Dealing with a changing world in Stan

or

How would an historian estimate time-series models?

Jim Savage
Data Science Lead
Lendable
@khakieconomist

Outline

- What are time-series modellers trying to do?
- Dealing with changes in the underlying DGP pre-Stan
- Dealing with changes in the underlying DGP post-Stan
- Introducing Analogy Weighting

What questions are time-series modellers asking?

- Is the future forecastable? If so, what should I expect? (**forecasting**)
- What elements of time series should I ignore? (signal processing)
- How uncertain should I be? (**Volatility modelling**)
    - Is it even sensible to be talking prediction intervals? (**Talebianism**)
- What happens when I exogenously alter some policy variable when historical changes in the policy variable have not been exogenous? (**time series econometrics**)

Basel eyes set periods for banks' risk models

Rule could set limit on time used for value at risk calculations

Morgan Stanley provided a real-time example when it said last October that it had switched from a four year-weighted model to one that was geared to the past 12 months. The shift lowered the bank's average VAR from $82m to $63m.

From Rob Hyndman: “Prediction intervals too narrow”

phenomenon and arises because they do not account for all sources of uncertainty. In my 2002 IJF paper, we measured the size of the problem by computing the actual coverage percentage of the prediction intervals on hold-out samples. We found that for ETS models, nominal 95% intervals may only provide coverage between 71% and 87%. The difference is due to missing sources of uncertainty.

There are at least four sources of uncertainty in forecasting using time series models:

1. The random error term;
2. The parameter estimates;
3. The choice of model for the historical data;
4. The continuation of the historical data generating process into the future.

Problem #1

- Estimating a time-series model implies that you think that history contains information about the parameters of your model.
  - You think that history must be **relevant**
  - Presumably, not all histories are **equally relevant**
  - **How do you select which histories to use in your estimation?**

Problem #2

- We estimate our models on a given history.
- Our estimates suffer from an inductive problem: the underlying DGP may not have revealed the scale of all probable outcomes. (Talebianism)
- We want to know when to ignore our models.
- We want to know when the world enters an unprecedented state, as soon as possible.

Dealing with problem 1

Problem: The world I'm modelling is changing

Unprincipled approaches

(rolling window, arbitrary choice of sample period)

Semi-principled approaches

(Weighting, exponential smoothing, change points)

Principled approaches

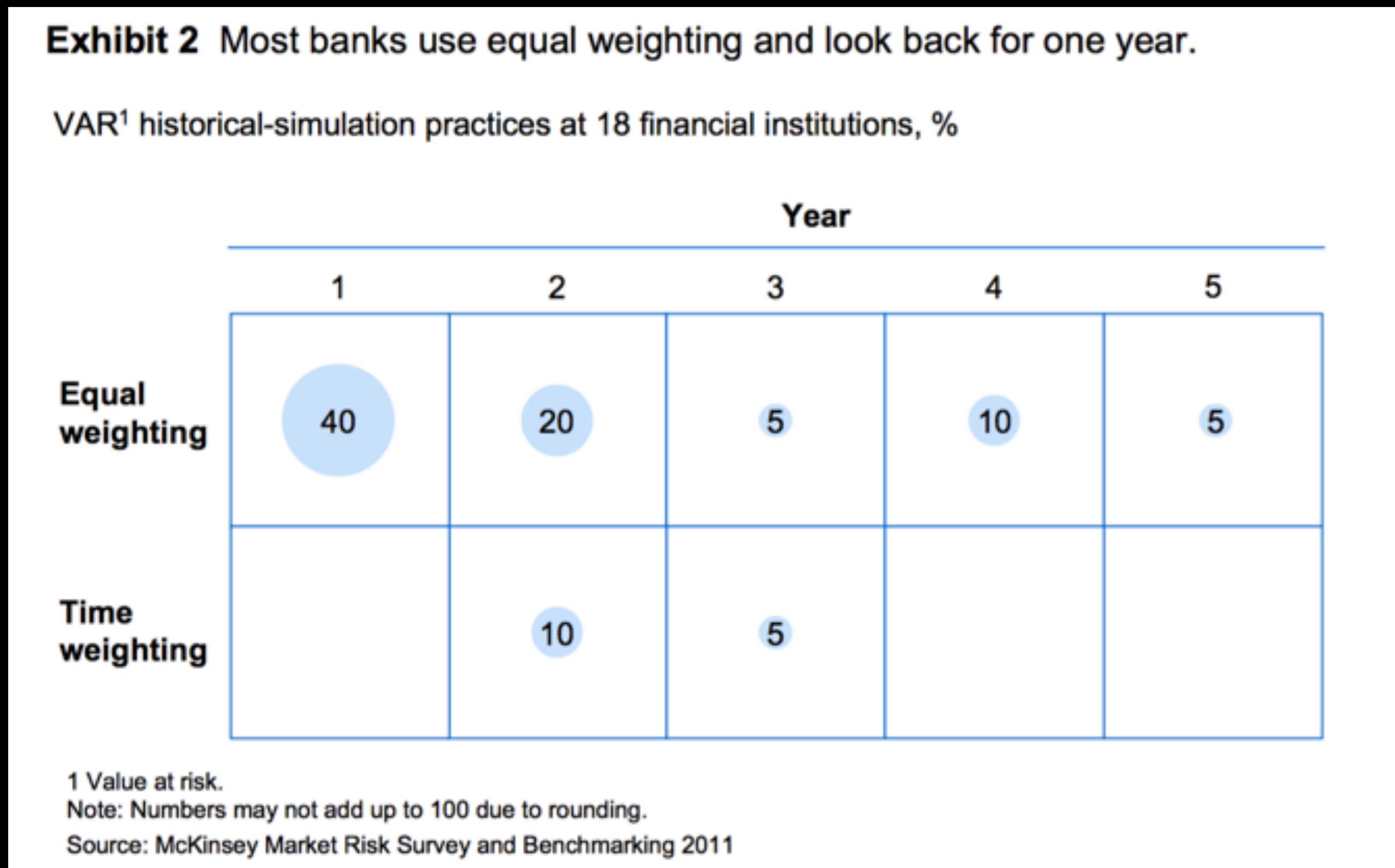
(Model state of world explicitly, either in regime-change model or continuous state)

