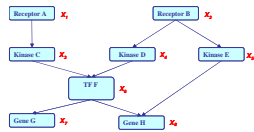


Factor Analysis and State Space Models

Probabilistic Graphical Models (10-708)

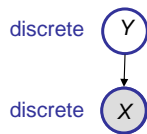
Lecture 13, Oct 31, 2007



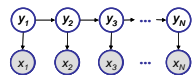
Eric Xing

Reading: J-Chap. 13,14

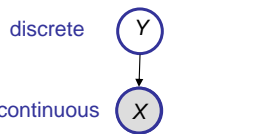
A road map to more complex dynamic models



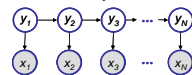
Mixture model
e.g., mixture of multinomials



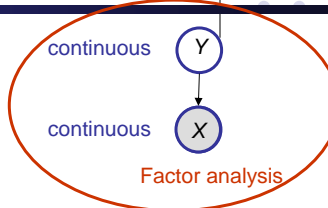
HMM
(for discrete sequential data, e.g., text)



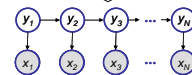
Mixture model
e.g., mixture of Gaussians



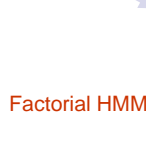
HMM
(for continuous sequential data, e.g., speech signal)



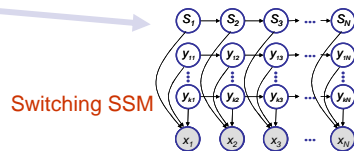
Factor analysis



State space model



Factorial HMM



Switching SSM

Review: A primer to multivariate Gaussian



- Multivariate Gaussian density:

$$p(\mathbf{x} | \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right\}$$

- A joint Gaussian:

$$p\left(\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \middle| \mu, \Sigma\right) = \mathcal{N}\left(\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \middle| \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$

- How to write down $p(\mathbf{x}_1)$, $p(\mathbf{x}_1|\mathbf{x}_2)$ or $p(\mathbf{x}_2|\mathbf{x}_1)$ using the block elements in μ and Σ ?

- Formulas to remember:

$$p(\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_2 | \mathbf{m}_2^m, \mathbf{V}_2^m)$$

$$p(\mathbf{x}_1|\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_1 | \mathbf{m}_{12}, \mathbf{V}_{12})$$

$$\mathbf{m}_2^m = \mu_2$$

$$\mathbf{m}_{12} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2)$$

$$\mathbf{V}_2^m = \Sigma_{22}$$

$$\mathbf{V}_{12} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Eric Xing

3

Review: The matrix inverse lemma



- Consider a block-partitioned matrix: $M = \begin{bmatrix} E & F \\ G & H \end{bmatrix}$

- First we diagonalize M

$$\begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} E & F \\ G & H \end{bmatrix} \begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix} = \begin{bmatrix} E-FH^{-1}G & 0 \\ 0 & H \end{bmatrix}$$

- Schur complement: $M/H = E-FH^{-1}G$

- Then we inverse, using this formula: $XYZ = W \Rightarrow Y^{-1} = ZW^{-1}X$

$$\begin{aligned} M^{-1} &= \begin{bmatrix} E & F \\ G & H \end{bmatrix}^{-1} = \begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix} \begin{bmatrix} (M/H)^{-1} & 0 \\ 0 & H^{-1} \end{bmatrix} \begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1} + H^{-1}G(M/H)^{-1}FH^{-1} \end{bmatrix} = \begin{bmatrix} E^{-1} + E^{-1}F(M/E)^{-1}GE^{-1} & -E^{-1}F(M/E)^{-1} \\ -(M/E)^{-1}GE^{-1} & (M/E)^{-1} \end{bmatrix} \end{aligned}$$

- Matrix inverse lemma

$$(E-FH^{-1}G)^{-1} = E^{-1} + E^{-1}F(H-GE^{-1}F)^{-1}GE^{-1}$$

Eric Xing

4

Review: Some matrix algebra



- Trace and derivatives $\text{tr}[A] \stackrel{\text{def}}{=} \sum_i a_{ii}$
 - Cyclical permutations

$$\text{tr}[ABC] = \text{tr}[CAB] = \text{tr}[BCA]$$

- Derivatives

$$\frac{\partial}{\partial A} \text{tr}[BA] = B^T$$

$$\frac{d}{dA} \text{tr}[x^T A x] = A + A^T$$

$$\frac{\partial}{\partial A} \text{tr}[x^T A x] = \frac{\partial}{\partial A} \text{tr}[x x^T A] = x x^T$$

- Determinants and derivatives

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

Factor analysis



- An unsupervised linear regression model **FA**



$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, I)$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}; \mu + \Lambda \mathbf{x}, \Psi)$$

where Λ is called a factor loading matrix, and Ψ is diagonal.

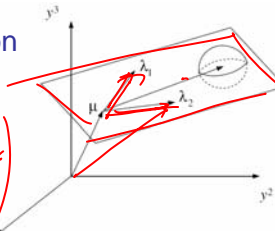
regression (x, y) obs.
 $y = \Lambda x + \epsilon$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} | & | \\ \Lambda & \Lambda \\ | & | \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \text{noise}$$

- Geometric interpretation

$$p(x) \Leftrightarrow p(x, y)$$

$$\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix} \quad \begin{pmatrix} \Sigma_x & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_y \end{pmatrix}$$



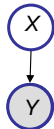
goal: $p(x|y)$

- To generate data, first generate a point within the manifold then add noise. Coordinates of point are components of latent variable.

