Mixing Dynamic Linear Models

Kevin R. Keane

krkeane@buffalo.edu

Computer Science & Engineering

University at Buffalo SUNY

Modeling label distributions on a lattice

- prior distribution of labels

$$p(L)$$

- likelihood of observed data given a label

$$p(D|L)$$

- posterior distribution of labels given data

$$p(L|D) \propto p(D|L)p(L)$$

- obtain other distributions of interest

$$p(X|D) \propto \int p(X|L)p(L|D)dL$$

A slight twist

Usual setup

- model distribution of *labels* at lattice sites

This project

- model distribution of *linear models* at lattice sites

Mixing Models

Why bother modeling distributions over *models*?

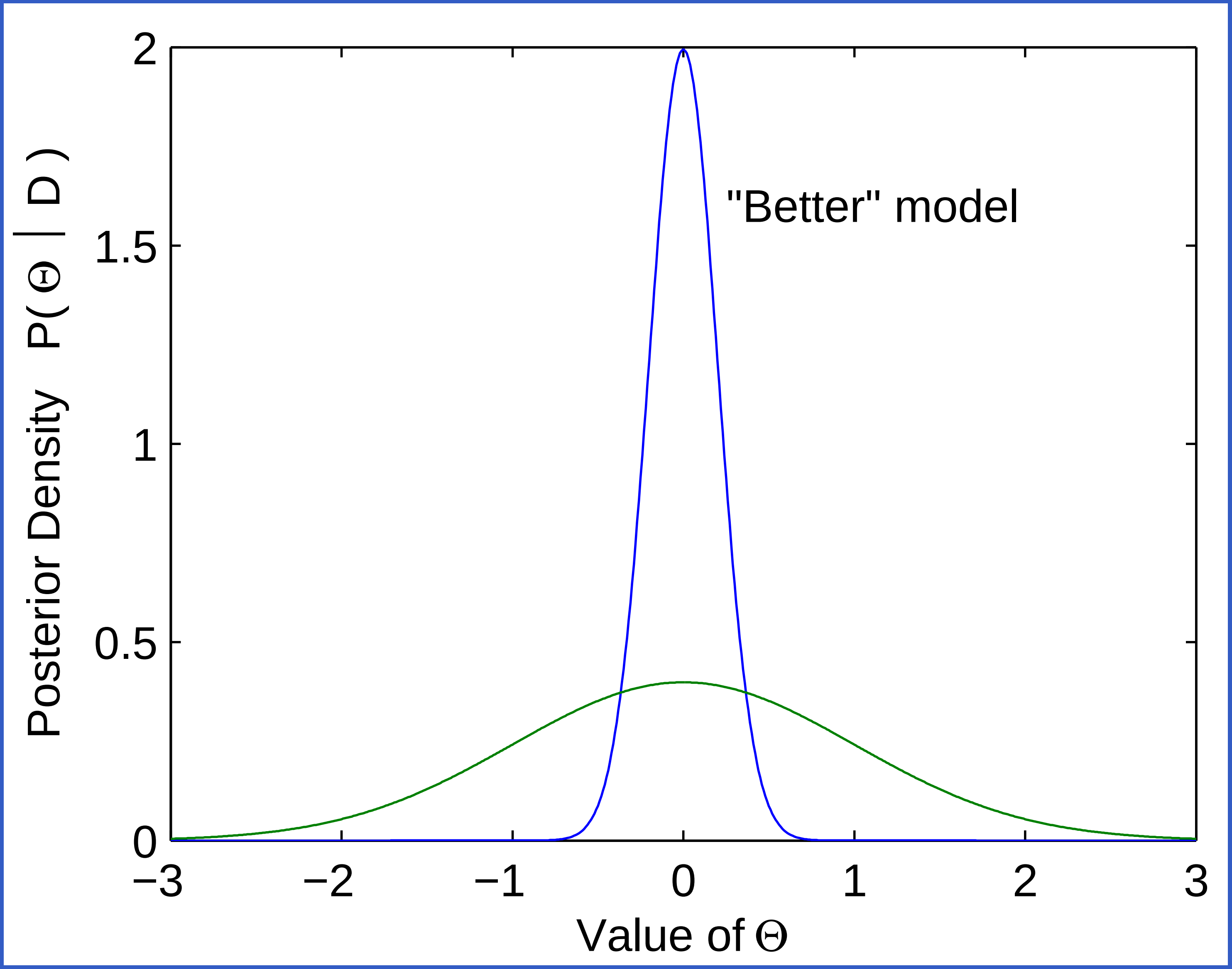
- Avoid the need to carefully specify model parameters in advance
  - use simple components
  - grid approximation spanning likely parameter values
  - let the data figure it out . . .
- To obtain *better* models . . .

