DS-GA 3001.008 Modelling time series data

Lab 3

Refresher: probabilities, graphical models

Derivation Kalman filtering/smoothing

Instructor: Cristina Savin

NYU, CNS & CDS

Basics probabilities

- Random variables **X** represent possible outcomes or states of the world.
- State space: the set of all possible outcomes (discrete/continuous/mixed).
- P(x) denotes probability that **X** takes the value **x**.
- Probability mass/density assigns a nonnegative value to any possible outcome; It sums to 1 (normalized). $\sum_x p(x) = 1$ or $\int_x p(x)dx = 1$.

We use probabilities to assign our beliefs about the state of the world.

Basic probabilities manipulations:

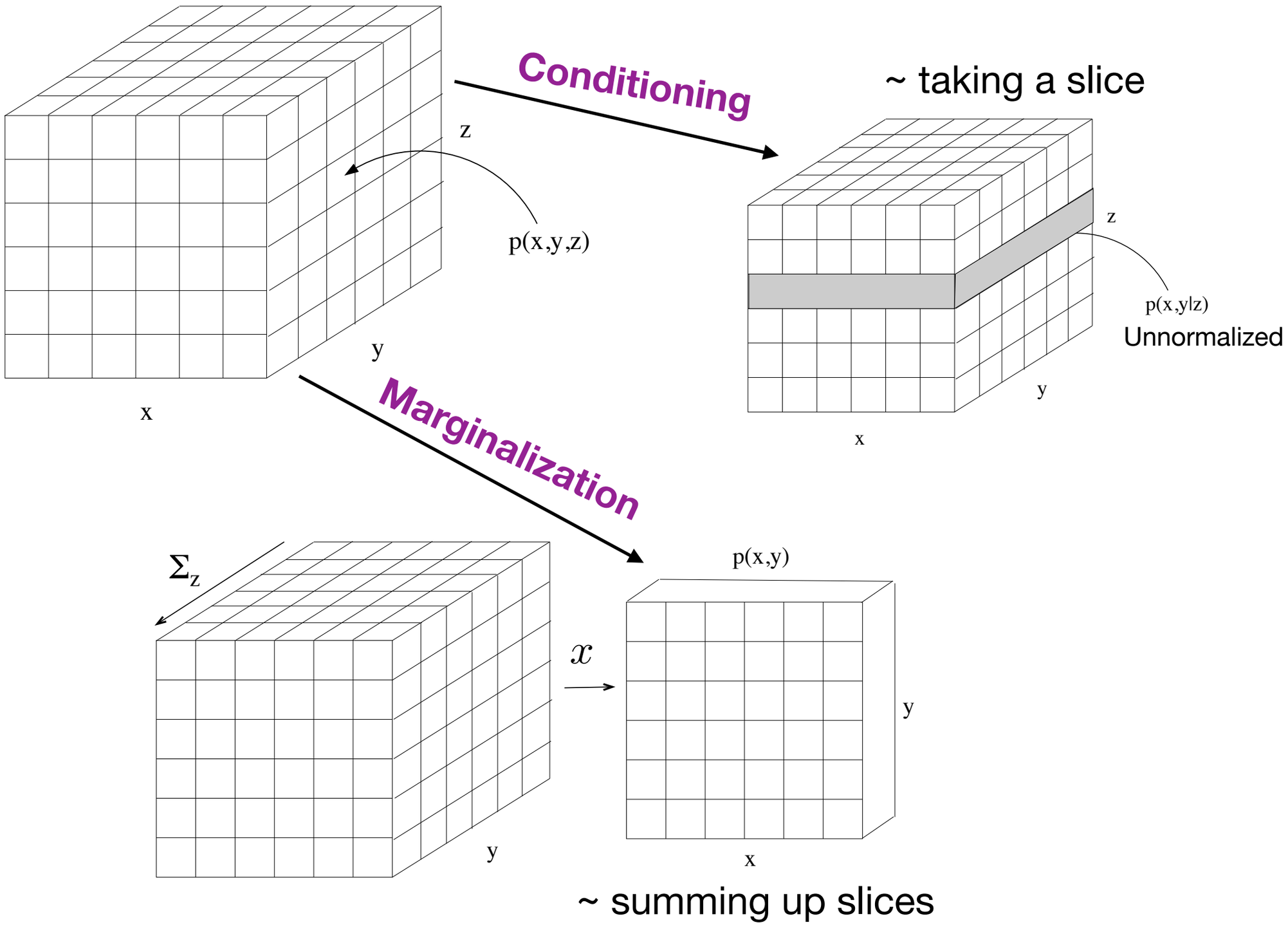
$$\mathrm{P}(A, B) = \mathrm{P}(A|B)\mathrm{P}(B) = \mathrm{P}(B|A)\mathrm{P}(A)$$