Marginal data distribution



- A marginal Gaussian (e.g., $p(\mathbf{x})$) times a conditional Gaussian (e.g., $p(\mathbf{y}|\mathbf{x})$) is a joint Gaussian
- Any marginal (e.g., $p(\mathbf{y})$) of a joint Gaussian (e.g., $p(\mathbf{x},\mathbf{y})$) is also a Gaussian
 - Since the marginal is Gaussian, we can determine it by just computing its mean and variance. (Assume noise uncorrelated with data.)



$$\begin{aligned}
 E[\mathbf{Y}] &= E[\mu + \Lambda \mathbf{X} + \mathbf{W}] \quad \text{where } \mathbf{W} \sim \mathcal{N}(\mathbf{0}, \Psi) \\
 &= \mu + \Lambda E[\mathbf{X}] + E[\mathbf{W}] \\
 &= \mu + \mathbf{0} + \mathbf{0} = \mu \\
 \text{Var}[\mathbf{Y}] &= E[(\mathbf{Y} - \mu)(\mathbf{Y} - \mu)^T] \\
 &= E[(\mu + \Lambda \mathbf{X} + \mathbf{W} - \mu)(\mu + \Lambda \mathbf{X} + \mathbf{W} - \mu)^T] \\
 &= E[(\Lambda \mathbf{X} + \mathbf{W})(\Lambda \mathbf{X} + \mathbf{W})^T] \\
 &= \Lambda E[\mathbf{X}\mathbf{X}^T] \Lambda^T + E[\mathbf{W}\mathbf{W}^T] \\
 &= \Lambda \Lambda^T + \Psi
 \end{aligned}$$

$p(\mathbf{x}) \sim \mathcal{N}(\mu, \Lambda \Lambda^T)$
 $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \Psi)$

Eric Xing

7

FA = Constrained-Covariance Gaussian

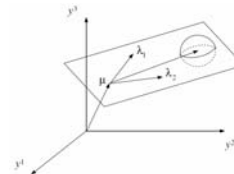


- Marginal density for factor analysis (\mathbf{y} is p -dim, \mathbf{x} is k -dim):

$$p(\mathbf{y} | \theta) = \mathcal{N}(\mathbf{y}; \mu, \Lambda \Lambda^T + \Psi)$$

- So the effective covariance is the low-rank outer product of two long skinny matrices plus a diagonal matrix:

$$\text{Cov}[\mathbf{y}] = \Lambda \Lambda^T + \Psi$$



- In other words, factor analysis is just a constrained Gaussian model. (If Ψ were not diagonal then we could model any Gaussian and it would be pointless.)

Eric Xing

8

FA joint distribution



- Model

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, I)$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}; \mu + \Lambda\mathbf{x}, \Psi)$$

- Covariance between \mathbf{x} and \mathbf{y}

$$\begin{aligned} \text{Cov}[\mathbf{X}, \mathbf{Y}] &= E[(\mathbf{X} - \mathbf{0})(\mathbf{Y} - \mu)^T] = E[\mathbf{X}(\mu + \Lambda\mathbf{X} + \mathbf{W} - \mu)^T] \\ &= E[\mathbf{X}\mathbf{X}^T\Lambda^T + \mathbf{X}\mathbf{W}^T] \\ &= \Lambda^T \end{aligned}$$

$p(\mathbf{y})$

- Hence the joint distribution of \mathbf{x} and \mathbf{y} :

$$p\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \middle| \begin{bmatrix} \mathbf{0} \\ \mu \end{bmatrix}, \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix}\right)$$

$p(\mathbf{x}, \mathbf{y})$

$p(\mathbf{y}|\mathbf{x})$

- Assume noise is uncorrelated with data or latent variables.

Eric Xing

9

Inference in Factor Analysis



- Apply the Gaussian conditioning formulas to the joint distribution we derived above, where

$$\Sigma_{11} = I$$

$$\Sigma_{12} = \Sigma_{12}^T = \Lambda^T$$

$$\Sigma_{22} = (\Lambda\Lambda^T + \Psi)$$

$x \rightarrow y \rightarrow x|y$

we can now derive the posterior of the latent variable \mathbf{x} given observation \mathbf{y} , $p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x} | \mathbf{m}_{12}, \mathbf{V}_{12})$, where

$$\mathbf{m}_{12} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y} - \mu_2)$$

$$\mathbf{V}_{12} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

$$= \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}(\mathbf{y} - \mu)$$

$$= I - \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}\Lambda$$

Applying the matrix inversion lemma $(E - FH^{-1}G)^{-1} = E^{-1} + E^{-1}F(H - GE^{-1}F)^{-1}GE^{-1}$

$$\Rightarrow \mathbf{V}_{12} = (I + \Lambda^T\Psi^{-1}\Lambda)^{-1}$$

$$\mathbf{m}_{12} = \mathbf{V}_{12}\Lambda^T\Psi^{-1}(\mathbf{y} - \mu)$$

- Here we only need to invert a matrix of size $|\mathbf{x}| \times |\mathbf{x}|$, instead of $|\mathbf{y}| \times |\mathbf{y}|$.