What's a *better* model?

A model with *higher posterior density* for quantities of interest. For example:

- parameters in a stock returns model
- parameters describing a plane
- . . .

What’s a *better* model?



Bayesian framework

- prior distribution of models

$$p(M)$$

- likelihood of observed data given a model

$$p(D|M)$$

- posterior distribution of models given data

$$p(M|D) \propto p(D|M)p(M)$$

- obtain other distributions of interest

$$p(X|D) \propto \int p(X|M)p(M|D)dM$$

Drilling down further . . .

Linear regression models

$$Y = F'\theta + \epsilon$$

- Y_i is the *observed response*
- F_i is the *regression vector* / independent variables
- θ is the *regression parameter vector*
- ϵ_i is the *error of fit*

FYI, ordinary least squares estimate of θ is

$$\hat{\theta} = (FF')^{-1} FY$$

Examples of linear models

Return of a stock is a function of market, industry and stock specific return components

$$r = F'\theta + \epsilon, \quad F = \begin{bmatrix} 1 \\ r_M \\ r_I \end{bmatrix}, \quad \theta = \begin{bmatrix} \alpha \\ \beta_M \\ \beta_I \end{bmatrix}.$$

Points on a plane

$$0 = F'\theta + \epsilon, \quad F = \begin{bmatrix} x \\ y \\ z \\ -1 \end{bmatrix}, \quad \theta = \begin{bmatrix} A \\ B \\ C \\ D \end{bmatrix}.$$

Dynamic Linear Models

Ordinary least squares yields a single estimate $\hat{\theta}$ of the regression parameter vector θ for the entire data set.

- we may not have a *finite* data set, but rather an *infinite* data stream!
- we expect / permit θ to vary (slightly) as we traverse the lattice, $\theta_s \approx \theta_t$. So, our models are *dynamic*.

Dynamic linear models (West & Harrison) are a generalized form of / able to represent:

- Kalman filters (Kalman)
- flexible least squares (Kalaba & Tesfatsion)
- linear dynamical systems (Bishop)
- ...

Specifying a dynamic linear model $\{F_t, G, V, W\}$

- F_t' is a row in the *design matrix*
  - vector of independent variables effecting Y_t
- G is the *evolution matrix*
  - captures deterministic changes to θ
  - $\theta_t \approx G\theta_{t-1}$
- V is the *observational variance*
  - a.k.a. Var(ϵ) in ordinary least squares
- W is the *evolution variance matrix*
  - captures random changes to θ
  - $\theta_t = G\theta_{t-1} + w_t, \quad w_t \sim N(0, W)$
- G and W make the linear model *dynamic*

