Temporal models

Rico Krueger

Filipe Rodrigues

DTU Management Engineering
Department of Management Engineering

Outline

- Case study: Modeling freeway occupancy rates in San Francisco
- Autoregressive models
- State-space models
  - Linear dynamical systems (LDSs)
  - Hidden Markov models (HMMs)
- Multivariate state-space models

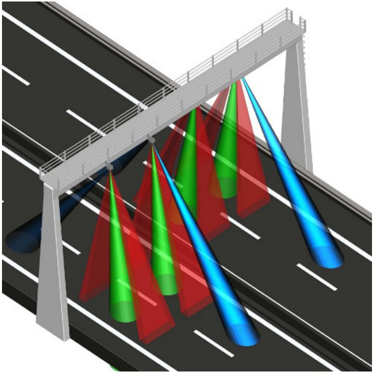
Learning objectives

At the end of this lecture, you should be able to:

- Explain what Markov models are and their underlying assumptions
- Explain the underlying concepts and assumptions behind state-space models
- Explain the difference between forecasting and imputation
- Relate different ways of modelling the temporal dependency between variables and justify their suitability for a problem
- Implement different temporal modelling techniques in STAN/Pyro

Modeling freeway occupancy rates in San Francisco

- Sensors measure occupancy rate (between 0 and 1) of different car lanes
- Time-series of measurements every 10 minutes (144 observations per day)
- Multivariate time-series data: one time-series per sensor
- Goal: **model freeway occupancy rates**
- Some possible applications:
    - Predict future occupancy rates for better routing
    - Identify problems and send alerts to road users
    - Understand drivers' behaviours



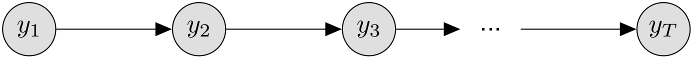
Modeling freeway occupancy rates in San Francisco (cont'd)

- Let's start thinking about the graphical model...
- We shall start by considering **only one sensor**
- Let y_t denote the occupancy rate in that sensor at time t

- How should we connect these (dependencies)?

Modeling freeway occupancy rates in San Francisco (cont'd)

- Let's start thinking about the graphical model...
- We shall start by considering **only one sensor**
- Let y_t denote the occupancy rate in that sensor at time t



- How should we connect these variables (dependencies)?
  - Simplest assumption is to assume that occupancy rate at time t is only dependent on the occupancy rate at time $t-1$
- What other variables could we have considered?
  - Seasonal information
  - Day of the week
  - Weather data
  - Information from other sensors
  - ...but for now we shall consider only data from the time-series $\{y_1, \ldots, y_T\}$
- How should we model the dependency of y_t on y_{t-1}?

Markov models (or Markov chains)

- Simplest mathematical models for random phenomena evolving in time
- Assume that **the present depends only on the recent past**
- Order of the Markov chain defines the number of past variables which directly influence a given variable
- In a first order Markov chain only y_{t-1} influences y_t

$$y_1 \longrightarrow y_2 \longrightarrow y_3 \longrightarrow \cdots \longrightarrow y_T$$

- Joint distribution of all variables factorizes as

$$p(y_1, y_2, \ldots, y_T) = p(y_1) \prod_{t=2}^{T} p(y_t | y_{t-1})$$

- This only requires us to specify a **transition probability** $p(y_t|y_{t-1})$ (well, and some prior distribution for the first observation y_1)

Note

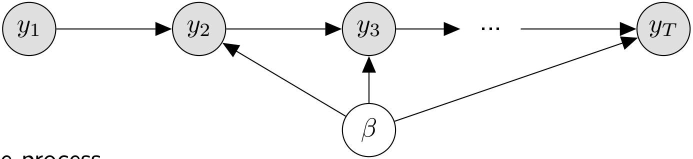
Transition probabilities are time-independent (same transition distribution for all t)

Autoregressive (AR) models

- Standard and widely used Markov model of continuous observations
- Assume that each y_t is a **linear function** of the M previous observations (plus some Gaussian-distributed observation noise term $\epsilon \sim \mathcal{N}(0, \sigma^2)$)