Eric Xing

10

Geometric interpretation: inference is linear projection

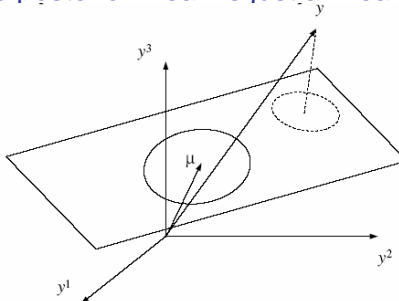


- The posterior is:

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}; \mathbf{m}_{12}, \mathbf{V}_{12})$$

$$\mathbf{V}_{12} = (\mathbf{I} + \mathbf{\Lambda}^T \mathbf{\Psi}^{-1} \mathbf{\Lambda})^{-1} \quad \mathbf{m}_{12} = \mathbf{V}_{12} \mathbf{\Lambda}^T \mathbf{\Psi}^{-1} (\mathbf{y} - \boldsymbol{\mu})$$

- Posterior covariance does not depend on observed data \mathbf{y} !
- Computing the posterior mean is just a linear operation:



Eric Xing

11

EM for Factor Analysis

$\eta. \quad \mathbf{x} =$
 $\mathbf{y} = \mathbf{\Lambda} \mathbf{x} + \boldsymbol{\phi}$



- Incomplete data log likelihood function (marginal density of \mathbf{y})

$$\begin{aligned} \ell(\theta, D) &= -\frac{N}{2} \log |\mathbf{\Lambda} \mathbf{\Lambda}^T + \mathbf{\Psi}| - \frac{1}{2} \sum_n (\mathbf{y}_n - \boldsymbol{\mu})^T (\mathbf{\Lambda} \mathbf{\Lambda}^T + \mathbf{\Psi})^{-1} (\mathbf{y}_n - \boldsymbol{\mu}) \\ &= -\frac{N}{2} \log |\mathbf{\Lambda} \mathbf{\Lambda}^T + \mathbf{\Psi}| - \frac{1}{2} \text{tr} [(\mathbf{\Lambda} \mathbf{\Lambda}^T + \mathbf{\Psi})^{-1} \mathbf{S}], \quad \text{where } \mathbf{S} = \sum_n (\mathbf{y}_n - \boldsymbol{\mu})(\mathbf{y}_n - \boldsymbol{\mu})^T \end{aligned}$$

$p(\mathbf{y}) = f(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$

- Estimating \mathbf{m} is trivial: $\hat{\boldsymbol{\mu}}^{ML} = \frac{1}{N} \sum_n \mathbf{y}_n$
- Parameters $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ are coupled nonlinearly in log-likelihood
- Complete log likelihood

$$\begin{aligned} \ell_c(\theta, D) &= \sum_n \log p(\mathbf{x}_n, \mathbf{y}_n) = \sum_n \log p(\mathbf{x}_n) + \log p(\mathbf{y}_n | \mathbf{x}_n) \\ &= -\frac{N}{2} \log |I| - \frac{1}{2} \sum_n \mathbf{x}_n^T \mathbf{x}_n - \frac{N}{2} \log |\mathbf{\Psi}| - \frac{1}{2} \sum_n (\mathbf{y}_n - \mathbf{\Lambda} \mathbf{x}_n)^T \mathbf{\Psi}^{-1} (\mathbf{y}_n - \mathbf{\Lambda} \mathbf{x}_n) \\ &= -\frac{N}{2} \log |\mathbf{\Psi}| - \frac{1}{2} \text{tr} [\mathbf{X}_n^T \mathbf{X}_n] - \frac{N}{2} \text{tr} [\mathbf{S} \mathbf{\Psi}^{-1}], \quad \text{where } \mathbf{S} = \frac{1}{N} \sum_n (\mathbf{y}_n - \mathbf{\Lambda} \mathbf{x}_n)(\mathbf{y}_n - \mathbf{\Lambda} \mathbf{x}_n)^T \end{aligned}$$

Eric Xing

12

E-step for Factor Analysis



- Compute $\langle \ell_c(\theta, \mathcal{D}) \rangle_{p(\mathbf{x}|\mathbf{y})}$

$$\langle \ell_c(\theta, \mathcal{D}) \rangle = -\frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_n \text{tr}[\langle \mathbf{X}_n \mathbf{X}_n^T \rangle] - \frac{N}{2} \text{tr}[\langle \mathbf{S} \rangle \Psi^{-1}]$$

$$\langle \mathbf{S} \rangle = \frac{1}{N} \sum_n (\mathbf{y}_n \mathbf{y}_n^T - \mathbf{y}_n \langle \mathbf{X}_n^T \rangle \Lambda^T - \Lambda \langle \mathbf{X}_n^T \rangle \mathbf{y}_n^T + \Lambda \langle \mathbf{X}_n \mathbf{X}_n^T \rangle \Lambda^T)$$

$$\langle \mathbf{X}_n \rangle = E[\mathbf{X}_n | \mathbf{y}_n]$$

$$\langle \mathbf{X}_n \mathbf{X}_n^T \rangle = \text{Var}[\mathbf{X}_n | \mathbf{y}_n] + E[\mathbf{X}_n | \mathbf{y}_n] E[\mathbf{X}_n | \mathbf{y}_n]^T$$

$P(\mathbf{x}|\mathbf{y})$

- Recall that we have derived:

$$\mathbf{V}_{1|2} = (\mathbf{I} + \Lambda^T \Psi^{-1} \Lambda)^{-1} \quad \mathbf{m}_{1|2} = \mathbf{V}_{1|2} \Lambda^T \Psi^{-1} (\mathbf{y} - \boldsymbol{\mu})$$

$$\Rightarrow \langle \mathbf{X}_n \rangle = \mathbf{m}_{\mathbf{x}_n | \mathbf{y}_n} = \mathbf{V}_{1|2} \Lambda^T \Psi^{-1} (\mathbf{y}_n - \boldsymbol{\mu}) \quad \text{and} \quad \langle \mathbf{X}_n \mathbf{X}_n^T \rangle = \mathbf{V}_{1|2} + \mathbf{m}_{\mathbf{x}_n | \mathbf{y}_n} \mathbf{m}_{\mathbf{x}_n | \mathbf{y}_n}^T$$