Specifying a dynamic linear model $\{F_t, G, V, W\}$

- The *observation equation* is

$$Y_t = F_t'\theta_t + \nu_t, \quad \nu_t \sim N(0, V)$$

- The *evolution equation* is

$$\theta_t = G\theta_{t-1} + \omega_t, \quad \omega_t \sim N(0, W)$$

- The initial information is summarized

$$(\theta_0 | D_0) \sim N\left(m_0, C_0\right)$$

- Information at time t

$$D_t = \{Y_t, D_{t-1}\}$$

Experiment #1

- Specify 3 DLMs

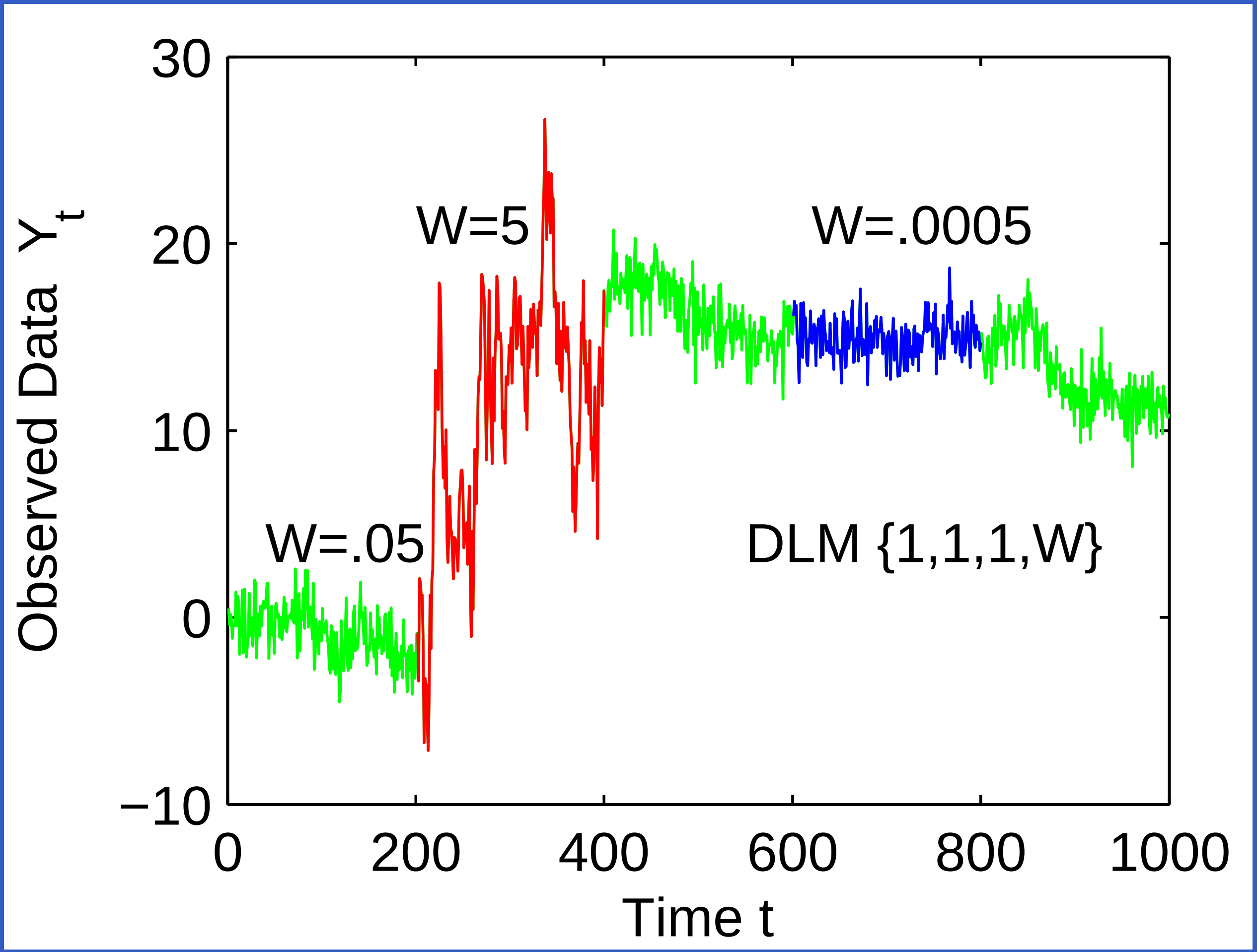
$$\{F = 1, G = 1, W = 5.0000, V = 1\}$$

$$\{F = 1, G = 1, W = 0.0500, V = 1\}$$

$$\{F = 1, G = 1, W = 0.0005, V = 1\}$$

- Generate 1000 observations Y_M
  - *Select a model, generate 200 observations*
  - *Select a model, generate 200 observations*
  - $\ldots$

Generated observations Y_t from 3 DLMs



local consistency is a function of W_m . . .

