# Lecture 12 and 13: Model Comparison

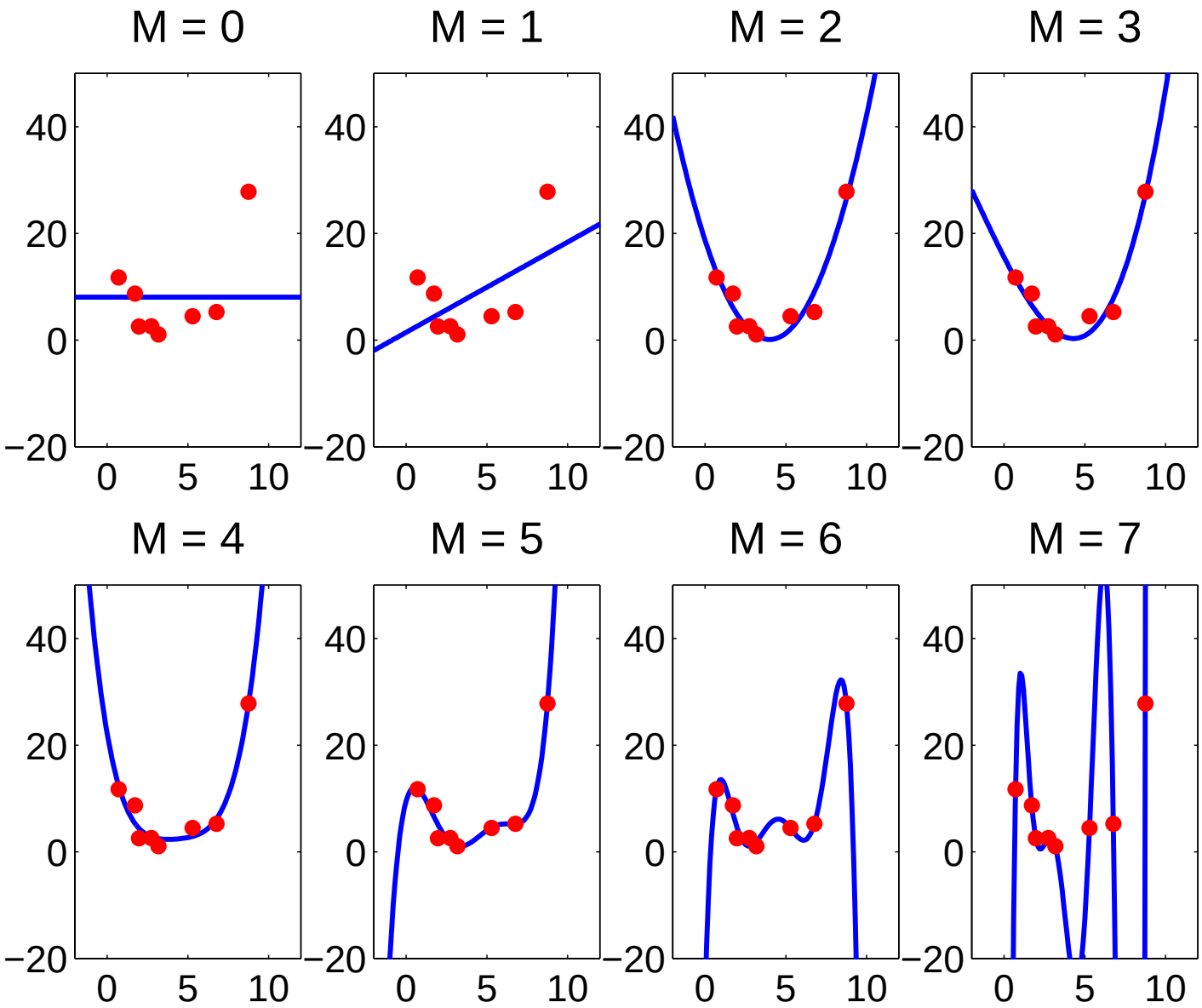
## 4F13: Machine Learning

Zoubin Ghahramani and Carl Edward Rasmussen

Department of Engineering, University of Cambridge

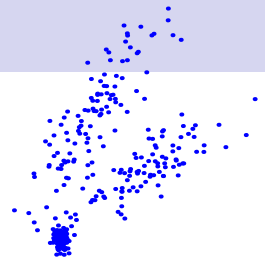
February 24th and 25th, 2010

# Model complexity and overfitting: a simple example

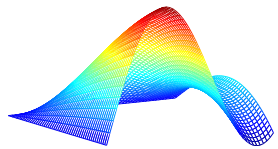


# Learning Model Structure

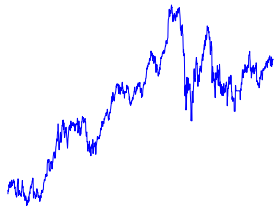
How many clusters in the data?



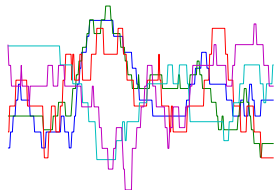
What is the intrinsic dimensionality of the data?



Is this input relevant to predicting that output?



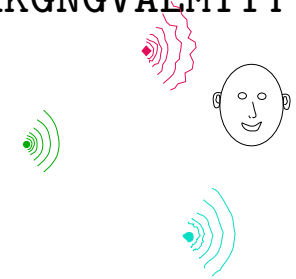
What is the order of a dynamical system?



How many states in a hidden Markov model?

SVYDAAAQLTADVKKDLRDSWKVIGSDKKGNGVALMTTY

How many auditory sources in the input?