Eric Xing

13

M-step for Factor Analysis



- Take the derivatives of the expected complete log likelihood wrt. parameters.
 - Using the trace and determinant derivative rules:

$$\begin{aligned} \frac{\partial}{\partial \Psi^{-1}} \langle \ell_c \rangle &= \frac{\partial}{\partial \Psi^{-1}} \left(-\frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_n \text{tr}[\langle \mathbf{X}_n \mathbf{X}_n^T \rangle] - \frac{N}{2} \text{tr}[\langle \mathbf{S} \rangle \Psi^{-1}] \right) \\ &= \frac{N}{2} \Psi - \frac{N}{2} \langle \mathbf{S} \rangle \stackrel{= \mathcal{U}}{\Rightarrow} \Psi^{t+1} = \langle \mathbf{S} \rangle \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \Lambda} \langle \ell_c \rangle &= \frac{\partial}{\partial \Lambda} \left(-\frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_n \text{tr}[\langle \mathbf{X}_n \mathbf{X}_n^T \rangle] - \frac{N}{2} \text{tr}[\langle \mathbf{S} \rangle \Psi^{-1}] \right) = -\frac{N}{2} \Psi^{-1} \frac{\partial}{\partial \Lambda} \langle \mathbf{S} \rangle \\ &= -\frac{N}{2} \Psi^{-1} \frac{\partial}{\partial \Lambda} \left(\frac{1}{N} \sum_n (\mathbf{y}_n \mathbf{y}_n^T - \mathbf{y}_n \langle \mathbf{X}_n^T \rangle \Lambda^T - \Lambda \langle \mathbf{X}_n^T \rangle \mathbf{y}_n^T + \Lambda \langle \mathbf{X}_n \mathbf{X}_n^T \rangle \Lambda^T) \right) \\ &= \Psi^{-1} \sum_n \mathbf{y}_n \langle \mathbf{X}_n^T \rangle - \Psi^{-1} \Lambda \sum_n \langle \mathbf{X}_n \mathbf{X}_n^T \rangle \Rightarrow \Lambda^{t+1} = \left(\sum_n \mathbf{y}_n \langle \mathbf{X}_n^T \rangle \right) \left(\sum_n \langle \mathbf{X}_n \mathbf{X}_n^T \rangle \right)^{-1} \end{aligned}$$

Eric Xing

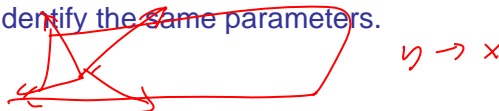
14



Model Invariance and Identifiability



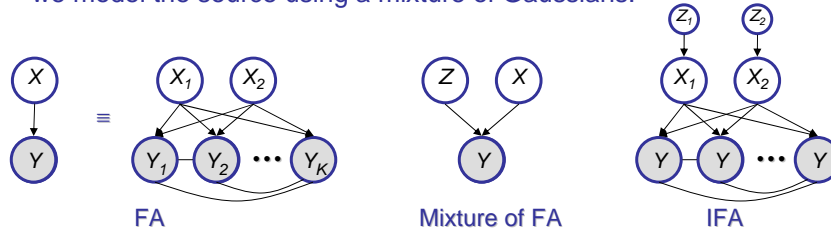
- There is *degeneracy* in the FA model. A.G.M
- Since Λ only appears as outer product $\Lambda\Lambda^T$, the model is invariant to rotation and axis flips of the latent space.
- We can replace Λ with ΛQ for any orthonormal matrix Q and the model remains the same: $(\Lambda Q)(\Lambda Q)^T = \Lambda(QQ^T)\Lambda^T = \Lambda\Lambda^T$.
- This means that there is no “one best” setting of the parameters. An infinite number of parameters all give the ML score!
- Such models are called *un-identifiable* since two people both fitting ML parameters to the identical data will not be guaranteed to identify the same parameters.



Independent Components Analysis (ICA)



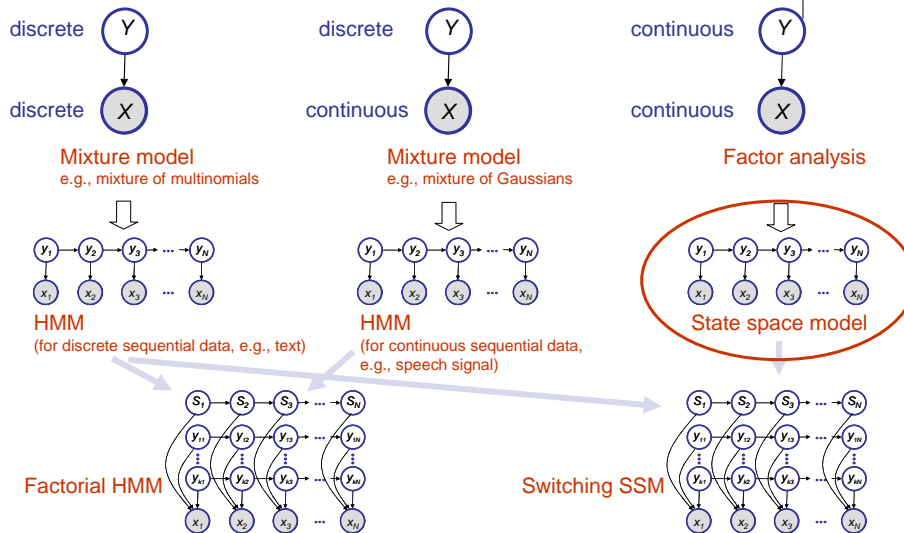
- ICA is similar to FA, except it assumes the latent source has non-Gaussian density.
- Hence ICA can extract higher order moments (not just second order).
- It is commonly used to solve blind source separation (cocktail party problem).
- Independent Factor Analysis (IFA) is an approximation to ICA where we model the source using a mixture of Gaussians.