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B)} \quad \textbf{Bayes rule}$$

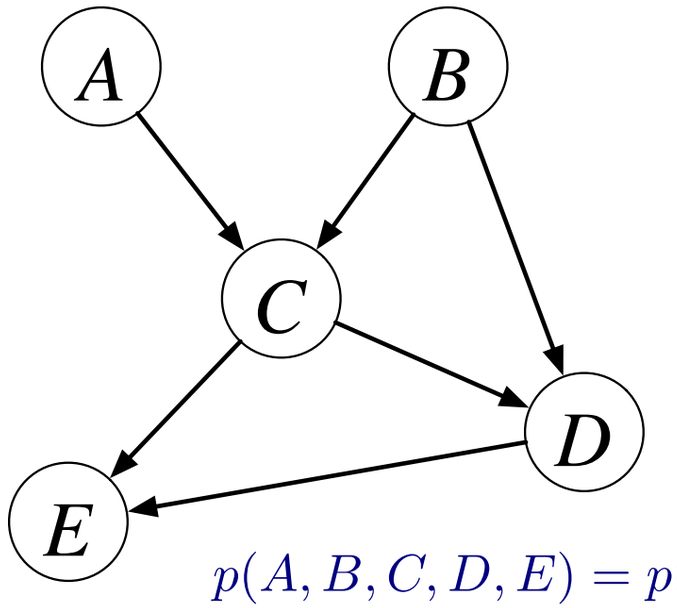
$$\mathrm{P}\left(\mathbf{x}_t, \mathbf{x}_{t-1}, \cdots, \mathbf{x}_0\right) = \prod_k \mathrm{P}\left(\mathbf{x}_k|\mathbf{x}_{k-1}, \cdots, \mathbf{x}_0\right) \quad \textbf{Chain rule}$$

$$p(x, y) = p(x)p(y) \quad \textbf{Independence}$$

Conditioning, marginalization visual



Graphical models



Nodes represent random variables

Edges represent statistical dependencies between variables

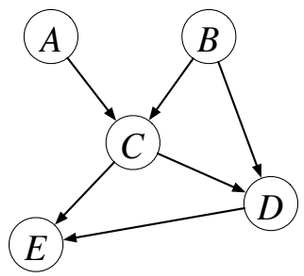
In general: $p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i | X_{\mathsf{pa}(i)})$

$$p(A, B, C, D, E) = p(A)p(B)p(C|A, B)p(D|B, C)p(E|C, D)$$

- Graphical models are an intuitive way of visualising relationships between many variables.
- allows us to abstract out the **conditional independence** relationships between the variables from the details of their parametric form. So we can answer questions like: “Is A dependent on B given that we know the value of C?” just by looking at the graph.
- In this framework we can derive general algorithms and prove their properties for any graph in a particular class, e.g message passing

Inference in graphical models

$$p(\text{variables of interest}|\text{observed variables})$$

Consider the following graph: 

which represents:

$$p(A,B,C,D,E) = p(A)p(B)p(C|A,B)p(D|B,C)p(E|C,D)$$

Computing $p(A|C=c)$.

