EM Algorithm

Jur van den Berg

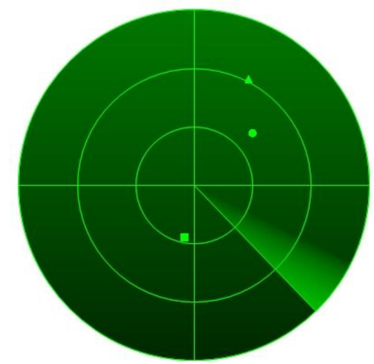
Kalman Filtering vs. Smoothing

- Dynamics and Observation model

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + \mathbf{w}_t, \quad \mathbf{w}_t \sim W_t = N(\mathbf{0}, Q)$$

$$\mathbf{y}_t = C\mathbf{x}_t + \mathbf{v}_t, \quad \mathbf{v}_t \sim V_t = N(\mathbf{0}, R)$$

- Kalman Filter:
  - Compute $\left(X_t \mid Y_0 = \mathbf{y}_0, \ldots, Y_t = \mathbf{y}_t\right)$
  - Real-time, given data so far



- Kalman Smoother:
  - Compute $\left(X_t \mid Y_0 = \mathbf{y}_0, \ldots, Y_T = \mathbf{y}_T\right), \quad 0 \le t \le T$
  - Post-processing, given all data

EM Algorithm

$$\begin{aligned}
\mathbf{x}_{t+1} &= A\mathbf{x}_t + \mathbf{w}_t, \quad \mathbf{w}_t \sim W_t = N(\mathbf{0}, Q) \\
\mathbf{y}_t &= C\mathbf{x}_t + \mathbf{v}_t, \quad \mathbf{v}_t \sim V_t = N(\mathbf{0}, R)
\end{aligned}$$

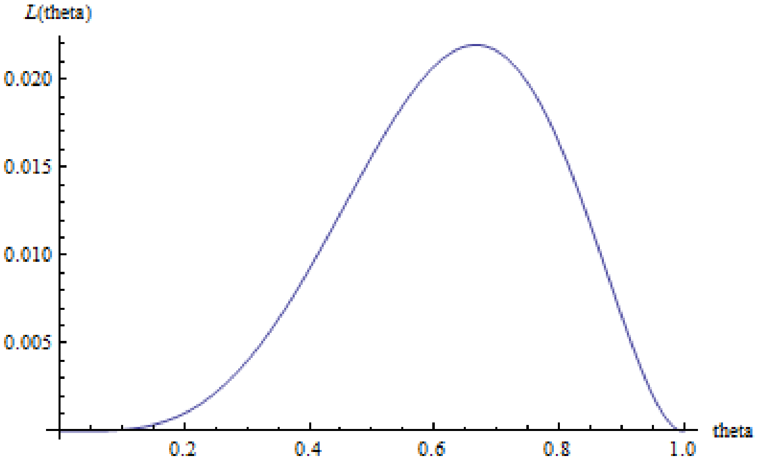
- Kalman smoother:
  - Compute distributions $X_0, \ldots, X_t$ given parameters A, C, Q, R, and data $\mathbf{y}_0, \ldots, \mathbf{y}_t$.
- EM Algorithm:
  - Simultaneously optimize $X_0, \ldots, X_t$ and A, C, Q, R given data $\mathbf{y}_0, \ldots, \mathbf{y}_t$.

Probability vs. Likelihood

- Probability: predict unknown *outcomes* based on known *parameters*:
  - $p(x \mid \theta)$
- Likelihood: estimate unknown *parameters* based on known *outcomes*:
  - $L(\theta \mid x) = p(x \mid \theta)$
- Coin-flip example:
  - θ is probability of "heads" (parameter)
  - x = HHHTTH is outcome