Eric Xing

17

A road map to more complex dynamic models



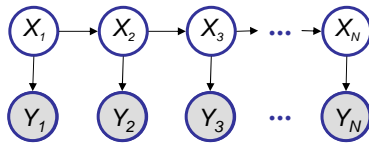
Eric Xing

18

State space models (SSM)



- A sequential FA or a continuous state HMM



$$\mathbf{x}_t = A\mathbf{x}_{t-1} + G\mathbf{w}_t$$

$$\mathbf{y}_t = C\mathbf{x}_{t-1} + V_t$$

$$\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}; Q), \quad v_t \sim \mathcal{N}(\mathbf{0}; R)$$

$$\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}; \Sigma_0),$$

This is a linear dynamic system.

- In general,

$$\mathbf{x}_t = f(\mathbf{x}_{t-1}) + G\mathbf{w}_t$$

$$\mathbf{y}_t = g(\mathbf{x}_{t-1}) + V_t$$

where f is an (arbitrary) dynamic model, and g is an (arbitrary) observation model

Eric Xing

19

The inference problem



- Filtering \rightarrow given $\mathbf{y}_1, \dots, \mathbf{y}_t$, estimate x_t
 - The Kalman filter is a way to perform exact online inference (sequential Bayesian updating) in an LDS. It is the Gaussian analog of the forwards algorithm for HMMs:

$$p(X_t = i | \mathbf{y}_{1:t}) \propto \alpha_t^i \propto p(y_t | X_t = i) \sum_j p(X_t = i | X_{t-1} = j) \alpha_{t-1}^j$$

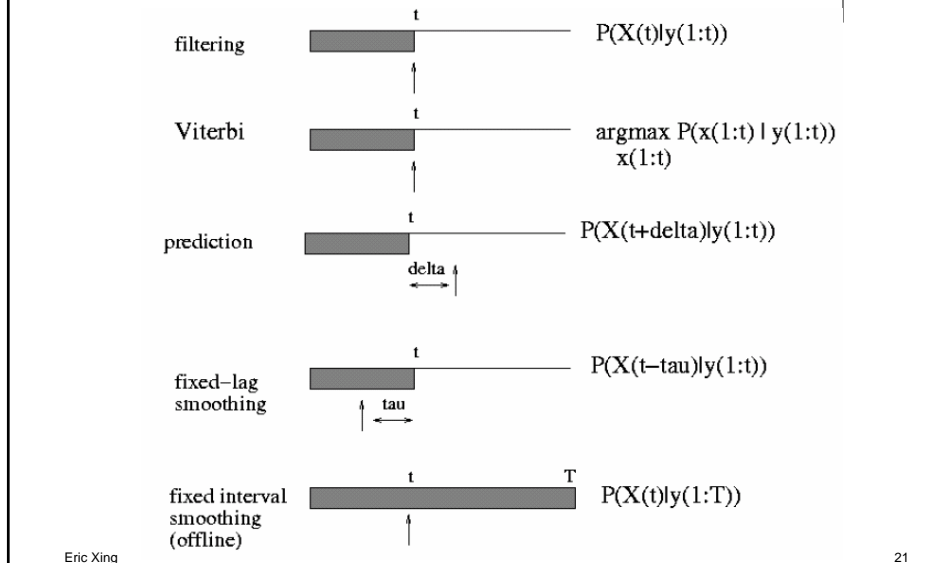
- Smoothing \rightarrow given $\mathbf{y}_1, \dots, \mathbf{y}_T$, estimate x_t ($t < T$)
 - The Rauch-Tung-Strievel smoother is a way to perform exact off-line inference in an LDS. It is the Gaussian analog of the forwards-backwards algorithm:

$$p(X_t = i | \mathbf{y}_{1:T}) \propto \alpha_t^i \beta_t^i$$

Eric Xing

20

Online vs offline inference



LDS for 2D tracking



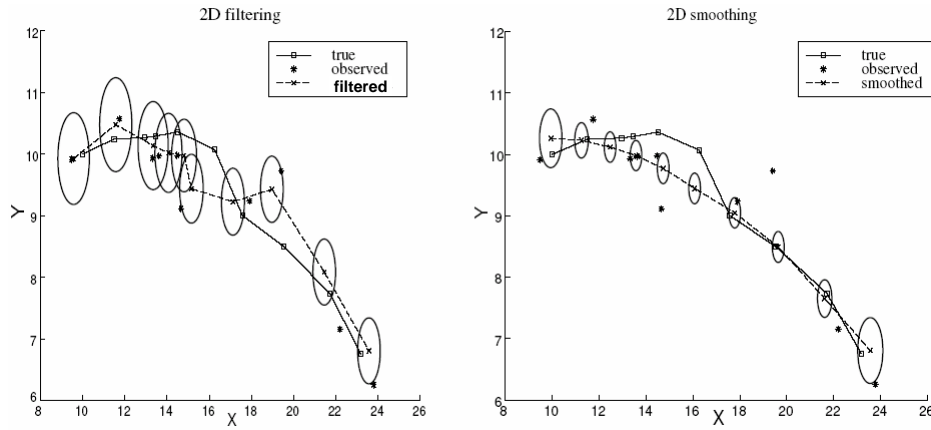
- Dynamics: new position = old position + $\Delta \times$ velocity + noise (constant velocity model, Gaussian noise)

$$\begin{pmatrix} x_t^1 \\ x_t^2 \\ \dot{x}_t^1 \\ \dot{x}_t^2 \end{pmatrix} = \Delta \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_{t-1}^1 \\ x_{t-1}^2 \\ \dot{x}_{t-1}^1 \\ \dot{x}_{t-1}^2 \end{pmatrix} + \text{noise}$$

- Observation: project out first two components (we observe Cartesian position of object - linear!)

$$\begin{pmatrix} y_t^1 \\ y_t^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_t^1 \\ x_t^2 \\ \dot{x}_t^1 \\ \dot{x}_t^2 \end{pmatrix} + \text{noise}$$