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_M y_{t-M} + \epsilon$$

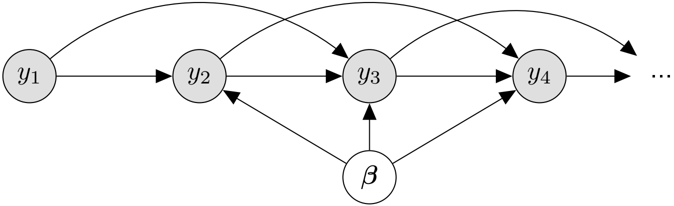
- Graphical model for (Bayesian) autoregressive model of order 1 (AR1)



- Generative process
  1. Draw transition coefficient $\beta \sim \mathcal{N}(\beta|0, \lambda)$
  2. Draw first observation $y_1 \sim \mathcal{N}(y_1|\mu_0, \sigma_0^2)$
  3. For each time $t \in \{2, \ldots, T\}$
     a. Draw observation $y_t \sim \mathcal{N}(y_t|\beta y_{t-1}, \sigma^2)$

Autoregressive (AR) models: higher-level dependencies

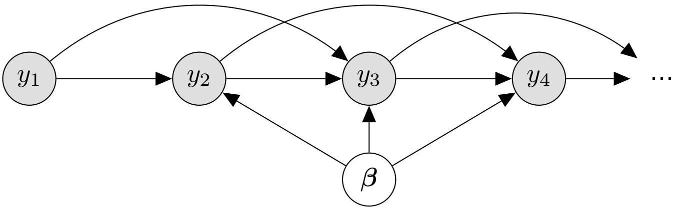
- Graphical model for (Bayesian) autoregressive model of order 2 (AR2)



- Generative process

  1. Draw transition coefficients $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}|\mathbf{0}, \lambda\mathbf{I})$
  2. Draw first observation $y_1 \sim \mathcal{N}(y_1|\mu_0, \sigma_0^2)$
  3. Draw second observation $y_2 \sim \mathcal{N}(y_2|\beta_1 y_{t-1}, \sigma^2)$
  4. For each time $t \in \{3, \ldots, T\}$
     a. Draw observation $y_t \sim \mathcal{N}(y_t|\beta_1 y_{t-1} + \beta_2 y_{t-2}, \sigma^2)$

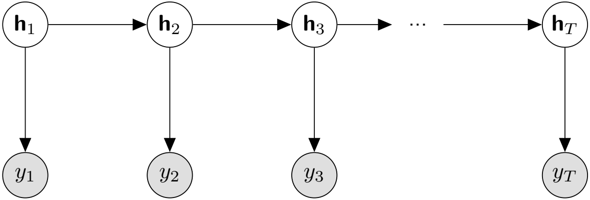
Autoregressive (AR) models: summary so far...



- In autoregressive models, each observation y_t is a linear function of the M previous observations $\{y_{t-1}, \ldots, y_{t-M}\}$ plus some noise
- But, if $\{y_{t-1}, \ldots, y_{t-M}\}$ are themselves noisy observations (e.g. due to measurement noise in sensors), aren't we building up noise over time?
- Perhaps there is a better way to model temporal data...

State-space models (SSMs)

- In AR models, only the observations $\{y_1, \ldots, y_T\}$ are modeled explicitly
- More general models of time-series introduce an extra set of unobserved (hidden) variables $\{\mathbf{h}_1, \ldots, \mathbf{h}_T\}$ from which the observations are assumed to be generated
- Typically the hidden variables are assumed to be (first order) Markovian
- y_t is assumed to be independent from all other variables given $\mathbf{h}_t$



- Notice that the latent **states** $\mathbf{h}_t$ can be vector-valued!
- Many physical processes can be expressed in this state-space framework

Linear dynamical systems (LDSs)

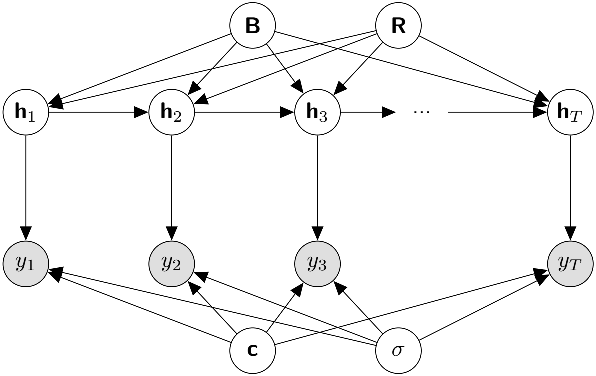
- State-space models require us to specify two probability distributions:
  - Transition probability: $p(\mathbf{h}_t|\mathbf{h}_{t-1})$
  - Observation (or emission) probability: $p(y_t|\mathbf{h}_t)$
- A popular choice is to assume **linear Gaussians**

$$\mathbf{h}_t \sim \mathcal{N}(\mathbf{h}_t|\mathbf{B}\mathbf{h}_{t-1}, \mathbf{R})$$
$$y_t \sim \mathcal{N}(y_t|\mathbf{c}^\mathsf{T}\mathbf{h}_t, \sigma^2)$$

