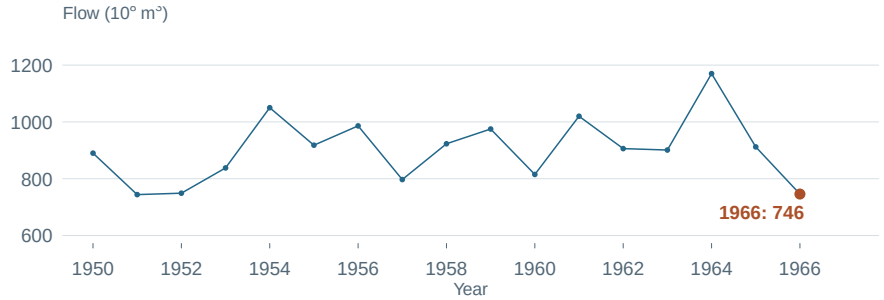
Forecast	Mean bin loss	RMSE of predicted mean (min)
Fixed histogram	2.31	14.2
Fixed Gaussian mixture	2.34	14.2
Hidden Markov model	2.13	11.2

Updating the state probabilities reduced the mean forecast's RMSE by about three minutes.

Real data: `MASS::geyser`, waits 201–299. Scores use exact bin probabilities and forecast means.

How much of a low Nile flow carries into the next year?

At Aswan we observe the volume of water that passes each year, in units of 10^8 cubic meters. We model its underlying mean as a continuous hidden state.



The volume fell from 912 in 1965 to 746 in 1966. How much should we lower our forecast for 1967?

Real data: `datasets::Nile`, annual volumes at Aswan, 1950–1966.

Two sources of variation in annual flow

Let L_t be the hidden mean flow in year t and X_t the recorded volume. In a *local-level model*, an unusually low volume can reflect a fall in the underlying mean or a fluctuation confined to that year.

$$L_t = L_{t-1} + Z_t, \quad Z_t \sim N(0, q),$$

$$X_t = L_t + \varepsilon_t, \qquad \varepsilon_t \sim N(0, r).$$

A change Z_t in the mean carries forward into later years. The departure ε_t affects the recorded volume in year t alone.

How large is a typical change in the mean compared with a single year's fluctuation?

A Gaussian dynamic linear model for a continuous state

For the river the state is $S_t = L_t$. A Gaussian dynamic linear model allows a vector of states $S_t \in \mathbb{R}^m$ and a vector of observations $X_t \in \mathbb{R}^d$. On a window beginning at $t = 1$,

$$S_1 \sim N_m(m_1, P_1),$$

$$S_t = G_t S_{t-1} + Z_t, \quad Z_t \sim N_m(0, Q_t) \quad (t \geq 2),$$

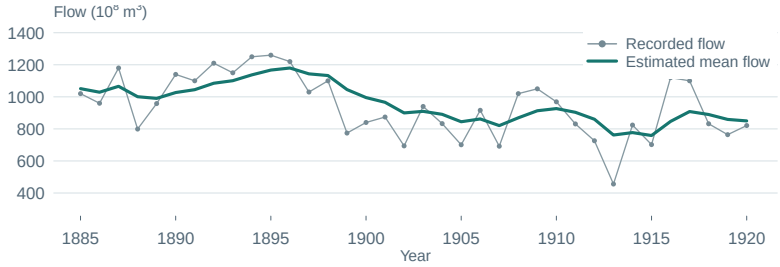
$$X_t = F_t S_t + \varepsilon_t, \qquad \varepsilon_t \sim N_d(0, H_t) \quad (t \geq 1).$$

$G_t s$ is the mean of S_t given $S_{t-1} = s$, and $F_t s$ is the mean of X_t given $S_t = s$. For the Nile, $G_t = F_t = 1$ and the two disturbance variances are $Q_t = q$ and $H_t = r$.

The matrices are deterministic, H_t is positive definite, and the initial state and all disturbances are mutually independent.

The fitted mean moves less than annual flow

For each (q, r), the Kalman filter updates the Gaussian law of the level after each flow. We fit the two variances by maximizing the likelihood from the one-year-ahead Gaussian forecast densities for the 1871–1965 record.



Fitted standard deviations: 35 for a change in mean flow, 124 for a year's departure from its mean.

Real data: `datasets::Nile`; local-level fit with `stats::StructTS`.

How much of a low reading should change the estimated level?

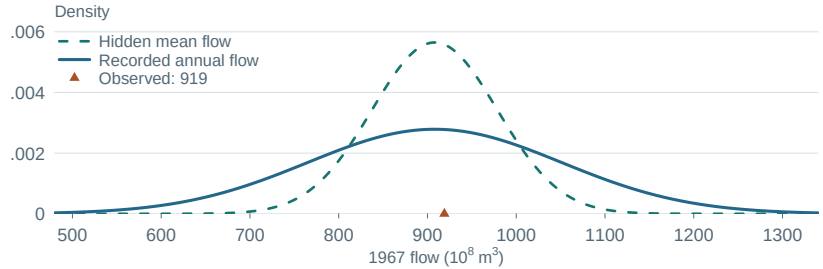
Before observing flow x_t, let m and V be the predicted mean and variance of the hidden level L_t. A single year's flow can deviate from that level with variance r.

$$\mathbb{E}[L_t \mid x_{1:t}] = m + \underbrace{\frac{V}{V+r}}_{\text{Kalman gain}} (x_t - m).$$

When r is large relative to V, the level estimate moves only a little toward the reading. The 1966 flow was 746 against a forecast of 960. The fitted gain carries about a quarter of that shortfall into the estimated level.

What do we forecast for the recorded 1967 flow?

Our updated estimate of the 1966 level is 908. The predicted change in level is zero, so we carry 908 forward to 1967. The annual deviation also has mean zero, giving the same forecast for the recorded volume.



The recorded flow has a wider forecast distribution because it includes the annual fluctuation. Its central 90% interval is $[673, 1143]$. The recorded volume was 919.

Real data: `datasets::Nile`. Forecasts use flows through 1966 and variances fitted through 1965.