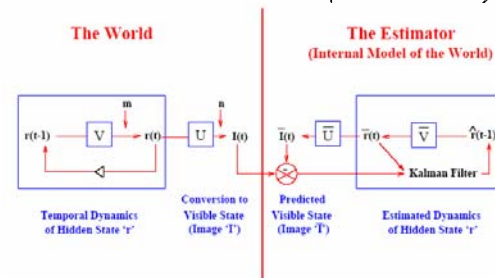
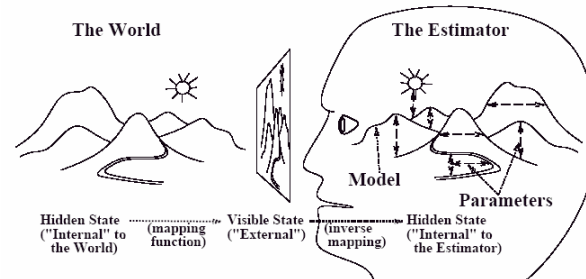
2D tracking



Eric Xing

23

Kalman filtering in the brain?



Eric Xing

24

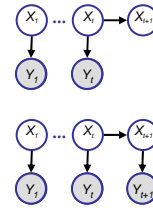
Kalman filtering derivation



- Since all CPDs are linear Gaussian, the system defines a large multivariate Gaussian.
 - Hence all marginals are Gaussian.
 - Hence we can represent the belief state $p(\mathbf{X}_t | \mathbf{y}_{1:t})$ as a Gaussian with mean $\hat{\mathbf{x}}_{t|t} \equiv E(\mathbf{X}_t | \mathbf{y}_1, \dots, \mathbf{y}_t)$ and covariance $P_{t|t} \equiv E(\mathbf{X}_t \mathbf{X}_t^T | \mathbf{y}_1, \dots, \mathbf{y}_t)$.
 - It is common to work with the inverse covariance (precision) matrix $P_{t|t}^{-1}$; this is called information form.

- Kalman filtering is a recursive procedure to update the belief state:

- Predict step: compute $p(\mathbf{X}_{t+1} | \mathbf{y}_{1:t})$ from prior belief $p(\mathbf{X}_t | \mathbf{y}_{1:t})$ and dynamical model $p(\mathbf{X}_{t+1} | \mathbf{X}_t)$ --- **time update**
- Update step: compute new belief $p(\mathbf{X}_{t+1} | \mathbf{y}_{1:t+1})$ from prediction $p(\mathbf{X}_{t+1} | \mathbf{y}_{1:t})$, observation \mathbf{y}_{t+1} and observation model $p(\mathbf{y}_{t+1} | \mathbf{X}_{t+1})$ --- **measurement update**



Eric Xing

25

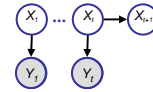
Predict step



- Dynamical Model: $\mathbf{x}_{t+1} = A\mathbf{x}_t + G\mathbf{w}_t, \quad \mathbf{w}_t \sim \mathcal{N}(0; Q)$

- One step ahead prediction of state:

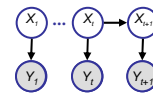
$$\begin{aligned} \hat{\mathbf{x}}_{t+1|t} &= E(\mathbf{X}_{t+1} | \mathbf{y}_1, \dots, \mathbf{y}_t) = A\hat{\mathbf{x}}_{t|t} \\ P_{t+1|t} &= E(\mathbf{X}_{t+1} - \hat{\mathbf{x}}_{t+1|t})(\mathbf{X}_{t+1} - \hat{\mathbf{x}}_{t+1|t})^T | \mathbf{y}_1, \dots, \mathbf{y}_t \\ &= E(A\mathbf{X}_t + G\mathbf{w}_t - \hat{\mathbf{x}}_{t+1|t})(A\mathbf{X}_t + G\mathbf{w}_t - \hat{\mathbf{x}}_{t+1|t})^T | \mathbf{y}_1, \dots, \mathbf{y}_t \\ &= AP_{t|t}A + GQG^T \end{aligned}$$



- Observation model: $\mathbf{y}_t = C\mathbf{x}_t + \mathbf{v}_t, \quad \mathbf{v}_t \sim \mathcal{N}(0; R)$

- One step ahead prediction of observation:

$$\begin{aligned} E(\mathbf{Y}_{t+1} | \mathbf{y}_1, \dots, \mathbf{y}_t) &= E(C\mathbf{X}_{t+1} + \mathbf{v}_{t+1} | \mathbf{y}_1, \dots, \mathbf{y}_t) = C\hat{\mathbf{x}}_{t+1|t} \\ E(\mathbf{Y}_{t+1} - \hat{\mathbf{y}}_{t+1|t})(\mathbf{Y}_{t+1} - \hat{\mathbf{y}}_{t+1|t})^T | \mathbf{y}_1, \dots, \mathbf{y}_t &= CP_{t+1|t}C^T + R \\ E(\mathbf{Y}_{t+1} - \hat{\mathbf{y}}_{t+1|t})(\mathbf{X}_{t+1} - \hat{\mathbf{x}}_{t+1|t})^T | \mathbf{y}_1, \dots, \mathbf{y}_t &= CP_{t+1|t} \end{aligned}$$



Eric Xing

26

Update step



- Summarizing results from previous slide, we have $p(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1} | \mathbf{Y}_{1:t}) \sim \mathcal{N}(m_{t+1}, V_{t+1})$, where

$$m_{t+1} = \begin{pmatrix} \hat{x}_{t+1|t} \\ C\hat{x}_{t+1|t} \end{pmatrix}, \quad V_{t+1} = \begin{pmatrix} P_{t+1|t} & P_{t+1|t}C^T \\ CP_{t+1|t} & CP_{t+1|t}C^T + R \end{pmatrix}$$