Inference with a dynamic linear model $\{F_t, G, V, W\}$

- Posterior at $t-1$, $\quad (\theta_{t-1}|D_{t-1}) \sim N\left(m_{t-1}, C_{t-1}\right)$
- Prior at t, $\quad (\theta_t|D_{t-1}) \sim N\left(a_t, R_t\right)$

$$a_t = Gm_{t-1}, \quad R_t = GC_{t-1}G' + W$$

Inference with a dynamic linear model $\{F_t, G, V, W\}$

- Posterior at $t-1$, $(\theta_{t-1}|D_{t-1}) \sim N(m_{t-1}, C_{t-1})$
- Prior at t, $(\theta_t|D_{t-1}) \sim N(a_t, R_t)$

$$a_t = Gm_{t-1}, \quad R_t = GC_{t-1}G' + W$$

- One-step forecast at t, $(Y_t|D_{t-1}) \sim N(f_t, Q_t)$

$$f_t = F_t' a_t \qquad Q_t = F_t' R_t F_t + V$$

Inference with a dynamic linear model $\{F_t, G, V, W\}$

- Posterior at $t-1$, $\quad (\theta_{t-1}|D_{t-1}) \sim N\left(m_{t-1}, C_{t-1}\right)$

- Prior at t, $\quad (\theta_t|D_{t-1}) \sim N\left(a_t, R_t\right)$

$$a_t = G m_{t-1}, \quad R_t = G C_{t-1} G' + W$$

- One-step forecast at t, $\quad (Y_t|D_{t-1}) \sim N\left(f_t, Q_t\right)$

$$f_t = F_t' a_t \qquad\qquad Q_t = F_t' R_t F_t + V$$

- One-step forecast is the *model likelihood*

$$(Y_t|D_{t-1}) = (Y_t, D_{t-1}|D_{t-1}, \{F_t, G, V, W\}) = (D|M)$$

Inference with a dynamic linear model $\{F_t, G, V, W\}$

- Prior at t, $\quad (\theta_t | D_{t-1}) \sim N(a_t, R_t)$

$$a_t = G m_{t-1}, \qquad R_t = G C_{t-1} G' + W$$

- One-step forecast at t, $\quad (Y_t | D_{t-1}) \sim N(f_t, Q_t)$

$$f_t = F_t' a_t \qquad\qquad Q_t = F_t' R_t F_t + V$$

- Posterior at t, $\quad (\theta_t | D_t) \sim N(m_t, C_t)$

$$e_t = Y_t - f_t \qquad\qquad A_t = R_t F_t Q_t^{-1}$$
$$m_t = a_t + A_t e_t \qquad\qquad C_t = R_t - A_t Q_t A_t'$$