# Using Occam's Razor to Learn Model Structure

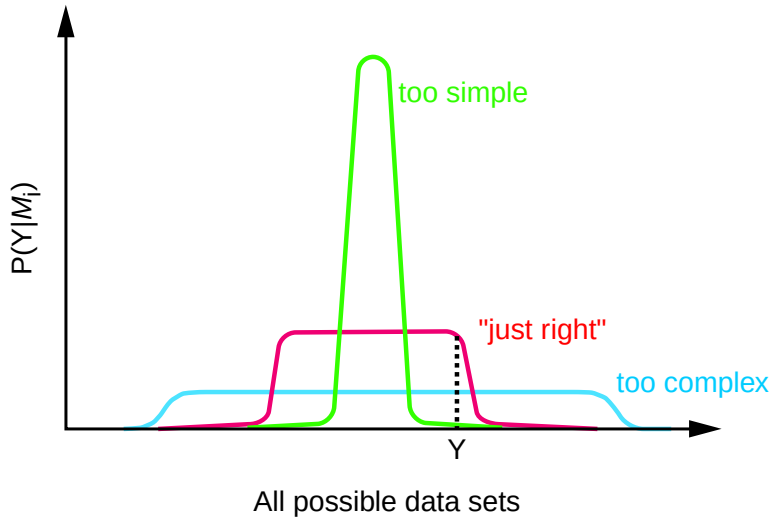
Compare model classes $m$ using their posterior probability given the data:

$$P(m|\mathbf{y}) = \frac{P(\mathbf{y}|m)P(m)}{P(\mathbf{y})}, \qquad P(\mathbf{y}|m) = \int_{\Theta_m} P(\mathbf{y}|\theta_m, m)P(\theta_m|m)\, d\theta_m$$

**Interpretation of $P(\mathbf{y}|m)$:** The probability that *randomly selected* parameter values from the model class would generate data set $\mathbf{y}$.

Model classes that are too simple are unlikely to generate the data set.

Model classes that are too complex can generate many possible data sets, so again, they are unlikely to generate that particular data set at random.



# Bayesian Model Comparison: Terminology

- A model class $m$ is a set of models parameterised by $\theta_m$, e.g. the set of all possible mixtures of $m$ Gaussians.
- The marginal likelihood of model class $m$:

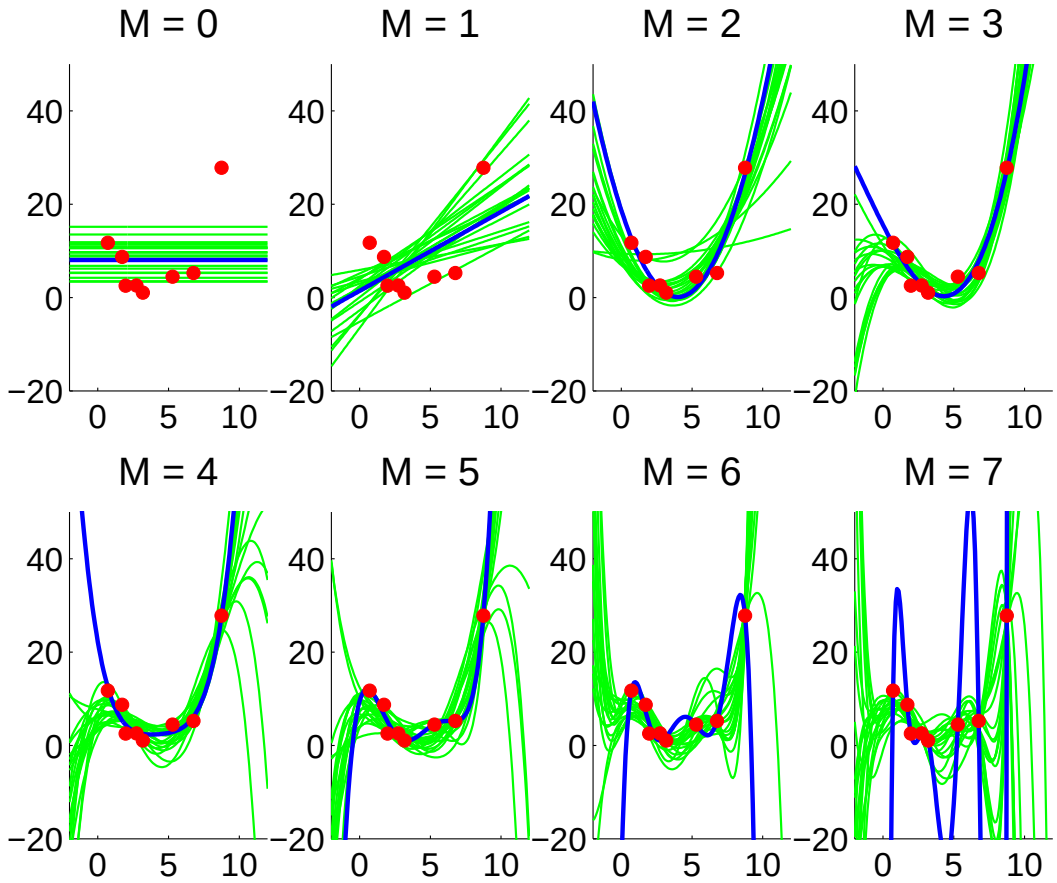
$$P(\mathbf{y}|m) = \int_{\Theta_m} P(\mathbf{y}|\theta_m, m) P(\theta_m|m) \, d\theta_m$$

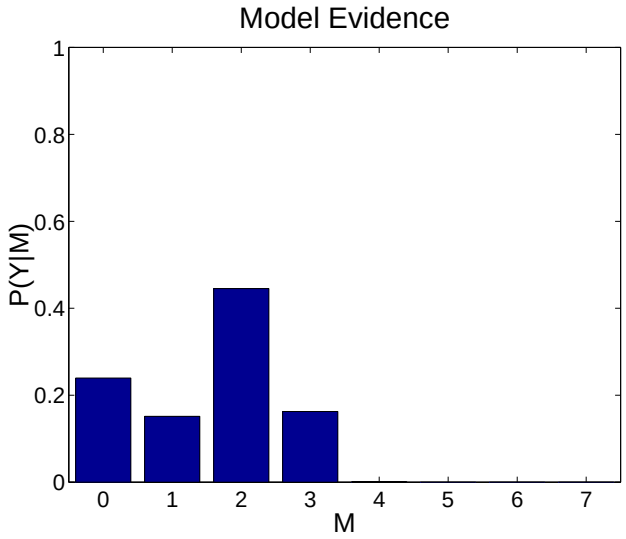
  is also known as the Bayesian evidence for model $m$.
- The ratio of two marginal likelihoods is known as the Bayes factor:

$$\frac{P(\mathbf{y}|m)}{P(\mathbf{y}|m')}$$

- The Occam's Razor principle is, roughly speaking, that one should prefer simpler explanations than more complex explanations.
- Bayesian inference formalises and *automatically* implements the Occam's Razor principle.