Likelihood for Coin-flip Example

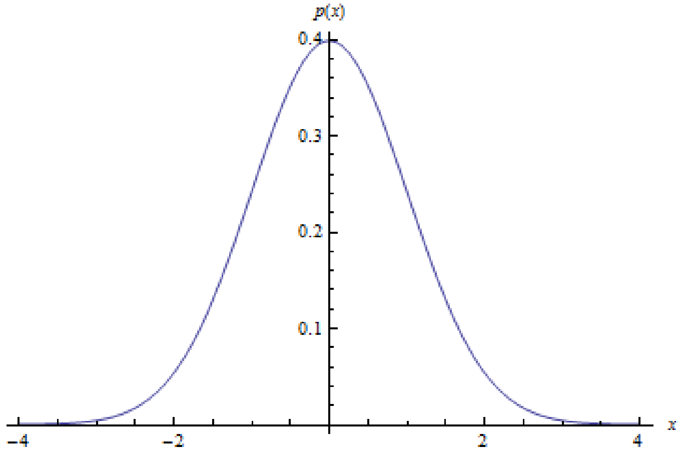
- Probability of outcome given parameter:
  - $p(x = HHHTTH \mid \theta = 0.5) = 0.5^6 = 0.016$
- Likelihood of parameter given outcome:
  - $L(\theta = 0.5 \mid x = HHHTTH) = p(x \mid \theta) = 0.016$

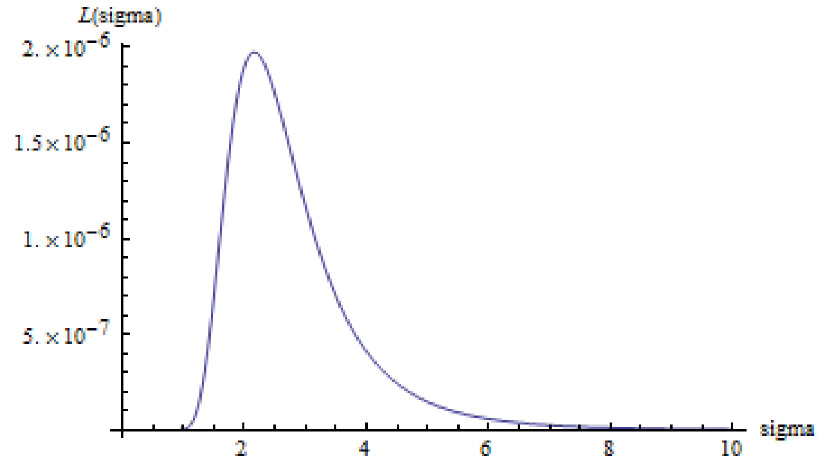


- Likelihood *maximal* when $\theta = 0.6666...$
- Likelihood function **not** a probability density

Likelihood for Cont. Distributions

- Six samples {-3, -2, -1, 1, 2, 3} believed to be drawn from some Gaussian N(0, σ^2)





- Likelihood of σ:

$$L(\sigma \mid \{-3,-2,-1,1,2,3\}) = p(x=-3 \mid \sigma) \cdot p(x=-2 \mid \sigma) \cdots p(x=3 \mid \sigma)$$

- Maximum likelihood:

$$\sigma = \sqrt{\frac{(-3)^2 + (-2)^2 + (-1)^2 + 1^2 + 2^2 + 3^2}{6}} = 2.16$$

Likelihood for Stochastic Model

- Dynamics model

$$\begin{aligned} \mathbf{x}_{t+1} &= A\mathbf{x}_t + \mathbf{w}_t, \quad \mathbf{w}_t \sim W_t = N(\mathbf{0}, Q) \\ \mathbf{y}_t &= C\mathbf{x}_t + \mathbf{v}_t, \quad \mathbf{v}_t \sim V_t = N(\mathbf{0}, R) \end{aligned}$$

- Suppose $\mathbf{x}_t$ and $\mathbf{y}_t$ are given for $0 \le t \le T$, what is likelihood of *A, C, Q* and *R*?
- $L(A, C, Q, R \mid \mathbf{x}, \mathbf{y}) = p(\mathbf{x}, \mathbf{y} \mid A, C, Q, R) = \prod_{t=0}^{T} p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) p(\mathbf{y}_t \mid \mathbf{x}_t)$
- Compute *log-likelihood*:

$$\log p(\mathbf{x}, \mathbf{y} \mid A, C, Q, R)$$

Log-likelihood

$$\log p(\mathbf{x},\mathbf{y} \mid A, C, Q, R) = \log \prod_{t=0}^{T} p(\mathbf{x}_t \mid \mathbf{x}_{t-1})\, p(\mathbf{y}_t \mid \mathbf{x}_t) =$$

$$\sum_{t=0}^{T-1} \log p(\mathbf{x}_{t+1} \mid \mathbf{x}_t) + \sum_{t=0}^{T} \log p(\mathbf{y}_t \mid \mathbf{x}_t) = \ldots$$