Experiment #1 continued / inference about θ_t

- Specify 3 DLMs

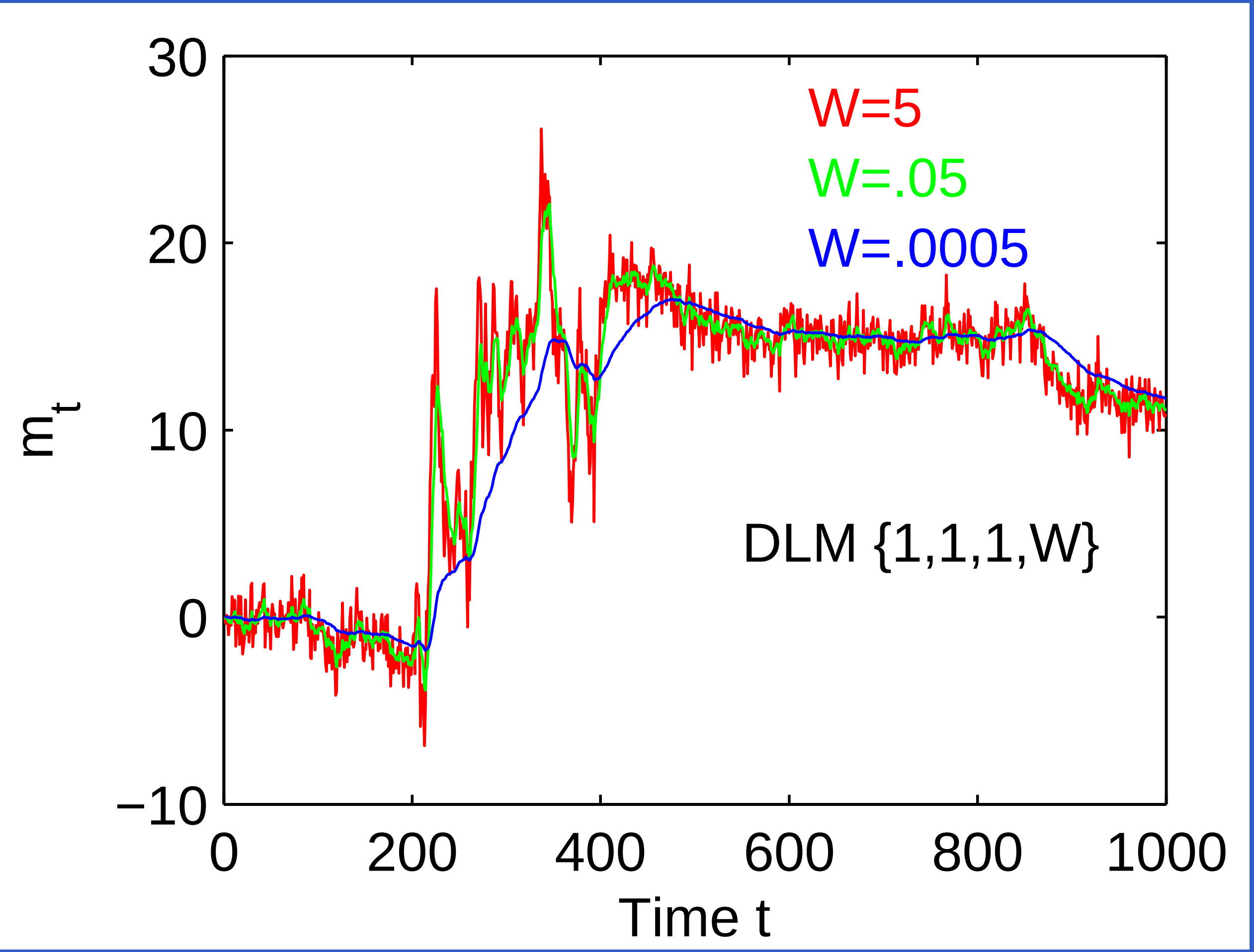
$$\{F = 1, G = 1, W = 5.0000, V = 1\}$$

$$\{F = 1, G = 1, W = 0.0500, V = 1\}$$

$$\{F = 1, G = 1, W = 0.0005, V = 1\}$$

- Generate 1000 observations Y_M
  - *Select a model, generate 200 observations*
  - *Select a model, generate 200 observations*
  - $\ldots$
- Compute posteriors $(\theta_{t-1} | M = m, D_{t-1})$