Dealing with problem 2

When should I ignore my model?

?

Unprincipled approaches



Unprincipled approaches

Unprincipled approaches

Unprincipled approaches

Unprincipled approaches

Unprincipled approaches

Upsides

- Extremely easy to deploy

In R's caret (*train()* function), easy to cross-validate:

```
trControl=trainControl(method='timeslice',
            initialWindow=12, fixedWindow=TRUE,
            horizon=12)
```

- Can result in good performance

Downsides

- No generative model (no idea whether it's estimating anything meaningful)
    - If history is not possible/ probable in your model, your model is definitely wrong.
- Choice of window length/start date?

A principled approach

Regime-shifting models

- Have a discrete number of unobserved states
- Have a model for state membership/transition between states
- Have a distinct set of model parameters for each state

Another principled approach

State space model

- Assumes parameters vary over time and are unobserved.

A very simple example for a matrix of time-varying parameters β and data vector *y*

$$y_t = \beta_t y_{t-1} + \epsilon_t$$
$$\text{vec}(\beta_t) = \text{vec}(\beta_{t-1}) + \eta_t$$
$$\eta \sim \mathcal{MVN}(0, \Sigma_\eta)$$
$$\epsilon \sim \mathcal{MVN}(0, \Sigma_\epsilon)$$

Advantages of regime-shifting and state-space models

- Model parameters can vary over time
  - We could include outside variables in our model for time-variation of model parameters
- Models are *generative*
  - Generative models are a joint probability model of the data and parameters
  - This means we can simulate future values of parameters and outcomes, getting better (wider) prediction intervals.

Disadvantages of these time-varying parameter approaches

- Before Stan, difficult to implement
  - Needed to write up your own filter
  - Canned functions tend to be about as hard to use as writing up your own particle filter
- Even with Stan, need to be careful (especially interpreting historical parameter estimates)
- Often they look like generative models but require hacks to actually behave like the data. (explosive states are very possible)