# Bayesian Model Comparison: Occam's Razor at Work




e.g. for quadratic (M=2): $y = a_0 + a_1 x + a_2 x^2 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \tau)$ and $\theta_2 = [a_0\ a_1\ a_2\ \tau]$

`demo: polybayes`

# Practical Bayesian approaches

- Laplace approximations:
  - Makes a Gaussian approximation about the maximum *a posteriori* parameter estimate.
- Bayesian Information Criterion (BIC)
  - an asymptotic approximation.
- Markov chain Monte Carlo methods (MCMC):
  - In the limit are guaranteed to converge, but:
  - Many samples required to ensure accuracy.
  - Sometimes hard to assess convergence.
- Variational approximations

Note: other deterministic approximations have been developed more recently and can be applied in this context: e.g. Bethe approximations and Expectation Propagation.

# Laplace Approximation

data set: $\mathbf{y}$ models: $m = 1 \ldots, M$ parameter sets: $\theta_1 \ldots, \theta_M$

Model Comparison: $P(m|\mathbf{y}) \propto P(m)P(\mathbf{y}|m)$

For large amounts of data (relative to number of parameters, d) the parameter posterior is approximately Gaussian around the MAP estimate $\hat{\theta}_m$:

$$P(\theta_m|\mathbf{y}, m) \simeq (2\pi)^{-d/2}|A|^{1/2} \exp\left(-\frac{1}{2}(\theta_m - \hat{\theta}_m)^\top A(\theta_m - \hat{\theta}_m)\right)$$

$$P(\mathbf{y}|m) = \frac{P(\theta_m, \mathbf{y}|m)}{P(\theta_m|\mathbf{y}, m)}$$

Evaluating the above for $\log P(\mathbf{y}|m)$ at $\hat{\theta}_m$ we get the Laplace approximation:

$$\log P(\mathbf{y}|m) \simeq \log P(\hat{\theta}_m|m) + \log P(\mathbf{y}|\hat{\theta}_m, m) + \frac{d}{2}\log 2\pi - \frac{1}{2}\log|A|$$

$-A$ is the $d \times d$ Hessian matrix of $\log P(\theta_m|\mathbf{y}, m)$:
$A_{k\ell} = -\frac{\partial^2}{\partial\theta_{mk}\partial\theta_{m\ell}} \log P(\theta_m|\mathbf{y}, m)|_{\hat{\theta}_m}$.

Can also be derived from $2^{nd}$ order Taylor expansion of log posterior.

The Laplace approximation can be used for model comparison.

# Bayesian Information Criterion (BIC)

BIC can be obtained from the Laplace approximation:

$$\log P(\mathbf{y}|m) \simeq \log P(\hat{\theta}_m|m) + \log P(\mathbf{y}|\hat{\theta}_m, m) + \frac{d}{2}\log 2\pi - \frac{1}{2}\log|A|$$

in the large sample limit ($N \to \infty$) where N is the number of data points, $A$ grows as $NA_0$ for some fixed matrix $A_0$, so $\log|A| \to \log|NA_0| = \log(N^d|A_0|) = d\log N + \log|A_0|$. Retaining only terms that grow in N we get:

$$\log P(\mathbf{y}|m) \simeq \log P(\mathbf{y}|\hat{\theta}_m, m) - \frac{d}{2}\log N$$

Properties:

- Quick and easy to compute, and does not depend on the prior
- We can use the ML estimate of $\theta$ instead of the MAP estimate
- It assumes that in the large sample limit, all the parameters are well-determined (i.e. the model is identifiable; otherwise, d should be the number of well-determined parameters)
- **Danger:** counting parameters can be deceiving! (c.f. sinusoid)
- It is equivalent to the "Minimum Description Length" (MDL) criterion

# Sampling Approximations

Let's consider a non-Markov chain method, **Importance Sampling:**

$$\begin{aligned}
\log P(\mathbf{y}|m) &= \log \int_{\Theta_m} P(\mathbf{y}|\theta_m, m) P(\theta_m|m)\, d\theta_m \\
&= \log \int_{\Theta_m} P(\mathbf{y}|\theta_m, m) \frac{P(\theta_m|m)}{Q(\theta_m)} Q(\theta_m)\, d\theta_m \\
&\simeq \log \frac{1}{K} \sum_k P(\mathbf{y}|\theta_m^{(k)}, m) \frac{P(\theta_m^{(k)}|m)}{Q(\theta_m^{(k)})}
\end{aligned}$$

where $\theta_m^{(k)}$ are i.i.d. draws from $Q(\theta_m)$. Assumes we can **sample from** and **evaluate** $Q(\theta_m)$ (incl. normalization!) and we can **compute the likelihood** $P(\mathbf{y}|\theta_m^{(k)}, m)$.