Naive method:

$$p(A,C=c) = \sum_{B,D,E} p(A,B,C=c,D,E) \qquad \text{[16 terms]}$$

$$p(C=c) = \sum_{A} p(A,C=c) \qquad \text{[2 terms]}$$

$$p(A|C=c) = \frac{p(A,C=c)}{p(C=c)} \qquad \text{[2 terms]}$$

Total: 16+2+2 = 20 terms have to be computed and summed

Learning parameters

Assume a data set $\mathcal{D} = \{\mathbf{x}^{(n)}\}_{n=1}^{N}$. How do we learn $\boldsymbol{\theta}$ from $\mathcal{D}$?

$$p(\mathbf{x}|\boldsymbol{\theta}) = p(x_1|\theta_1)p(x_2|x_1,\theta_2)p(x_3|x_1,\theta_3)p(x_4|x_2,\theta_4)$$

Likelihood:

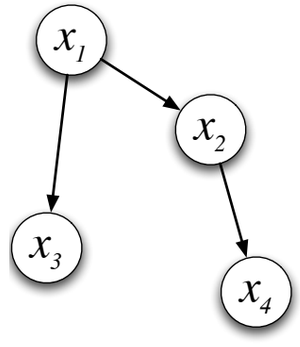
$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\mathbf{x}^{(n)}|\boldsymbol{\theta})$$

Log Likelihood:

$$\log p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{n=1}^{N}\sum_{i} \log p(x_i^{(n)}|x_{\mathrm{pa}(i)}^{(n)},\theta_i)$$

Maximum likelihood

Learning parameters: everything observed



$p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)$

Assume each variable x_i is discrete and can take on K_i values.

The parameters of this model can be represented as 4 tables: θ_1 has K_1 entries, θ_2 has $K_1 \times K_2$ entries, etc.

These are called **conditional probability tables** (CPTs) with the following semantics:

$$p(x_1 = k) = \theta_{1,k} \qquad p(x_2 = k'|x_1 = k) = \theta_{2,k,k'}$$

Log Likelihood:

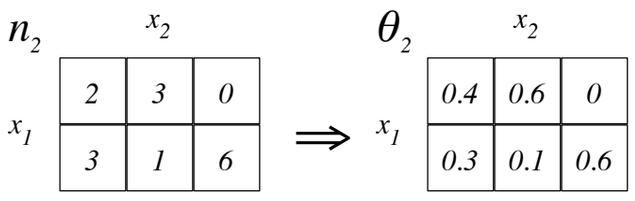
$$\log p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{n=1}^{N} \sum_{i} \log p(x_i^{(n)}|x_{\mathrm{pa}(i)}^{(n)}, \theta_i)$$

This decomposes into sum of functions of θ_i. Each θ_i can be optimized separately:

$$\hat{\theta}_{i,k,k'} = \frac{n_{i,k,k'}}{\sum_{k''} n_{i,k,k''}}$$

where $n_{i,k,k'}$ is the number of times in $\mathcal{D}$ where $x_i = k'$ and $x_{\mathrm{pa}(i)} = k$, where k represents a joint configuration of all the parents of i (i.e. takes on one of $\prod_{j \in \mathrm{pa}(i)} K_j$ values)

ML solution: Simply calculate frequencies!



Learning parameters: with latent variables - use EM

EM alternates between finding a good approximation Q,
and then changing the parameters to improve the approx likelihood

This is done as coordinate-wise ascent on free energy
(lower bound for log likelihood)
We improve one keeping the other fixed, then swap.