- This choice is attractive for computational tractability
- This sub-class of SSMs are called *linear dynamical systems* (LDSs)
- **B**, **R**, **c** and σ are often treated as parameters rather random variables
  - Maximum likelihood estimation (MLE) or maximum-a-posteriori (MAP) rather than a fully Bayesian approach
  - Efficient (exact!) inference of latent states $\mathbf{h}_t$ (the Kalman filter)
- We shall consider a fully Bayesian approach
  - Assign priors to **B**, **R**, **c** and σ
  - Use Bayesian inference to compute respective posteriors

Linear dynamical systems (LDSs)

- Graphical model



- The goal of inference is to compute the posterior distribution over all the latent variables in the model: $\{\mathbf{h}_1, \ldots, \mathbf{h}_T\}$, $\mathbf{B}$, $\mathbf{R}$, $\mathbf{c}$ and σ
  - Can be very challenging (e.g. identifiability issues)!
  - We often need to impose some restrictions on the structure of $\mathbf{B}$, $\mathbf{R}$, $\mathbf{c}$.

AR models as linear dynamical systems (LDSs)

- By appropriately defining the latent states, many popular time-series models can be cast as an LDS
- E.g. an autoregressive model of order k (AR-k) can be written as:

$$\underbrace{\begin{pmatrix} h_t \\ h_{t-1} \\ \vdots \\ h_{t-k+1} \end{pmatrix}}_{\mathbf{h}_t} = \underbrace{\begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_k \\ 1 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}}_{\mathbf{B}} \underbrace{\begin{pmatrix} h_{t-1} \\ h_{t-2} \\ \vdots \\ h_{t-k} \end{pmatrix}}_{\mathbf{h}_{t-1}} + \begin{pmatrix} \tau_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

with the emission probability defined as

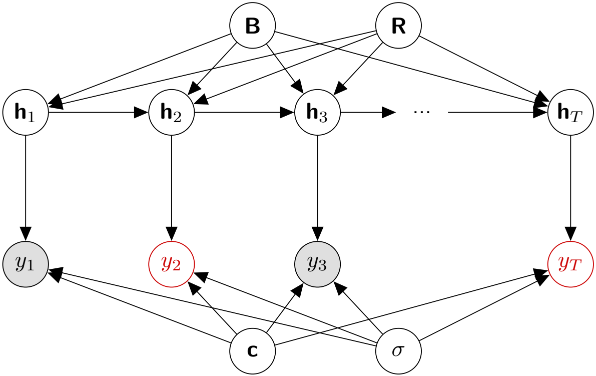
$$y_t = \underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}}_{\mathbf{c}} \mathbf{h}_t + \epsilon_t$$

where $\tau_t \sim \mathcal{N}(0, r)$ and $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$

- Notice the restrictions that we imposed on the structures of **B** and **c**

Missing data

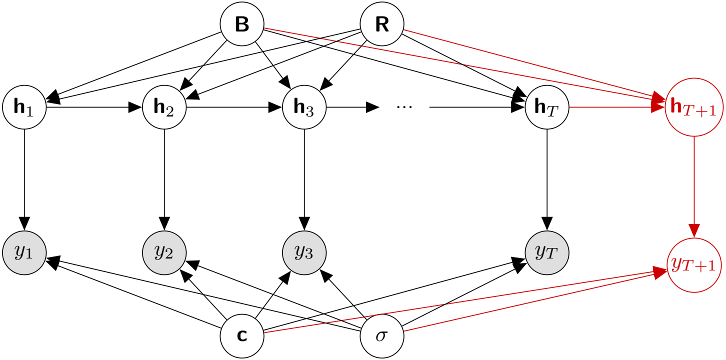
- What if we have **missing** observations?



- Just treat the missing observations as latent variables
  - Marginalize out their values
  - Compute their posterior distribution

Forecasting

- What if we want to predict the value of y_{T+1}?