Posterior mean m_t from 3 DLMs



$$(\theta_t | D_t) \sim N\left(m_t, C_t\right)$$

Experiment #1 continued / inference about models

- Specify 3 DLMs

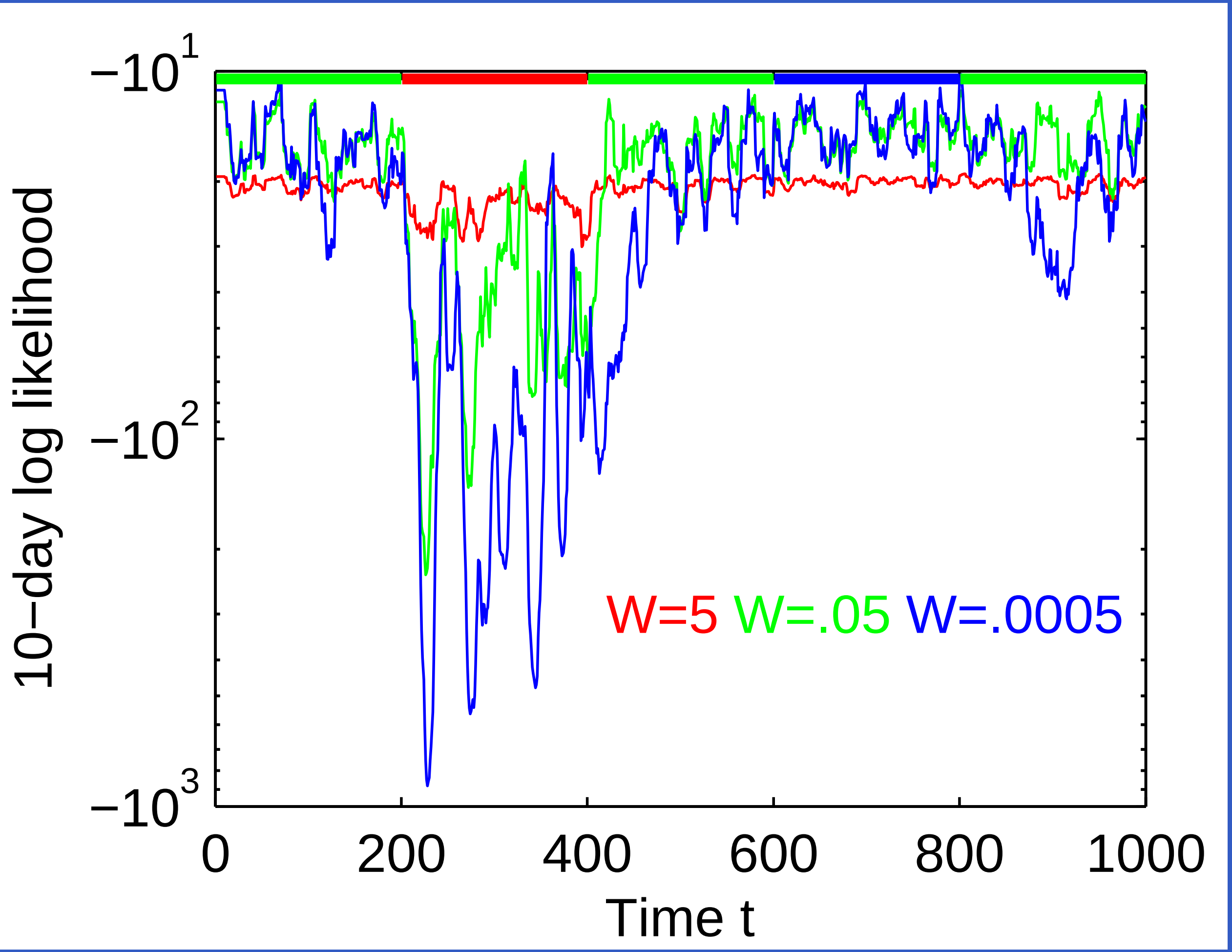
$$\{F = 1, G = 1, W = 5.0000, V = 1\}$$

$$\{F = 1, G = 1, W = 0.0500, V = 1\}$$

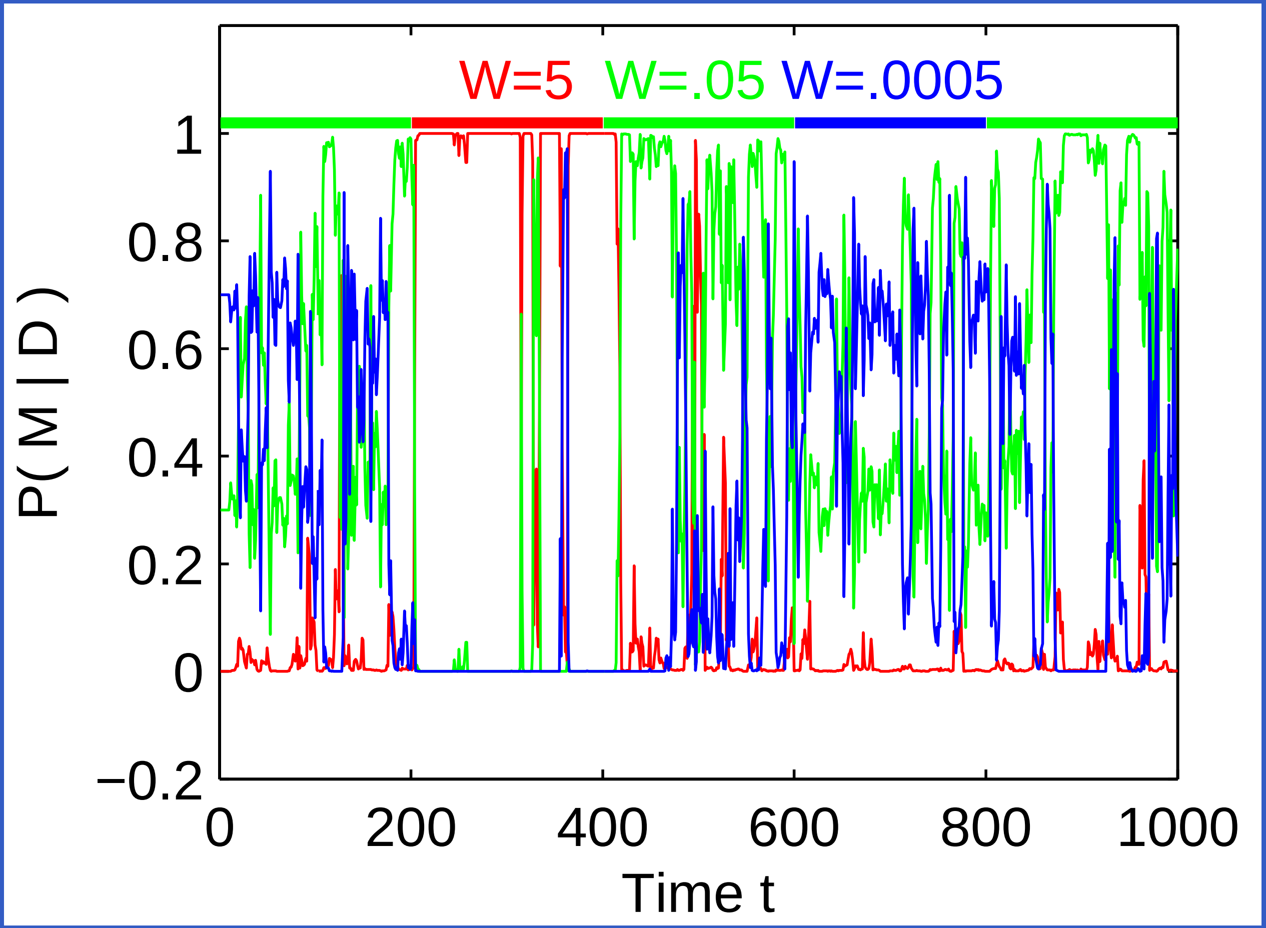
$$\{F = 1, G = 1, W = 0.0005, V = 1\}$$

- Generate 1000 observations Y_M
  - *Select a model, generate 200 observations*
  - *Select a model, generate 200 observations*
  - $\ldots$
- Compute posteriors $(\theta_{t-1} | M = m, D_{t-1})$
- Assume prior $p(M = m) = \frac{1}{3}$
- Compute likelihoods $(D_t | M = m)$
- Compute posterior model probabilities $p(m | D_t)$

Trailing interval model log likelihoods



Posterior model probabilities and ground truth



Modeling interpretation / tricks

- Large W permits the state vector θ to vary abruptly in a stochastic manner
- Examples of when is this important?
  - *stock price model* — one company aquires another company