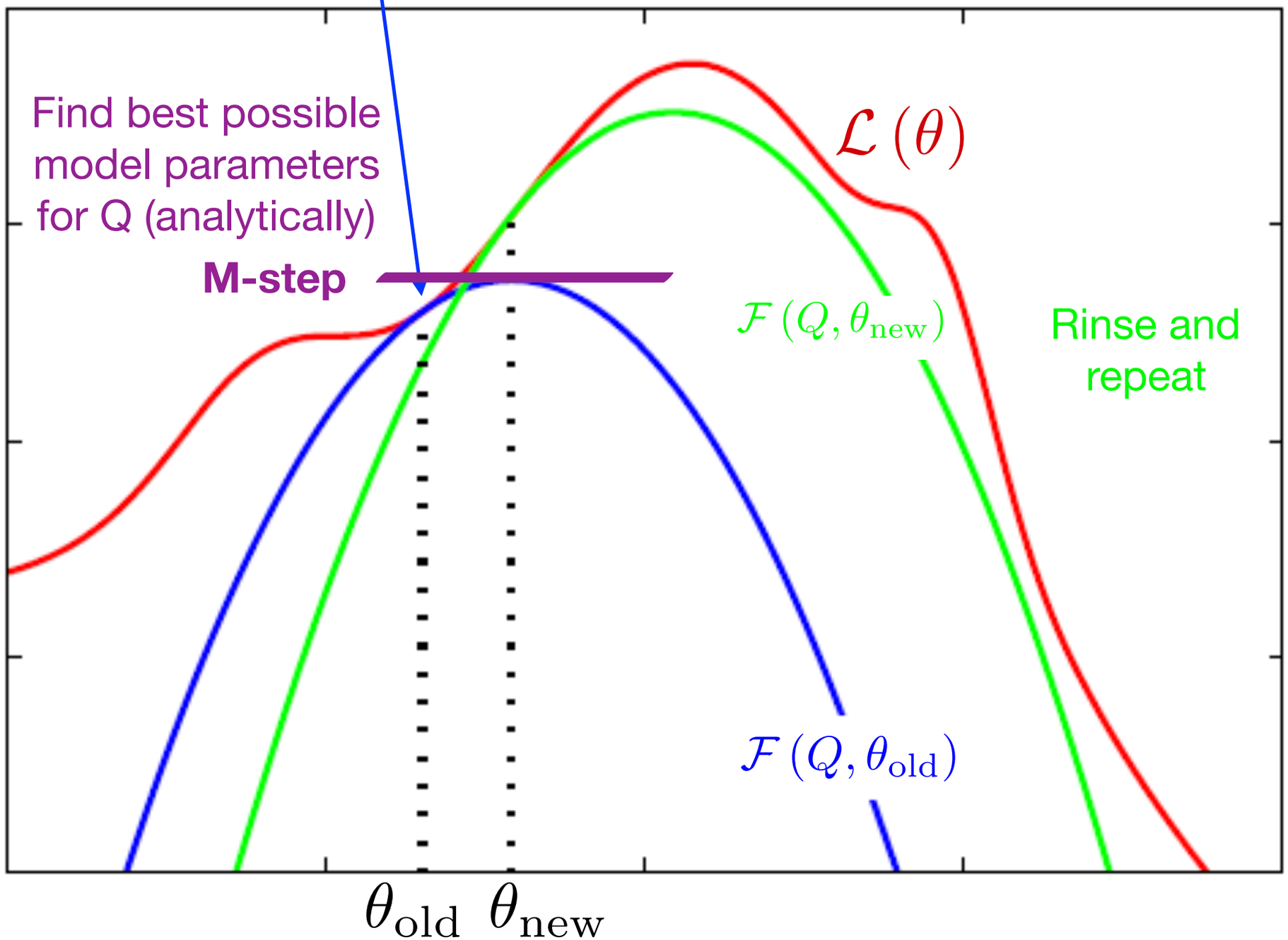
$$\mathcal{F}(Q,\theta) = \int_Q Q(\mathbf{z}) \log \mathrm{P}(\mathbf{x},\mathbf{z}|\theta))\mathbf{dz} - \int_Q Q(\mathbf{z}) \log Q(\mathbf{z})\mathbf{dz}$$

$$\textbf{E step:} \qquad Q_{k+1} \leftarrow \arg\max_{Q} \mathcal{F}(Q,\theta_k)$$

$$\textbf{M step:} \qquad \theta_{k+1} \leftarrow \arg\max_{\theta} \mathcal{F}(Q_{k+1},\theta)$$