Estimating state space models before Stan

Kalman Filter

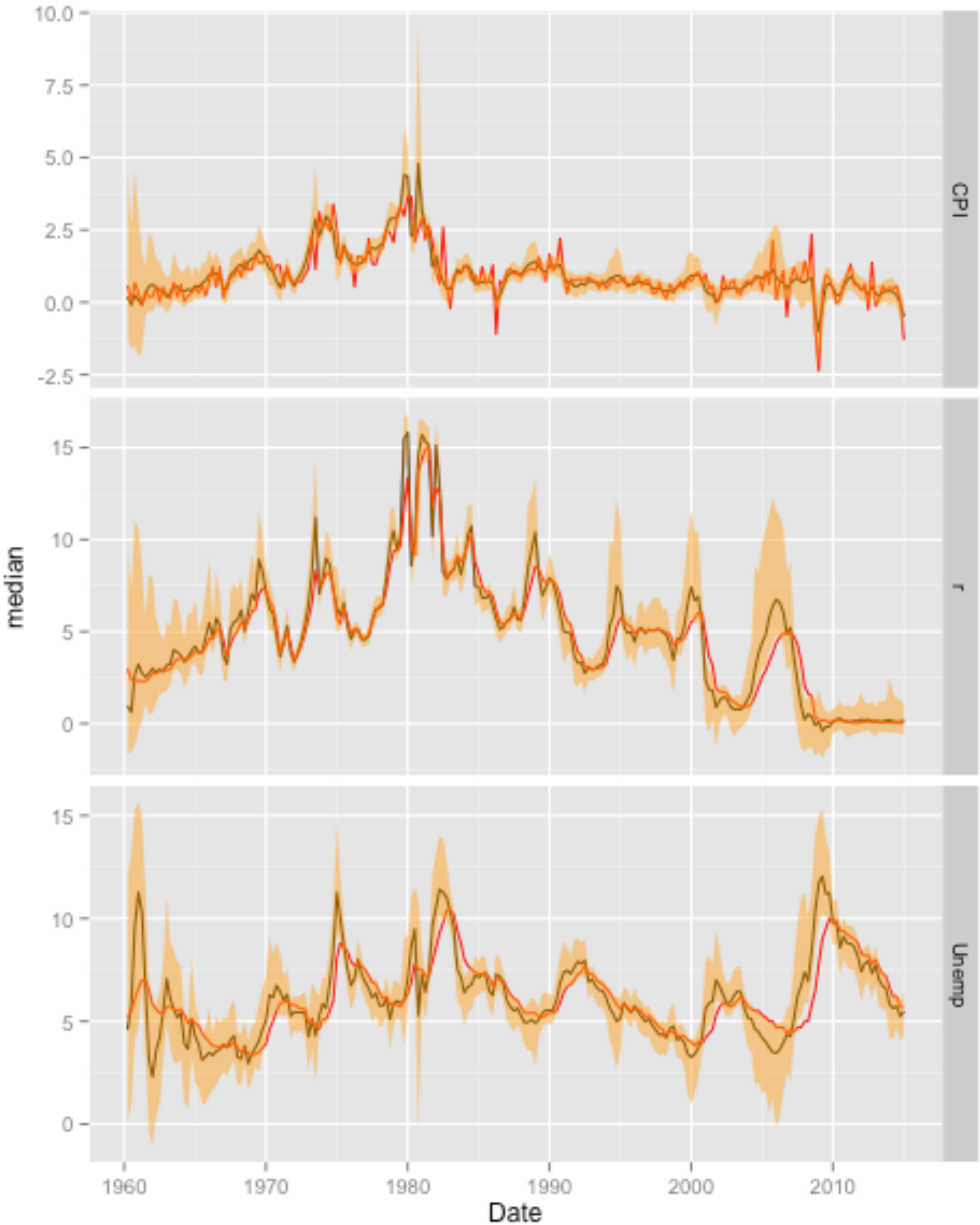
- What do I want? **$p(\beta|y)$** — a time series of estimates of β
- Start with a multi-normal prior for initial state $p(\beta(0), \Sigma(0))$
- Observe data, take multi-normal likelihood $p(y(1)|\beta(0), \Sigma(0))$
- Update posterior $p(\beta(1), \Sigma(1)|y(1))$ estimate of β, Σ using Gaussian product trick (product of two multivariate Gaussians is proportional to a multivariate Gaussian)
- Generate forecast of $\beta(2|1)$ from state model, take forecast variance $\Sigma(2|1)$
- Use these forecasts as priors for the next time step

Kalman filter

- You could do it in Stan
- I find it easier to write it out in R
- Impact of priors dissipates with long time series
- Only a for loop with one matrix inversion per loop
- **For estimate at t, only uses data available at t**