- Multivariate normal distribution N($\boldsymbol{\mu}$, Σ) has pdf: $p(\mathbf{x}) = (2\pi)^{-k/2} \left|\Sigma^{-1}\right|^{1/2} \exp(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}))$
- From model: $\mathbf{x}_{t+1} \sim N(A\mathbf{x}_t, Q) \quad \mathbf{y}_t \sim N(C\mathbf{x}_t, R)$

$$= \left(\sum_{t=0}^{T-1} \frac{1}{2} \log\left|Q^{-1}\right| - \frac{1}{2} (\mathbf{x}_{t+1} - A\mathbf{x}_t)^T Q^{-1} (\mathbf{x}_{t+1} - A\mathbf{x}_t) \right) +$$

$$\left(\sum_{t=0}^{T} \frac{1}{2} \log\left|R^{-1}\right| - \frac{1}{2} (\mathbf{y}_t - C\mathbf{x}_t)^T R^{-1} (\mathbf{y}_t - C\mathbf{x}_t) \right) + \text{const}$$

Log-likelihood #2

$$\left(\sum_{t=0}^{T-1}\frac{1}{2}\log\left|Q^{-1}\right|-\frac{1}{2}(\mathbf{x}_{t+1}-A\mathbf{x}_t)^T Q^{-1}(\mathbf{x}_{t+1}-A\mathbf{x}_t)\right)+$$

$$\left(\sum_{t=0}^{T}\frac{1}{2}\log\left|R^{-1}\right|-\frac{1}{2}(\mathbf{y}_t-C\mathbf{x}_t)^T R^{-1}(\mathbf{y}_t-C\mathbf{x}_t)\right)+\text{const}=\ldots$$

- a = Tr(a) if a is scalar
- Bring summation inward

$$=\frac{T}{2}\log\left|Q^{-1}\right|-\frac{1}{2}\left(\sum_{t=0}^{T-1}\mathrm{Tr}((\mathbf{x}_{t+1}-A\mathbf{x}_t)^T Q^{-1}(\mathbf{x}_{t+1}-A\mathbf{x}_t))\right)+$$

$$\frac{T+1}{2}\log\left|R^{-1}\right|-\frac{1}{2}\left(\sum_{t=0}^{T}\mathrm{Tr}((\mathbf{y}_t-C\mathbf{x}_t)^T R^{-1}(\mathbf{y}_t-C\mathbf{x}_t))\right)+\text{const}$$

Log-likelihood #3

$$\frac{T}{2}\log\left|Q^{-1}\right| - \frac{1}{2}\left(\sum_{t=0}^{T-1}\mathrm{Tr}((\mathbf{x}_{t+1} - A\mathbf{x}_t)^T Q^{-1}(\mathbf{x}_{t+1} - A\mathbf{x}_t))\right) +$$

$$\frac{T+1}{2}\log\left|R^{-1}\right| - \frac{1}{2}\left(\sum_{t=0}^{T}\mathrm{Tr}((\mathbf{y}_t - C\mathbf{x}_t)^T R^{-1}(\mathbf{y}_t - C\mathbf{x}_t))\right) + \mathrm{const} = \ldots$$

- Tr(AB) = Tr(BA)
- Tr(A) + Tr(B) = Tr(A+B)

$$= \frac{T}{2}\log\left|Q^{-1}\right| - \frac{1}{2}\mathrm{Tr}\left(Q^{-1}\left(\sum_{t=0}^{T-1}(\mathbf{x}_{t+1} - A\mathbf{x}_t)(\mathbf{x}_{t+1} - A\mathbf{x}_t)^T\right)\right) +$$

$$\frac{T+1}{2}\log\left|R^{-1}\right| - \frac{1}{2}\mathrm{Tr}\left(R^{-1}\left(\sum_{t=0}^{T}(\mathbf{y}_t - C\mathbf{x}_t)(\mathbf{y}_t - C\mathbf{x}_t)^T\right)\right) + \mathrm{const}$$