E-step: $Q_{k+1}(\mathbf{z}) = \mathrm{P}(\mathbf{z}|\mathbf{x}, \theta_k)$

The approximation is locally exact



Bayesian learning

Data set: $\mathcal{D} = \{x_1, \ldots, x_n\}$ Models: m, m' etc. Model parameters: θ

Prior probability of models: $P(m)$, $P(m')$ etc.
Prior probabilities of model parameters: $P(\theta|m)$
Model of data given parameters (likelihood model): $P(x|\theta, m)$

If the data are independently and identically distributed then:

$$P(\mathcal{D}|\theta, m) = \prod_{i=1}^{n} P(x_i|\theta, m)$$

Posterior probability of model parameters:

$$P(\theta|\mathcal{D}, m) = \frac{P(\mathcal{D}|\theta, m)P(\theta|m)}{P(\mathcal{D}|m)}$$

Posterior probability of models:

$$P(m|\mathcal{D}) = \frac{P(m)P(\mathcal{D}|m)}{P(\mathcal{D})}$$

Basic Gaussian identities

a d-dimensional multidimensional gaussian (normal) density for $\mathbf{x}$ is:

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

E1. Rescaling and shifting gaussian variables

$$\mathrm{E}[\mathbf{A}\mathbf{x}+\mathbf{y}] = \mathbf{A}(\mathrm{E}[\mathbf{x}])+\mathbf{y}$$
$$\mathrm{Covar}[\mathbf{A}\mathbf{x}+\mathbf{y}] = \mathbf{A}(\mathrm{Covar}[\mathbf{x}])\mathbf{A}^T$$

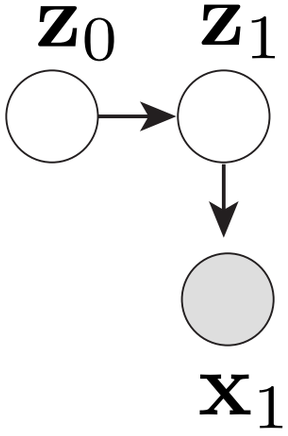
E2. Sum of 2 independent gaussian variables

$$\Sigma_c = \Sigma_a + \Sigma_b$$
$$\mu_c = \mu_a + \mu_b$$

E3. Conditioning

$$\mathbf{z} = \begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix}, \begin{bmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^T & \mathbf{B}\end{bmatrix}\right)$$
$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}\left(\mathbf{a}+\mathbf{C}\mathbf{B}^{-1}(\mathbf{y}-\mathbf{b}), \mathbf{A}-\mathbf{C}\mathbf{B}^{-1}\mathbf{C}^T\right)$$

Let's derive Kalman filter in detail



Forecasting (marginalize z_0)

$$\mu_{1|0} = A\mu_0$$

$$\Sigma_{1|0} = A\Sigma_0 A^{\mathrm{T}} + Q$$

Combine with new evidence

$$\mu_{1|1} = \mu_{1|0} + \mathbf{K}(\mathbf{x}_1 - \mathbf{C}\mu_{1|0})$$

$$\Sigma_{1|1} = \Sigma_{1|0} - \mathbf{K}\mathbf{C}\Sigma_{1|0}$$

$$\mathbf{K} = \Sigma_{1|0}\mathbf{C}^{\mathrm{T}}(\mathbf{C}\Sigma_{1|0}\mathbf{C}^{\mathrm{T}} + R)^{-1}$$

Steps

Use what we know about z_0 to forecast prior for z_1

$$\mathbf{z_1} = \mathbf{A}z_0 + \mathbf{w}_1$$

$$\mathrm{P}(\mathbf{z_1}) = \int P(z_0, z_1) dz_0 = \mathcal{N}(\mathbf{A}\mu_0, \mathbf{A}\Sigma_0\mathbf{A}^{\mathrm{T}} + Q)$$