State space model in Stan

- *Can* estimate Kalman filter
- Easier to write out state space model directly and generate estimates of the posterior using HMC
- Rather than estimate the moments of the Gaussian posterior, we generate draws from it and infer the moment.
- State model can take whatever multivariate distribution you want (not relying on Gaussian products anymore)
- Downside: generates **smoothed** estimates. Estimate of state at t uses data from after t. States can often appear to precede outcomes (see next slide; red is actual)



10.0
7.5
5.0
2.5
0.0
-2.5
CPI
15
10
5
0
r
15
10
5
0
Unemp
median
1960
1970
1980
1990
2000
2010
Date

Implementing it in Stan

```
data {
  int T; // number of observations
  int K; // number of variables
  vector[K] Y[T]; // data vector (matrix) T*K
}
parameters {
  matrix[K,K] beta[T];
}
model {
  matrix[K*K,K*K] Sigma_eta;
  matrix[K,K] Sigma_epsilon;
  // fill in your user-provided covariances
  Sigma_eta <- diag_matrix(rep_vector(0.05, K*K));
  Sigma_epsilon <- diag_matrix(rep_vector(0.1, K));
  // give prior to initial observation
  to_vector(beta[1]) ~ normal(0, 1);

  // State and measurement models
  for(t in 2:T){
    Y[t] ~ multi_normal(beta[t]*Y[t-1], Sigma_epsilon);

    to_vector(beta[t]) ~ multi_normal(to_vector(beta[t-1]), Sigma_eta);
  }
}
```

Implementing it in Stan

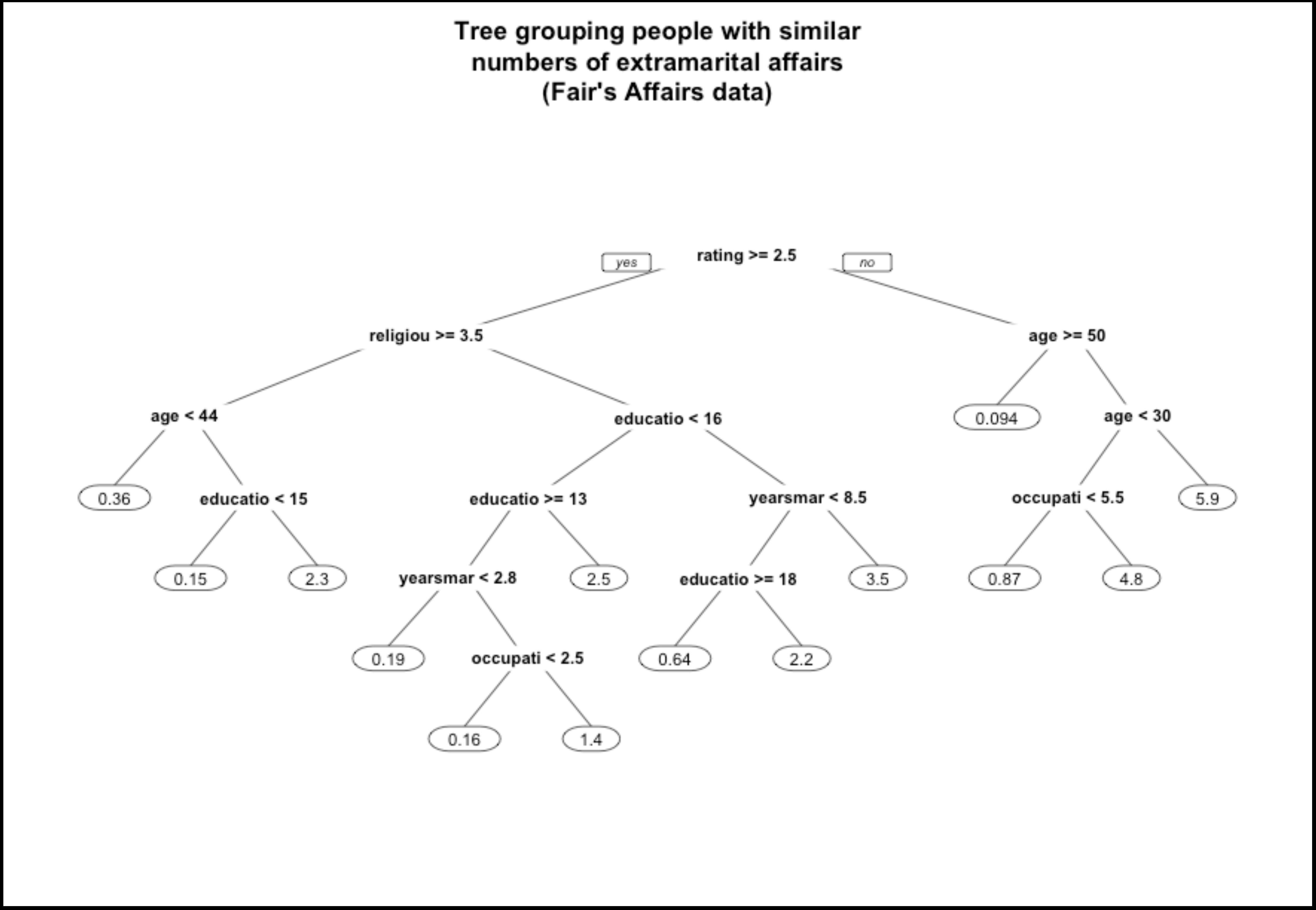
- Extremely easy way of building very rich models
- Beware—the identification of these models is highly dependent on the structure of the model.
  - You could estimate the covariances of the innovation and measurement equations, but this would require strong priors.