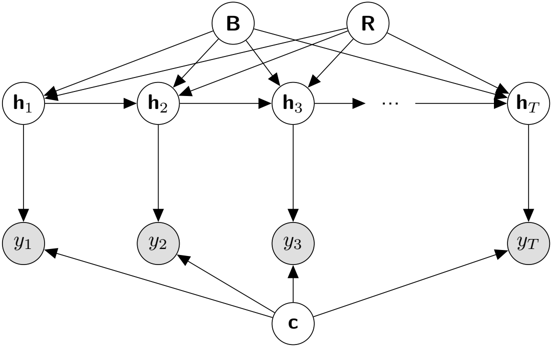


- **Extend** the model for one (or more) time-steps with latent variables
  - Compute the posterior distribution over y_{T+1}

Playtime!

- Ancestral sampling from Bayesian LDS model
  - See “07 - Temporal models - Part 1.ipynb" notebook
  - Expected duration: 10 minutes
- Bayesian LDS model of freeway occupancy rates: forecasting
  - See “07 - Temporal models - Part 2.ipynb" notebook
  - Expected duration: 45 minutes
- Bayesian LDS model of freeway occupancy rates: imputation
  - See “07 - Temporal models - Part 3.ipynb" notebook
  - Expected duration: 45 minutes

Extensions: Poisson LDSs



- Suppose we want to model a time-series of counts
  - E.g. forecasting the number of arrivals/departures at a metro station
- A **Poisson distribution** could be a better choice for the emission probabilities

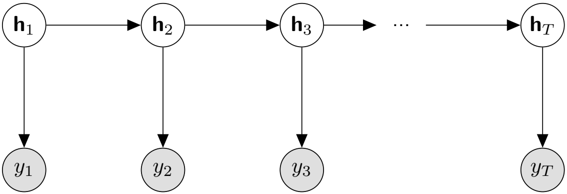
$$\mathbf{h}_t \sim \mathcal{N}(\mathbf{h}_t | \mathbf{B}\mathbf{h}_{t-1}, \mathbf{R})$$

$$y_t \sim \text{Poisson}(y_t | \exp(\mathbf{c}^\mathsf{T} \mathbf{h}_t))$$

- In fact, as with generalized linear models (GLMs), we could use **any** general exponential family for the response distribution

Extensions: Linear dynamical systems (LDSs) with inputs

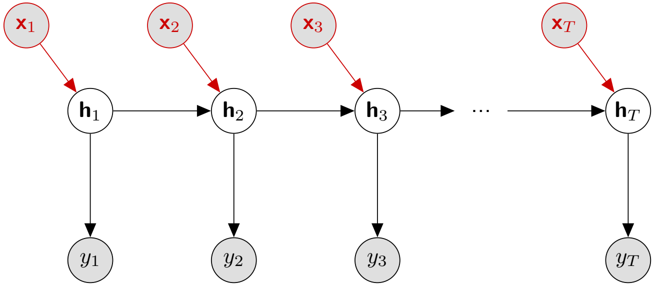
- Graphical model representation of a linear dynamical system:



(for simplicity, we omitted the variables **B**, **R**, **c** and σ from the graph)

- What if we have additional input information available at each time t that can help explain the observations y_t?
  - In our case study of freeway occupancy rates, this could be event information, weather, seasonal information, etc.

Extensions: Linear dynamical systems (LDSs) with inputs



(for simplicity, we omitted the variables **B**, **R**, **c**, σ and **W** from the graph)

- $\mathbf{x}_t$ are **external inputs** at time t
- Transition probability becomes:

$$\mathbf{h}_t \sim \mathcal{N}(\mathbf{h}_t | \mathbf{B}\mathbf{h}_{t-1} + \mathbf{W}\mathbf{x}_t, \mathbf{R})$$

Extensions: Extended Kalman filter (EKF)

- So far we have assumed linear relations between states and from state to observation

$$\mathbf{h}_t \sim \mathcal{N}(\mathbf{h}_t | \mathbf{B}\mathbf{h}_{t-1} + \mathbf{W}\mathbf{x}_t, \mathbf{R})$$
$$y_t \sim \mathcal{N}(y_t | \mathbf{c}^\mathsf{T}\mathbf{h}_t, \sigma^2)$$