E1+E2

Combine with evidence

$$\mathbf{x_1} = \mathbf{C}z_1 + \mathbf{v}_1$$

Construct joint distribution $\mathrm{P}(\mathbf{z_1}, \mathbf{x_1})$ which requires $\mathrm{P}(\mathbf{x_1})$ **E1+E2**

$\mathrm{cov}(\mathbf{z_1}, \mathbf{x_1})$

Covariance of linear combinations

Then condition on $\mathbf{x}_1$ **E3**

Let's also derive the Kalman smoother

$$\mu_{i|t} = \mu_{i|i} + \mathbf{F}_i(\mu_{i+1|t} - \mu_{i+1|i})$$

$$\Sigma_{i|t} = \mathbf{F}_i(\Sigma_{i+1|t} - \Sigma_{i+1|i})\mathbf{F}_i^{\mathrm{T}} + \Sigma_{i|i}$$

$$\mathbf{F}_i = \Sigma_{i|i}\mathbf{A}^{\mathrm{T}}\Sigma_{i+1|i}^{-1}$$

Proof (Ainsley&Kohn, tsa4 ref in lecture)

Notation: $\mathbf{x}_{1:i} = \{\mathbf{x}_1, \ldots \mathbf{x}_i\}$

$$\eta_t = \{\mathbf{w}_{i+2}, \ldots \mathbf{w}_t; \mathbf{v}_{i+1}, \ldots \mathbf{v}_t\}$$

We compute: $m_t = \mathbb{E}[\mathbf{z}_i | \mathbf{x}_{1:t}, \mathbf{z}_{i+1} - \mathbf{z}_{i+1|i}, \eta_t]$

independent

Mutually independent

Finally $\mu_{i|t} = \mathbb{E}[m_t | \mathbf{x}_{1:t}]$

Details

We are going to use the conditioning rule again for joint

$$\mathrm{P}(\mathbf{z_i}, \mathbf{z_{i+1}} - \mathbf{z_{i+1|t}} | \mathbf{x}_{1:i})$$

We need the means

$$\mathbb{E}[\mathbf{z_i}|\mathbf{x_{1:i}}] = \mu_{i|i}$$
$$\mathbb{E}[\mathbf{z_{i+1}} - \mathbf{z_{i+1|i}}] = 0$$

The variance $\mathrm{Var}[\mathbf{z_{i+1}} - \mathbf{z_{i+1|i}}] = \Sigma_{i+1|i}$

And covariance $\mathrm{cov}[\mathbf{z_i}, \mathbf{z_{i+1}} - \mathbf{z_{i+1|i}}] = \Sigma_{i|i} A^{\mathrm{t}}$

Together this gives (**E3**): $m_t = \mu_{i|i} + \mathbf{F}_i(\mathbf{z}_{i+1} - \mathbf{z}_{i+1|i})$

Finally $\mu_{i|t} = \mathbb{E}[m_t|\mathbf{x}_{1:t}] = \mu_{i|i} + \mathbf{F}_i(\mathbf{z}_{i+1|t} - \mathbf{z}_{i+1|i})$

For second moment we use the mean derivation and the definition of Cov

$$\mathbf{z}_i - \mathbf{z}_{i|t} = \mathbf{z}_i - \mathbf{z}_{i|i} - \mathbf{F}_i(\mathbf{z}_{i+1|t} - \mathbf{A}\mathbf{z}_{i|i})$$

$$\mathbf{z}_i - \mathbf{z}_{i|t} + \mathbf{F}_i\mathbf{z}_{i+1|t} = \mathbf{z}_i - \mathbf{z}_{i|i}\mathbf{F}_i\mathbf{A}\mathbf{z}_{i|i})$$

multiply each side with its transpose and massage terms

Definition of cov as difference of moments will help

$$\mathbb{E}[\mathbf{x}_{i|i}\mathbf{x}_{i|i}^{\mathrm{t}}] = \mathbb{E}[\mathbf{x}_i\mathbf{x}_i^{\mathrm{t}}] - \Sigma_{i|i}$$