  - *planar geometry model* — a building's plane intersects the ground plane
- Easy to obtain parameter estimates from a mixture model. For example,evolution variance W_M is obtained from individual model variances, likelihoods, and posterior probabilities

$$W_M = \int_M W_m p(W_m|m) p(m|D) dm = \int_M W_m p(m|D) dm$$

A statistical arbitrage application

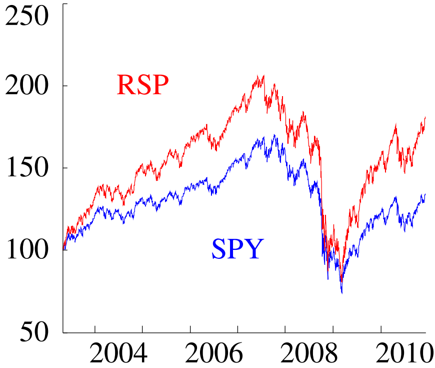
- Montana, Triantafyllopoulos, and Tsagaris. Flexible least squares for temporal data mining and statistical arbitrage. *Expert Systems with Applications*, 36(2):2819-2830, 2009.
- Parameter $\delta = \frac{W}{W+V}$ controls adaptiveness. $W = 0, \delta = 0$ equivalent to ordinary least squares.
- Published results for 11 constant parameter models
- Model S&P 500 Index return as a function of the largest principal component return (score) of the underlying stocks

$$r_{\text{s\&p 500}} = F'\theta + \epsilon, \quad F = \begin{bmatrix} r_{\text{pca}} \end{bmatrix}, \quad \theta = \begin{bmatrix} \beta_{\text{pca}} \end{bmatrix}.$$