Introduction to analogy weighting

- Motivation 1: want parameters most relevant to today.

- Motivation 2: want to know when model is least likely to do a good job.

CART


Tree grouping people with similar numbers of extramarital affairs (Fair's Affairs data)
rating >= 2.5
yes
no
religiou >= 3.5
age >= 50
age < 44
educatio < 16
0.094
age < 30
0.36
educatio < 15
educatio >= 13
yearsmar < 8.5
occupati < 5.5
5.9
0.15
2.3
yearsmar < 2.8
2.5
educatio >= 18
3.5
0.87
4.8
0.19
occupati < 2.5
0.64
2.2
0.16
1.4

The Random Forest

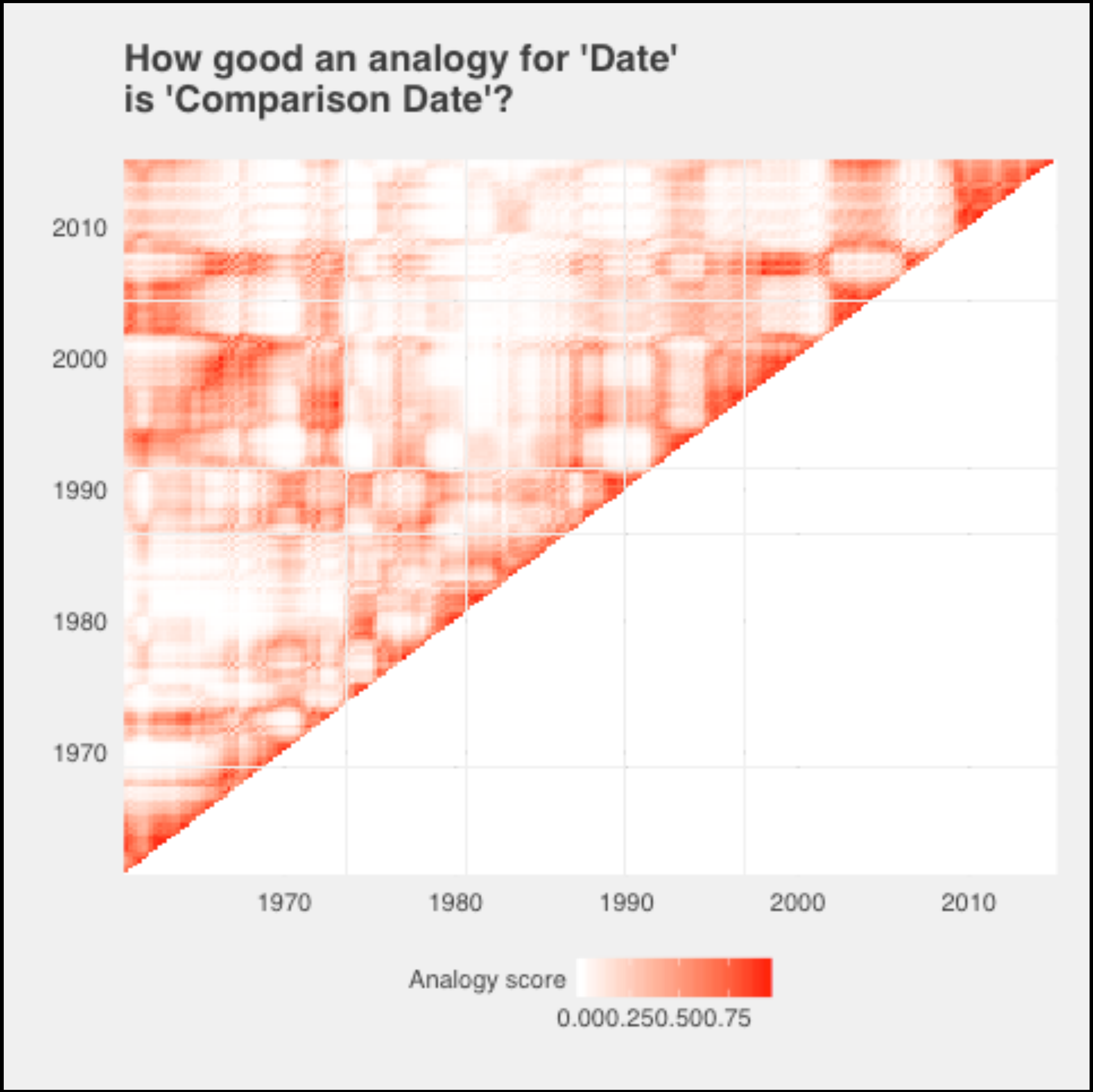
- Essentially a collection of CART models
  - Each estimated on a random subset of the data
  - In each node, a sample of possible Xs drawn to be considered for a split
- Each tree fairly different.