Log-likelihood #4

$$\frac{T}{2}\log\left|Q^{-1}\right| - \frac{1}{2}\mathrm{Tr}\left(Q^{-1}\left(\sum_{t=0}^{T-1} (\mathbf{x}_{t+1} - A\mathbf{x}_t)(\mathbf{x}_{t+1} - A\mathbf{x}_t)^T \right)\right) +$$

$$\frac{T+1}{2}\log\left|R^{-1}\right| - \frac{1}{2}\mathrm{Tr}\left(R^{-1}\left(\sum_{t=0}^{T} (\mathbf{y}_t - C\mathbf{x}_t)(\mathbf{y}_t - C\mathbf{x}_t)^T \right)\right) + \mathrm{const} = \ldots$$

- Expand

$$l(A, C, Q, R \mid \mathbf{x}, \mathbf{y}) =$$

$$\frac{T}{2}\log\left|Q^{-1}\right| - \frac{1}{2}\mathrm{Tr}\left(Q^{-1}\left(\sum_{t=0}^{T-1} \mathbf{x}_{t+1}\mathbf{x}_{t+1}^T - \mathbf{x}_{t+1}\mathbf{x}_t^T A^T - A\mathbf{x}_t\mathbf{x}_{t+1}^T + A\mathbf{x}_t\mathbf{x}_t^T A^T \right)\right) +$$

$$\frac{T+1}{2}\log\left|R^{-1}\right| - \frac{1}{2}\mathrm{Tr}\left(R^{-1}\left(\sum_{t=0}^{T} \mathbf{y}_t\mathbf{y}_t^T - \mathbf{y}_t\mathbf{x}_t^T C^T - C\mathbf{x}_t\mathbf{y}_{t+1}^T + C\mathbf{x}_t\mathbf{x}_t^T C^T \right)\right) + \mathrm{const}$$

Maximize likelihood

- log is monotone function
  - max log(f(x)) ⇔ max f(x)
- Maximize $l(A, C, Q, R \mid \mathbf{x}, \mathbf{y})$ in turn for A, C, Q and R.
  - Solve $\frac{\partial l(A, C, Q, R \mid x, y)}{\partial A} = 0$ for A
  - Solve $\frac{\partial l(A, C, Q, R \mid x, y)}{\partial C} = 0$ for C
  - Solve $\frac{\partial l(A, C, Q, R \mid x, y)}{\partial Q} = 0$ for Q
  - Solve $\frac{\partial l(A, C, Q, R \mid x, y)}{\partial R} = 0$ for R

Matrix derivatives

- Defined for scalar functions f : $\mathbf{R}^{n*m}$ **->** $\mathbf{R}$

$$\frac{\partial f}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial f}{\partial X_{1,1}} & \cdots & \frac{\partial f}{\partial X_{n,1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial X_{1,m}} & \cdots & \frac{\partial f}{\partial X_{n,m}} \end{bmatrix}.$$

- Key identities

$$\frac{\partial \mathbf{x}^T A \mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^T (A^T + A)$$

$$\frac{\partial B^T A B}{\partial B} = B^T (A^T + A)$$

$$\frac{\partial \mathrm{Tr}(AB)}{\partial A} = \frac{\partial \mathrm{Tr}(BA)}{\partial A} = \frac{\partial \mathrm{Tr}(B^T A^T)}{\partial A} = B^T$$

$$\frac{\partial \log |A|}{\partial A} = A^{-T}$$

Optimizing A

- Derivative

$$\frac{\partial l(A, C, Q, R \mid x, y)}{\partial A} = \frac{1}{2} Q^{-1} \left(\sum_{t=0}^{T-1} 2\mathbf{x}_{t+1}\mathbf{x}_t^T - 2A\mathbf{x}_t\mathbf{x}_t^T \right)$$

- Maximizer

$$A = \left(\sum_{t=0}^{T-1} \mathbf{x}_{t+1}\mathbf{x}_t^T \right) \left(\sum_{t=0}^{T-1} \mathbf{x}_t\mathbf{x}_t^T \right)^{-1}$$

Optimizing C

- Derivative

$$\frac{\partial l(A,C,Q,R \mid x,y)}{\partial C} = \frac{1}{2} R^{-1} \left(\sum_{t=0}^{T} 2\mathbf{y}_t \mathbf{x}_t^T - 2C\mathbf{x}_t \mathbf{x}_t^T \right)$$

- Maximizer

$$C = \left(\sum_{t=0}^{T} \mathbf{y}_t \mathbf{x}_t^T \right) \left(\sum_{t=0}^{T} \mathbf{x}_t \mathbf{x}_t^T \right)^{-1}$$