- Remember the formulas for conditional Gaussian distributions:

$$p\left(\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \mid \mu, \Sigma\right) = \mathcal{N}\left(\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \mid \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$

$$p(\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_2 \mid \mathbf{m}_2^m, \mathbf{V}_2^m) \quad p(\mathbf{x}_1 \mid \mathbf{x}_2) = \mathcal{N}(\mathbf{x}_1 \mid \mathbf{m}_{1|2}, \mathbf{V}_{1|2})$$

$$\mathbf{m}_2^m = \mu_2 \quad \mathbf{m}_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2)$$

$$\mathbf{V}_2^m = \Sigma_{22} \quad \mathbf{V}_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Eric Xing

27

Kalman Filter



- Measurement updates:

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

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

- where K_{t+1} is the *Kalman gain matrix*

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

- Time updates:

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

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

- K_t can be pre-computed (since it is independent of the data).

Eric Xing

28

Example of KF in 1D

- Consider noisy observations of a 1D particle doing a random walk:

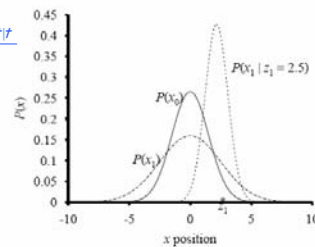
$$x_{t|t-1} = x_{t-1} + w, \quad w \sim \mathcal{N}(0, \sigma_x) \quad z_t = x_t + v, \quad v \sim \mathcal{N}(0, \sigma_z)$$

- KF equations: $P_{t+1|t} = AP_{t|t}A + GQG^T = \sigma_t + \sigma_x$, $\hat{x}_{t+1|t} = A\hat{x}_{t|t} = \hat{x}_{t|t}$

$$K_{t+1} = P_{t+1|t}C^T(CP_{t+1|t}C^T + R)^{-1} = (\sigma_t + \sigma_x)(\sigma_t + \sigma_x + \sigma_z)$$

$$\hat{x}_{t+1|t+1} = \hat{x}_{t+1|t} + K_{t+1}(z_{t+1} - C\hat{x}_{t+1|t}) = \frac{(\sigma_t + \sigma_x)z_{t+1} + \sigma_z\hat{x}_{t|t}}{\sigma_t + \sigma_x + \sigma_z}$$

$$P_{t+1|t+1} = P_{t+1|t} - KCP_{t+1|t} = \frac{(\sigma_t + \sigma_x)\sigma_z}{\sigma_t + \sigma_x + \sigma_z}$$



Eric Xing

29

KF intuition

- The KF update of the mean is

$$\hat{x}_{t+1|t+1} = \hat{x}_{t+1|t} + K_{t+1}(z_{t+1} - C\hat{x}_{t+1|t}) = \frac{(\sigma_t + \sigma_x)z_{t+1} + \sigma_z\hat{x}_{t|t}}{\sigma_t + \sigma_x + \sigma_z}$$

- the term $(z_{t+1} - C\hat{x}_{t+1|t})$ is called the *innovation*
- New belief is convex combination of updates from prior and observation, weighted by Kalman Gain matrix:

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

- If the observation is unreliable, σ_z (i.e., R) is large so K_{t+1} is small, so we pay more attention to the prediction.
- If the old prior is unreliable (large σ_t) or the process is very unpredictable (large σ_x), we pay more attention to the observation.

Eric Xing

30

KF, RLS and LMS



- The KF update of the mean is

$$\hat{\mathbf{x}}_{t+1|t+1} = A\hat{\mathbf{x}}_{t|t} + K_{t+1}(y_{t+1} - C\hat{\mathbf{x}}_{t+1|t})$$

- Consider the special case where the hidden state is a constant, $\mathbf{x}_t = \theta$, but the “observation matrix” C is a time-varying vector, $C = \mathbf{x}_t^T$.
 - Hence the observation model at each time slide, $y_t = \mathbf{x}_t^T \theta + v_t$, is a linear regression

- We can estimate θ recursively using the Kalman filter:

$$\hat{\theta}_{t+1} = \hat{\theta}_t + P_{t+1}R^{-1}(y_{t+1} - \mathbf{x}_t^T \hat{\theta}_t)\mathbf{x}_t$$

This is called the recursive least squares (RLS) algorithm.

- We can approximate $P_{t+1}R^{-1} \approx \eta_{t+1}$ by a scalar constant. This is called the least mean squares (LMS) algorithm.
- We can adapt η_t online using stochastic approximation theory.

Eric Xing

31

Complexity of one KF step



- Let $\mathbf{x}_t \in \mathbb{R}^{N_x}$ and $y_t \in \mathbb{R}^{N_y}$,
- Computing $P_{t+1|t} = AP_{t|t}A + GQG^T$ takes $O(N_x^2)$ time, assuming dense P and dense A .
- Computing $K_{t+1} = P_{t+1|t}C^T(CP_{t+1|t}C^T + R)^{-1}$ takes $O(N_y^3)$ time.
- So overall time is, in general, $\max\{N_x^2, N_y^3\}$

Eric Xing

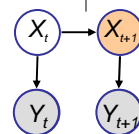
32

Rauch-Tung-Strievel smoother



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

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



- General structure: KF results + the difference of the "smoothed" and predicted results of the next step
- Backward computation: Pretend to know things at t+1 — such conditioning makes things simple and we can remove this condition finally

- The difficulty: $X_t | y_1, \dots, y_T$
- The trick: $E[X | Z] = E[E[X | Y, Z] | Z]$ (Hw!)