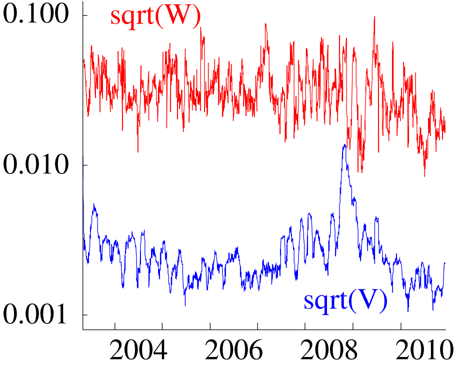
Experiment #2

- Data used is the log price return series for SPY and RSP
- SPY is the capitalization weighted S&P 500 ETF
- RSP is the equal-weighted S&P 500 ETF; our proxy for Montana's β_{pca}

$$r_{\text{spy}} = F'\theta + \epsilon, \quad F = \begin{bmatrix} r_{\text{rsp}} \end{bmatrix}, \quad \theta = \begin{bmatrix} \beta_{\text{rsp}} \end{bmatrix}.$$



250
200
150
100
50
2004
2006
2008
2010
RSP
SPY

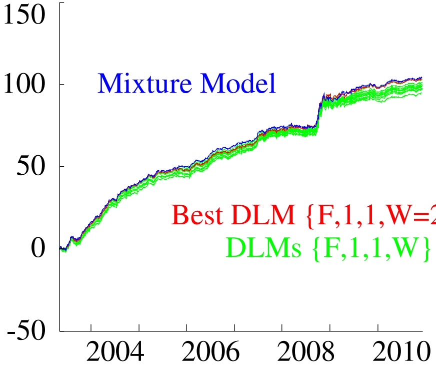

0.100
0.010
0.001
sqrt(W)
sqrt(V)
2004
2006
2008
2010

Experiment #2 continued

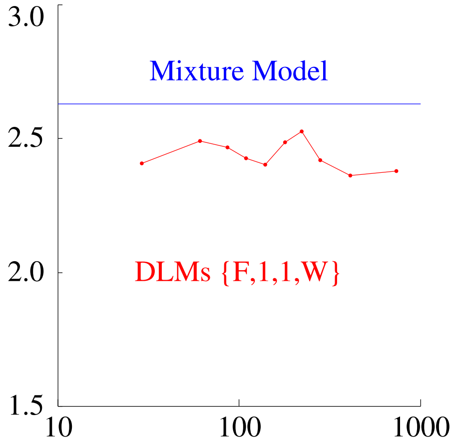
- Data used is the log price return series for SPY and RSP
- SPY is the capitalization weighted S&P 500 ETF
- RSP is the equal-weighted S&P 500 ETF; our proxy for Montana's β_{pca}

$$r_{\text{spy}} = F'\theta + \epsilon, \quad F = \begin{bmatrix} r_{\text{rsp}} \end{bmatrix}, \quad \theta = \begin{bmatrix} \beta_{\text{rsp}} \end{bmatrix}.$$

- Can a mixture model outperform the *ex post* best single DLM?



150
100
50
0
-50
2004
2006
2008
2010
Mixture Model
Best DLM {F,1,1,W=221}
DLMs {F,1,1,W}


3.0
2.5
2.0
1.5
10
100
1000
Mixture Model
DLMs {F,1,1,W}

Future work

- Longer term (after this semester), generalize the one-dimensional DLM framework to permit application to images and video.
- For video, would probably need to consider variation W in different directions,

$$\frac{\delta\theta}{\delta x}, \quad \frac{\delta\theta}{\delta y}, \quad \text{and} \quad \frac{\delta\theta}{\delta t} \quad .$$

(image coordinates x and y, frame t).

- THANK YOU.