Random forest proximity

- When two people end up in the same terminal node of a tree, they are said to be *proximate*
- The *proximity score* (i, j) is the proportion of terminal nodes shared by individuals *i* and *j*.
  - We calculate it on held-out observations
  - It is a measure of similarity between two individuals in terms of their Xs
  - **But only the similarity in terms of the Xs that matter to y**
  - **A metric-free, scale invariant supervised similarity score**

Analogy weighting: the idea

- Train a random forest on the dependent variable of interest with potentially many Xs
- Take the proximity matrix from the random forest
- Use the relevant row from this matrix to weight the observations in your parametric model
- This is akin to training your model on the **relevant history**



How good an analogy for 'Date'
is 'Comparison Date'?
2010
2000
1990
1980
1970
1970
1980
1990
2000
2010
Analogy score
0.000.250.500.75

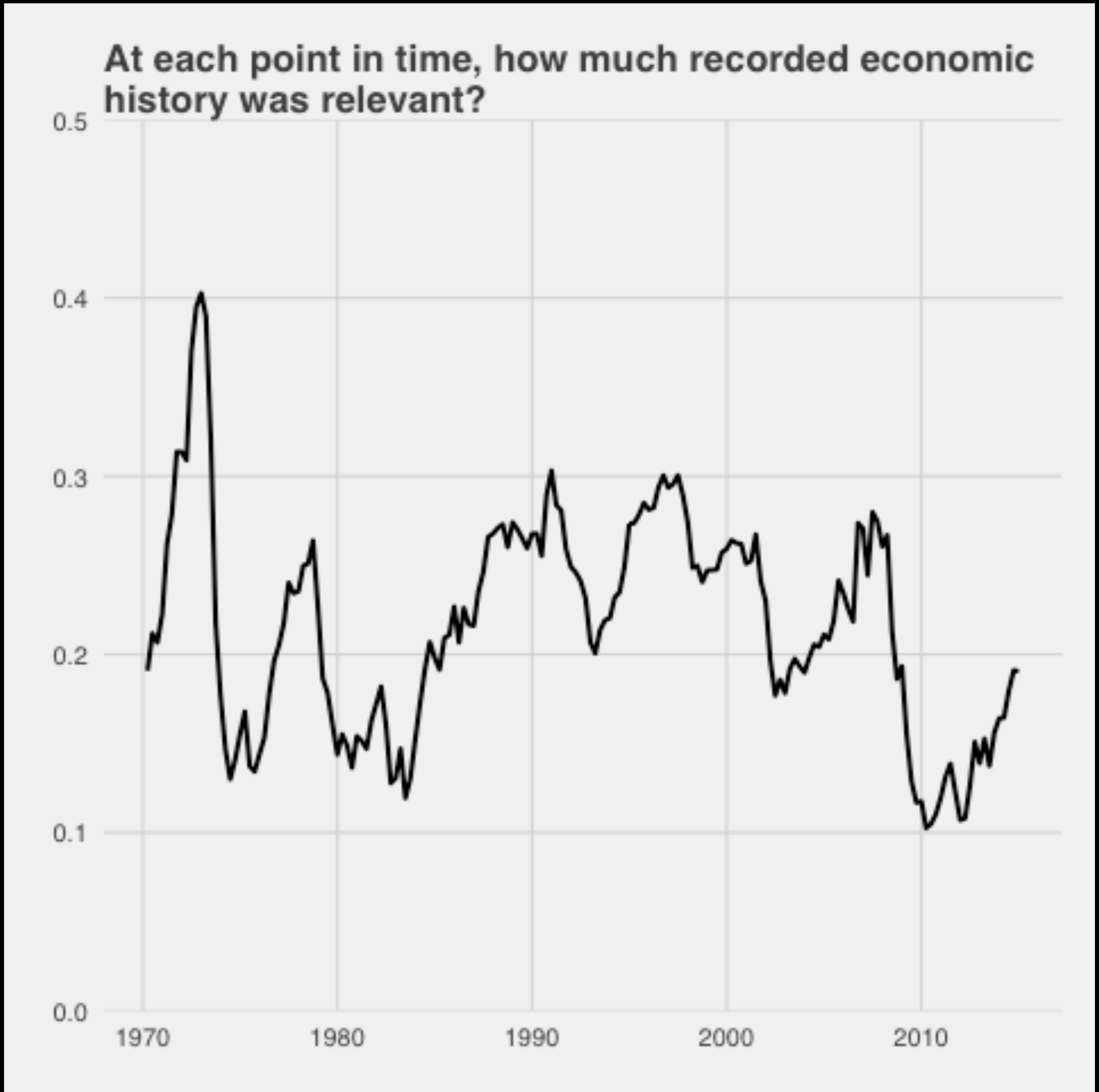
Implementing

- For very simple models, canned functions normally take a weights argument.
- For complex models, weights are not normally included.
  - Use Stan
  - Direct call to increment_log_prob rather than sampling notation

```