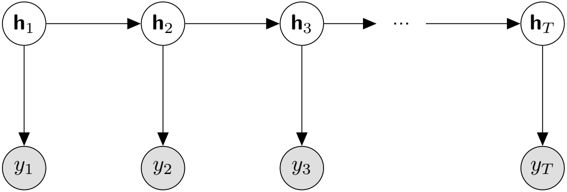
- The *extended Kalman filter* (EKF) relaxes this assumption by allowing these relations to be defined by arbitrary **differentiable functions** f and g

$$\mathbf{h}_t \sim \mathcal{N}(\mathbf{h}_t | f(\mathbf{h}_{t-1}, \mathbf{x}_t), \mathbf{R})$$
$$y_t \sim \mathcal{N}(y_t | g(\mathbf{h}_t), \sigma^2)$$

- f and g can even be neural networks! (search for "Deep K)
- EKF allows for a **non-linear** version of the Kalman filter
- Extremely popular in navigation systems and GPS
  (e.g. your phone uses it all the time!)
- Specialized algorithms have been developed for learning and performing inference in EKFs - not the focus of this course! We will keep relying on Stan

Extensions: Different regimes

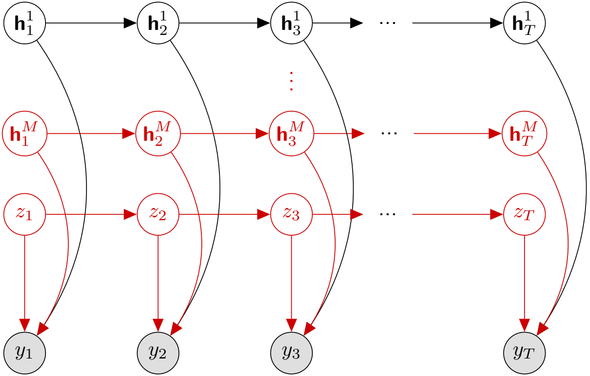
- Graphical model representation of a linear dynamical system:



(for simplicity, we omitted the variables $\mathbf{B}$, $\mathbf{R}$, $\mathbf{c}$ and σ from the graph)

- Suppose that the time-series data can behave quite differently in different periods (**regimes**)
  - E.g. do traffic conditions evolve the same way over time in congestion periods as they do in free flow?
- How can we change the model to incorporate this prior knowledge?

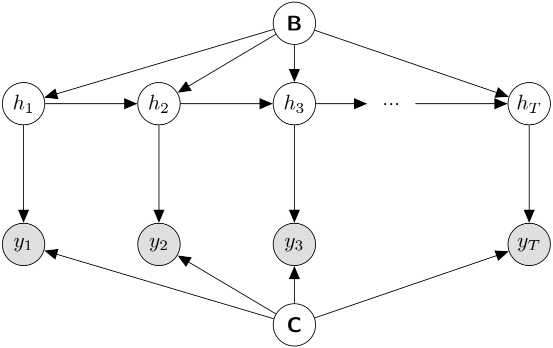
Extensions: Switching linear dynamical systems (SLDSs)



- Multiple latent state chains: $\{\mathbf{h}_1^1, \dots, \mathbf{h}_T^1\} \dots \{\mathbf{h}_1^M, \dots, \mathbf{h}_T^M\}$
- At each time t, a discrete latent variable z_t selects one of the latent state chains to be “active" and produce an observation y_t
- Useful if your observations can alternate between regimes with different dynamics
  - E.g. road segments with multiple traffic regimes (congested vs. free flow)

Hidden Markov models (HMMs)

- Discrete analog of the LDS is the *hidden Markov model* (HMM)



- Hidden states h_t and observations y_t are **discrete random variables**
- Transition probabilities $p(h_t|h_{t-1}, \mathbf{B})$ and emission probabilities $p(y_t|h_t, \mathbf{C})$ can be simply represented using conditional probability tables (CPTs)
    - Parameter matrices **B** and **C** are CPTs!

Hidden Markov models (HMMs)

- Suppose that the latent state h_t can take K possible values: $h_t \in \{1, \ldots, K\}$
- We can specify the transition probabilities using a $K \times K$ matrix **B**
    - Element b_{ij} of matrix **B** denotes the probability of transition to state j given that the current state is i
    - Given that the previous state h_{t-1} is i, the row $\mathbf{b}_i$ specifies the probabilities for transitioning to each of the possible K states at time t

- Suppose that the observations y_t can take V possible values: $y_t \in \{1, \ldots, V\}$
- We can specify the emission probabilities using a $K \times V$ matrix **C**
    - Element c_{ij} of matrix **C** denotes the probability of observing the value $y_t = j$ given that the current state h_t is i
    - Given that the current state h_t is i, the row $\mathbf{c}_i$ specifies the probabilities of observing each of the possible V values at time t ($y_t \in \{1, \ldots, V\}$)

- Notice that the rows $\mathbf{b}_i$ and $\mathbf{c}_i$ must sum to 1 in order to be valid probabilities!