Optimizing Q

- Derivative with respect to inverse

$$\frac{\partial l(A,C,Q,R \mid x,y)}{\partial Q^{-1}} = \frac{T}{2}Q - \frac{1}{2}\left(\sum_{t=0}^{T-1}\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T - \mathbf{x}_{t+1}\mathbf{x}_t^T A^T - A\mathbf{x}_t\mathbf{x}_{t+1}^T + A\mathbf{x}_t\mathbf{x}_t^T A^T\right)^T$$

- Maximizer

$$Q = \frac{1}{T}\left(\sum_{t=0}^{T-1}\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T - \mathbf{x}_{t+1}\mathbf{x}_t^T A^T - A\mathbf{x}_t\mathbf{x}_{t+1}^T + A\mathbf{x}_t\mathbf{x}_t^T A^T\right)$$

Optimizing *R*

- Derivative with respect to inverse

$$\frac{\partial l(A,C,Q,R \mid x,y)}{\partial R^{-1}} = \frac{T+1}{2}R - \frac{1}{2}\left(\sum_{t=0}^{T}\mathbf{y}_t\mathbf{y}_t^T - \mathbf{y}_t\mathbf{x}_t^T C^T - C\mathbf{x}_t\mathbf{y}_t^T + C\mathbf{x}_t\mathbf{x}_t^T C^T\right)^T$$

- Maximizer

$$R = \frac{1}{T+1}\left(\sum_{t=0}^{T}\mathbf{y}_t\mathbf{y}_t^T - \mathbf{y}_t\mathbf{x}_t^T C^T - C\mathbf{x}_t\mathbf{y}_t^T + C\mathbf{x}_t\mathbf{x}_t^T C^T\right)$$

EM-algorithm

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + \mathbf{w}_t, \quad \mathbf{w}_t \sim W_t = N(\mathbf{0}, Q)$$

$$\mathbf{y}_t = C\mathbf{x}_t + \mathbf{v}_t, \quad \mathbf{v}_t \sim V_t = N(\mathbf{0}, R)$$

- Initial guesses of *A*, *C*, *Q*, *R*
- Kalman smoother (E-step):
  - Compute distributions X_0, …, X_T given data $\mathbf{y}_0$, …, $\mathbf{y}_T$ and *A*, *C*, *Q*, *R*.
- Update parameters (M-step):
  - Update *A*, *C*, *Q*, *R* such that *expected log-likelihood* is maximized
- Repeat until convergence (local optimum)

Kalman Smoother

- for (t = 0; t < T; ++t)   // Kalman filter

$$\hat{\mathbf{x}}_{t+1|t} = A\hat{\mathbf{x}}_{t|t}$$

$$P_{t+1|t} = AP_{t|t}A^T + Q$$

$$K_{t+1} = P_{t+1|t}C^T\left(CP_{t+1|t}C^T + R\right)^{-1}$$

$$\hat{\mathbf{x}}_{t+1|t+1} = \hat{\mathbf{x}}_{t+1|t} + K_{t+1}\left(\mathbf{y}_{t+1} - C\hat{\mathbf{x}}_{t+1|t}\right)$$

$$P_{t+1|t+1} = P_{t+1|t} - K_{t+1}CP_{t+1|t}$$

- for (t = T – 1; t ≥ 0; --t)   // Backward pass

$$L_t = P_{t|t}A^T P_{t+1|t}^{-1}$$

$$\hat{\mathbf{x}}_{t|T} = \hat{\mathbf{x}}_{t|t} + L_t\left(\hat{\mathbf{x}}_{t+1|T} - \hat{\mathbf{x}}_{t+1|t}\right)$$

$$P_{t|T} = P_{t|t} + L_t(P_{t+1|T} - P_{t+1|t})L_t^T$$

Update Parameters

- Likelihood in terms of **x**, but only X available

$$l(A,C,Q,R\,|\,\mathbf{x},\mathbf{y}) =$$

$$\frac{T}{2}\log\left|Q^{-1}\right| - \frac{1}{2}\mathrm{Tr}\left(Q^{-1}\left(\sum_{t=0}^{T-1}\mathbf{x}_{t+1}\mathbf{x}_{t+1}^{T} - \mathbf{x}_{t+1}\mathbf{x}_{t}^{T}A^{T} - A\mathbf{x}_{t}\mathbf{x}_{t+1}^{T} + A\mathbf{x}_{t}\mathbf{x}_{t}^{T}A^{T}\right)\right) +$$