to_vector(theta_22) ~ normal(linear_est_2, 0.5);

for(t in 3:T){
  increment_log_prob(weights2[t]*multi_normal_log(Y[t], intercept2 + theta_12*Y[t-1] + theta_22*Y[t-2], Sig_VAR2));
  increment_log_prob(weights[t]*multi_normal_log(Y[t], intercept + theta_1*Y[t-1] + theta_2*Y[t-2], Sig_VAR));
}
```

And when history is not relevant?



Covariance in scale-correlation form

$$\Sigma = \mathrm{diag}(\sigma)\Omega\mathrm{diag}(\sigma)$$

- Here, sigma is a vector of standard deviations, and Omega is a correlation matrix
- We can give sigma a non-negative prior (say, half Cauchy), and Omega an LKJ prior
- LKJ is a one-parameter distribution of correlation matrices.
  - Low values of the parameter give (approaching 1) give uniform prior over correlations.
  - High values (approaching infinity) give an identity matrix.

Application: volatility modelling during financial crisis

- Most volatility models work like so:

returns vector(t) ~ multivariate distribution(expected return(t), covariance(t))

- Expected returns model is just a forecasting model
- Covariance needs to be explicitly modelled
  - Multivariate GARCH common.
    - CCC Garch allows time varying shock magnitudes
    - DCC allows time varying correlations that update with correlated shocks