Hidden Markov models (HMMs)

- In order to formally specify the HMM, we need to assign probability distributions to the variables in the model
- Lets start with the hidden states $h_t \in \{1, \dots, K\}$
  - What distribution do we know for discrete observations with K possible values? Multinomial!

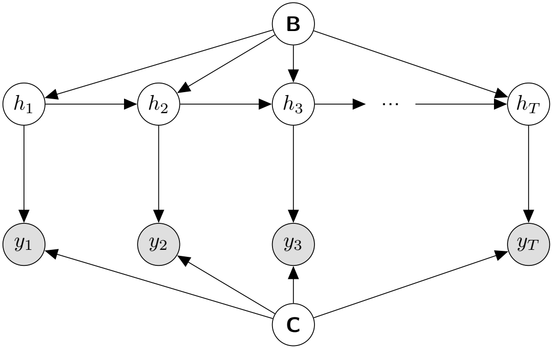
$$h_t \sim \text{Multinomial}(h_t | \mathbf{b}_{h_{t-1}})$$

  - Notice that the previous hidden state h_{t-1} is specifying at which row $\mathbf{b}_i$ of $\mathbf{B}$ the model should look at in order to decide the next hidden state h_t
- Similarly, for the observations $y_t \in \{1, \dots, V\}$ we have

$$y_t \sim \text{Multinomial}(y_t | \mathbf{c}_{h_t})$$

- Lastly, we need to assign priors to the parameters of these multinomials
- What is the conjugate prior for the multinomial distribution? Dirichlet!

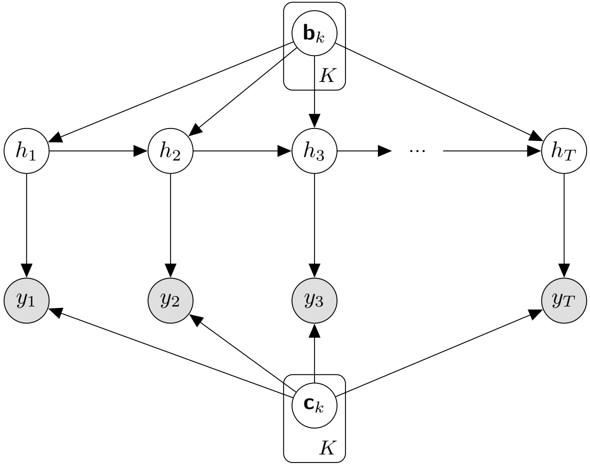
Hidden Markov models (HMMs)



- Putting everything together, the **generative process** becomes:
  1. For each possible hidden state value $k \in \{1, \ldots, K\}$
     a. Draw transition probabilities $\mathbf{b}_k \sim \text{Dirichlet}(\mathbf{b}_k|\boldsymbol{\alpha})$
     b. Draw emission probabilities $\mathbf{c}_k \sim \text{Dirichlet}(\mathbf{c}_k|\boldsymbol{\gamma})$
  2. For each time $t \in \{1, \ldots, T\}$
     a. Draw new hidden state $h_t \sim \text{Multinomial}(h_t|\mathbf{b}_{h_{t-1}})$
     b. Draw observation $y_t \sim \text{Multinomial}(y_t|\mathbf{c}_{h_t})$

Hidden Markov models (HMMs)

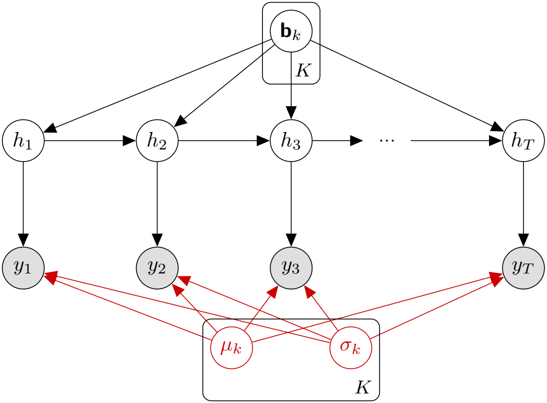
- Alternatively, we can represent the graphical model of the HMM as:



- Both ways are correct... This is one is more similar to the generative process

Extensions: HMMs with continuous observations

- We can **mix discrete** hidden states with **continuous** observations



- Hidden state value h_t defines from which of K Gaussians the observation y_t comes from

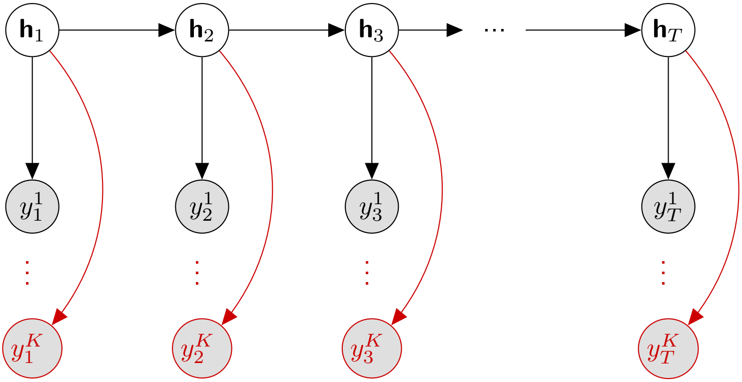
$$y_t \sim \mathcal{N}(y_t | \mu_{h_t}, \sigma_{h_t})$$

- Popular for modeling speech: y_t are sounds (continuous) and h_t are words

Multivariate state-space models

- What if, rather than a single time-series, we observed multiple time-series? (e.g. multiple sensors)
- Typically, multivariate time-series can exhibit **strong correlations** between the different component series
  - E.g., in our case study, occupancy rates at nearby segments of the freeway should be highly correlated
- Multivariate state-space models allow us to model these correlations

Multivariate state-space models



- We can group all the K observed outputs at time t in a single vector $\mathbf{y}_t = (y_t^1, \ldots, y_t^K)$
- Model becomes:

$$\mathbf{h}_t \sim \mathcal{N}(\mathbf{h}_t | \mathbf{B}\mathbf{h}_{t-1}, \mathbf{R})$$
$$\mathbf{y}_t \sim \mathcal{N}(\mathbf{y}_t | \mathbf{C}\mathbf{h}_t, \mathbf{\Sigma})$$

Multivariate state-space models

- A particularly important aspect is how we define the covariance matrices $\mathbf{R}$ and $\mathbf{\Sigma}$
- Consider a vector $\mathbf{y}$ of size N distributed as a multivariate Gaussian:

$$\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$$

  - An isotropic covariance of the form $\mathbf{\Sigma} = \sigma\mathbf{I}$ assumes a single variance for all elements of $\mathbf{y}$ (1 parameter: σ)
  - A diagonal covariance of the form $\mathbf{\Sigma} = \text{diag}(\boldsymbol{\sigma})$ assumes a different variance for each element of $\mathbf{y}$ (N parameters: $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)$)
  - A full covariance matrix $\mathbf{\Sigma}$ also models covariance between the different elements of $\mathbf{y}$ ($N \times N$ parameters! But there are ways to reduce this...)

Multivariate state-space models

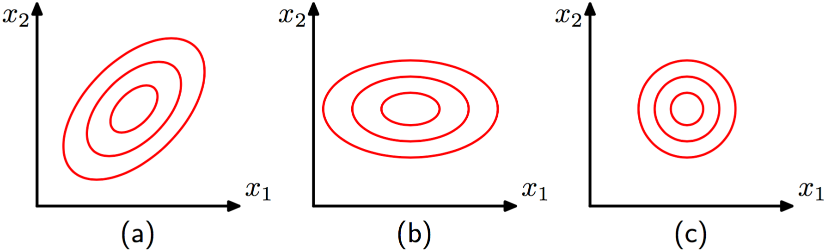


Figure: Bishop, C., 2006. Pattern Recognition and Machine Learning.

(a) A full covariance matrix $\mathbf{\Sigma}$ (models covariances!)

(b) A diagonal covariance of the form $\mathbf{\Sigma} = \text{diag}(\boldsymbol{\sigma})$

(c) An isotropic covariance of the form $\mathbf{\Sigma} = \sigma\mathbf{I}$

Playtime!

- **Multivariate** LDS model of freeway occupancy rates: imputation
- See “07 - Temporal models - Part 4.ipynb" notebook