$$\frac{T+1}{2}\log\left|R^{-1}\right| - \frac{1}{2}\mathrm{Tr}\left(R^{-1}\left(\sum_{t=0}^{T}\mathbf{y}_{t}\mathbf{y}_{t}^{T} - \mathbf{y}_{t}\mathbf{x}_{t}^{T}C^{T} - C\mathbf{x}_{t}\mathbf{y}_{t+1}^{T} + C\mathbf{x}_{t}\mathbf{x}_{t}^{T}C^{T}\right)\right) + \mathrm{const}$$

- Likelihood-function linear in $\mathbf{x}_t, \mathbf{x}_t\mathbf{x}_t^T, \mathbf{x}_t\mathbf{x}_{t+1}^T$
- Expected likelihood: replace them with:

$$E(X_t\,|\,\mathbf{y}) = \hat{\mathbf{x}}_{t|T}$$

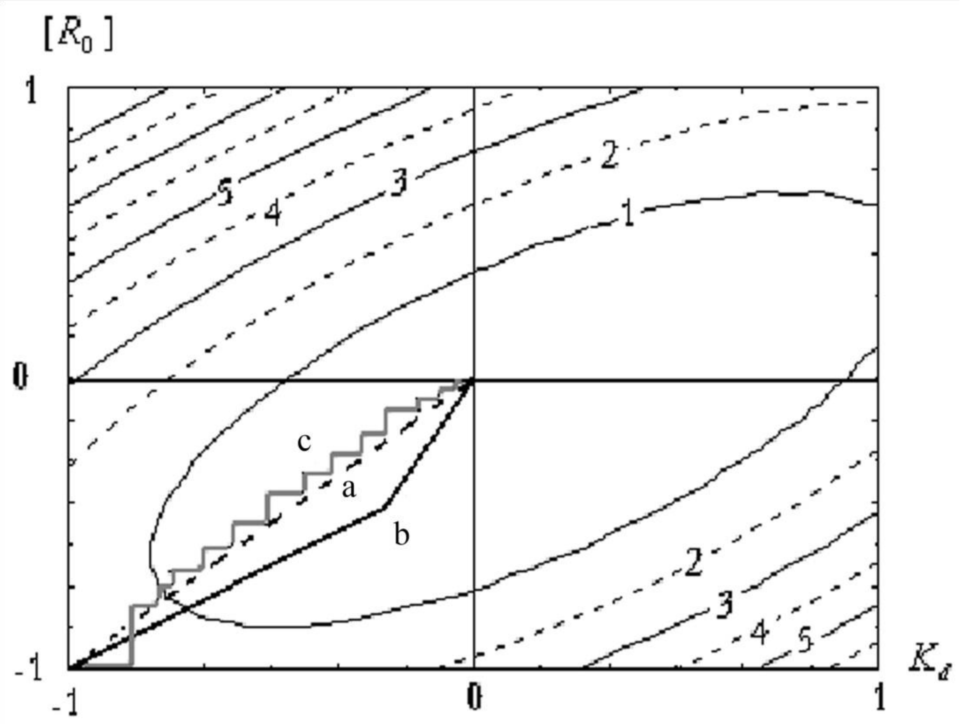
$$E(X_t X_t^T\,|\,\mathbf{y}) = P_{t|T} + \hat{\mathbf{x}}_{t|T}\hat{\mathbf{x}}_{t|T}^T$$

$$E(X_t X_{t+1}^T\,|\,\mathbf{y}) = \hat{\mathbf{x}}_{t|t}\hat{\mathbf{x}}_{t+1|T}^T + L_t\left(P_{t+1|T} + (\hat{\mathbf{x}}_{t+1|T} - \hat{\mathbf{x}}_{t+1|t})\hat{\mathbf{x}}_{t+1|T}^T\right)$$

- Use maximizers to update A, C, Q and R.

Convergence

- Convergence is guaranteed to local optimum
- Similar to coordinate ascent



Conclusion

- EM-algorithm to simultaneously optimize state estimates and model parameters
- Given ``training data'', EM-algorithm can be used (off-line) to *learn* the model for subsequent use in (real-time) Kalman filters

Next time

- Learning from demonstrations
- Dynamic Time Warping