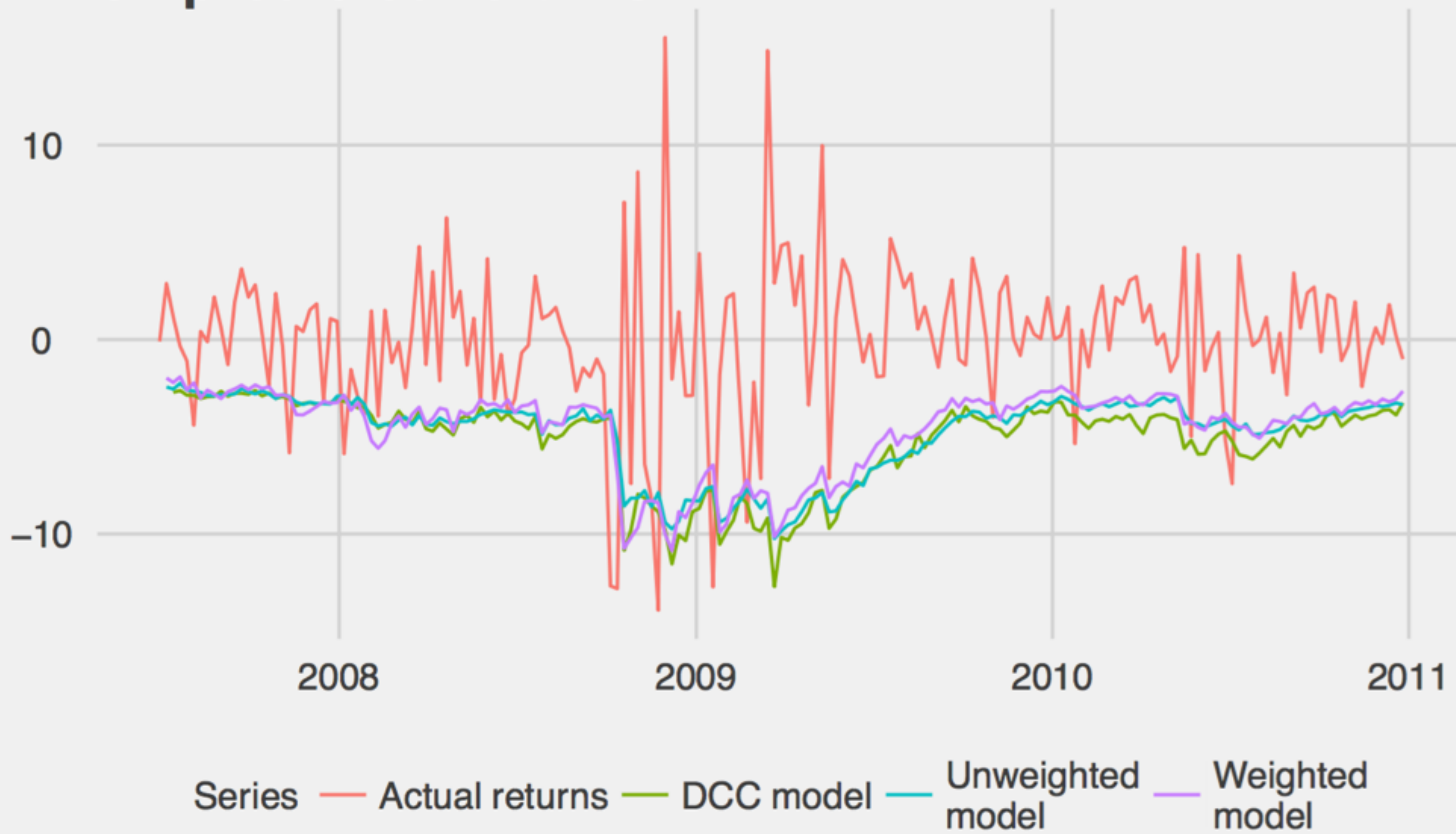
LKJ as a “danger prior” in volatility models

- Idea: when we have relevant histories, we learn correlation structure from the data.

- When we have no relevant history, our likelihood does not impact the posterior and we revert to the prior.

    - Using an LKJ prior with low degrees of freedom gives us highly correlated returns in unprecedented states.

Actual returns and lower 95% bounds on predicted returns



Loss functions evaluated for week-ahead forecasts, July 2007–June 2013

Loss function	Unweighted	Weighted	DCC
Quadratic loss	34.31	**32.52**	38.54
Mean absolute value	18.39	**17.20**	21.56
Heteroskedasticity-adjusted absolute value	0.85	0.86	**0.83**
Heteroskedasticity-adjusted quadratic loss	1.34	1.32	**1.09**
Log loss	9.93	**9.50**	10.79

Questions?