$$\text{Var}[X | Z] = \text{Var}[E[X | Y, Z] | Z] + E[\text{Var}[X | Y, Z] | Z]$$

$$\hat{x}_{t|T} \stackrel{\text{def}}{=} E[X_t | y_1, \dots, y_T] = E[E[X_t | X_{t+1}, y_1, \dots, y_T] | y_1, \dots, y_T]$$

$$= E[E[X_t | X_{t+1}, y_1, \dots, y_t] | y_1, \dots, y_t]$$

$$= E[X_t | X_{t+1}, y_1, \dots, y_t]$$

Same for $P_{t|T}$

RTS derivation



- Following the results from previous slide, we need to derive $p(\mathbf{X}_{t+1}, \mathbf{X}_t | y_{1:t}) \sim \mathcal{N}(m, V)$, where

$$m = \begin{pmatrix} \hat{x}_{t|T} \\ \hat{x}_{t+1|T} \end{pmatrix}, \quad V = \begin{pmatrix} P_{t|T} & P_{t|T} A^T \\ A P_{t|T} & P_{t+1|T} \end{pmatrix}$$

- all the quantities here are available after a forward KF pass
- Remember the formulas for conditional Gaussian distributions:

$$p\left(\begin{matrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{matrix} \middle| \mu, \Sigma\right) = \mathcal{N}\left(\begin{matrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{matrix} \middle| \begin{matrix} \mu_1 \\ \mu_2 \end{matrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right), \quad \begin{matrix} p(\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_2 | \mathbf{m}_2^m, \mathbf{V}_2^m) \\ \mathbf{m}_2^m = \mu_2 \\ \mathbf{V}_2^m = \Sigma_{22} \end{matrix}$$

$$p(\mathbf{x}_1 | \mathbf{x}_2) = \mathcal{N}(\mathbf{x}_1 | \mathbf{m}_{12}, \mathbf{V}_{12}) \quad \begin{matrix} \mathbf{m}_{12} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (\mathbf{x}_2 - \mu_2) \\ \mathbf{V}_{12} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \end{matrix}$$

- The RTS smoother

$$\hat{x}_{t|T} = E[X_t | X_{t+1}, y_1, \dots, y_t]$$

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

$$P_{t|T} \stackrel{\text{def}}{=} \text{Var}[\hat{x}_{t|T} | y_{1:T}] + E[\text{Var}[X_t | X_{t+1}, y_{1:T}] | y_{1:T}]$$

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

Learning SSMs



- Complete log likelihood

$$\begin{aligned} \ell_c(\theta, D) &= \sum_n \log p(x_n, y_n) = \sum_n \log p(x_1) + \sum_n \sum_t \log p(x_{n,t} | x_{n,t-1}) + \sum_n \sum_t \log p(y_{n,t} | x_{n,t}) \\ &= f_1(x_1; \Sigma_0) + f_2(\{x_t, x_{t-1}^T, x_t x_t^T, x_t; \forall t\}; A, Q, G) + f_3(\{x_t, x_t^T, x_t; \forall t\}; C, R) \end{aligned}$$

- EM

- E-step: compute $\langle x_t, x_{t-1}^T \rangle, \langle x_t, x_t^T \rangle, \langle x_t \rangle | y_1, \dots, y_T$

these quantities can be inferred via KF and RTS filters, etc.,

$$e, g, \langle x_t, x_t^T \rangle \equiv \text{var}(x_t, x_t^T) + E(x_t)^2 = P_{t|T} + \hat{x}_{t|T}^2$$

- M-step: MLE using

$$\langle \ell_c(\theta, D) \rangle = f_1(\langle x_1 \rangle; \Sigma_0) + f_2(\langle \{x_t, x_{t-1}^T \rangle, \langle x_t, x_t^T \rangle, \langle x_t \rangle; \forall t \rangle; A, Q, G) + f_3(\langle \{x_t, x_t^T \rangle, \langle x_t \rangle; \forall t \rangle; C, R)$$

c.f., M-step in factor analysis

Eric Xing

35

Nonlinear systems



- In robotics and other problems, the motion model and the observation model are often nonlinear:

$$x_t = f(x_{t-1}) + w_t, \quad y_t = g(x_t) + v_t$$

- An optimal closed form solution to the filtering problem is no longer possible.
- The nonlinear functions f and g are sometimes represented by neural networks (multi-layer perceptrons or radial basis function networks).
- The parameters of f and g may be learned offline using EM, where we do gradient descent (back propagation) in the M step, c.f. learning a MRF/CRF with hidden nodes.
- Or we may learn the parameters online by adding them to the state space: $x_t' = (x_t, \theta)$. This makes the problem even more nonlinear.

Eric Xing

36

Extended Kalman Filter (EKF)



- The basic idea of the EKF is to linearize f and g using a second order Taylor expansion, and then apply the standard KF.
- i.e., we approximate a stationary nonlinear system with a non-stationary linear system.

$$x_t = f(\hat{x}_{t-1|t-1}) + A_{\hat{x}_{t-1|t-1}}(x_{t-1} - \hat{x}_{t-1|t-1}) + w_t$$

$$y_t = g(\hat{x}_{t-1|t-1}) + C_{\hat{x}_{t-1|t-1}}(x_t - \hat{x}_{t-1|t-1}) + v_t$$

$$\text{where } \hat{x}_{t-1|t-1} = f(\hat{x}_{t-1|t-1}) \text{ and } A_{\hat{x}} \stackrel{\text{def}}{=} \left. \frac{\partial f}{\partial x} \right|_{\hat{x}} \text{ and } C_{\hat{x}} \stackrel{\text{def}}{=} \left. \frac{\partial g}{\partial x} \right|_{\hat{x}}$$

- The noise covariance (Q and R) is not changed, i.e., the additional error due to linearization is not modeled.