Although importance sampling does not work well in high dimensions, it inspires the following approach: Create a **Markov chain,** $Q_k \to Q_{k+1} \ldots$ for which:

- $Q_k(\theta)$ can be evaluated including normalization
- $\lim_{k\to\infty} Q_k(\theta) = P(\theta|\mathbf{y}, m)$

# Variational Bayesian Learning

Lower Bounding the Marginal Likelihood

Let the hidden latent variables be $\mathbf{x}$, data $\mathbf{y}$ and the parameters $\theta$.

Lower bound the marginal likelihood (Bayesian model evidence) using Jensen's inequality:

$$\begin{aligned}
\log P(\mathbf{y}) &= \log \int d\mathbf{x}\, d\theta\; P(\mathbf{y}, \mathbf{x}, \theta) \qquad |m \\
&= \log \int d\mathbf{x}\, d\theta\; Q(\mathbf{x}, \theta) \frac{P(\mathbf{y}, \mathbf{x}, \theta)}{Q(\mathbf{x}, \theta)} \\
&\geqslant \int d\mathbf{x}\, d\theta\; Q(\mathbf{x}, \theta) \log \frac{P(\mathbf{y}, \mathbf{x}, \theta)}{Q(\mathbf{x}, \theta)}.
\end{aligned}$$

Use a simpler, factorised approximation to $Q(\mathbf{x}, \theta)$:

$$\begin{aligned}
\log P(\mathbf{y}) &\geqslant \int d\mathbf{x}\, d\theta\; Q_{\mathbf{x}}(\mathbf{x}) Q_{\theta}(\theta) \log \frac{P(\mathbf{y}, \mathbf{x}, \theta)}{Q_{\mathbf{x}}(\mathbf{x}) Q_{\theta}(\theta)} \\
&= \mathcal{F}(Q_{\mathbf{x}}(\mathbf{x}), Q_{\theta}(\theta), \mathbf{y}).
\end{aligned}$$

Maximize this lower bound.

# Variational Bayesian Learning . . .

Maximizing this lower bound, $\mathcal{F}$, leads to **EM-like** updates:

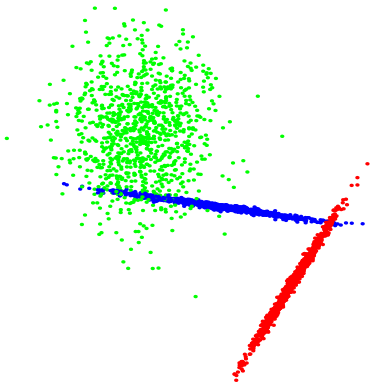
$$\begin{aligned} Q_x^*(x) &\propto \exp\langle \log P(x,y|\theta)\rangle_{Q_\theta(\theta)} && E-like\ step \\ Q_\theta^*(\theta) &\propto P(\theta)\exp\langle \log P(x,y|\theta)\rangle_{Q_x(x)} && M-like\ step \end{aligned}$$

Maximizing $\mathcal{F}$ is equivalent to minimizing KL-divergence between the *approximate posterior*, $Q(\theta)Q(x)$ and the *true posterior*, $P(\theta, x|y)$.

$$\begin{aligned} \log P(y) - \mathcal{F}(Q_x(x), Q_\theta(\theta), y) &= \\ \log P(y) - \int dx\, d\theta\; Q_x(x)Q_\theta(\theta)\log\frac{P(y,x,\theta)}{Q_x(x)Q_\theta(\theta)} &= \\ \int dx\, d\theta\; Q_x(x)Q_\theta(\theta)\log\frac{Q_x(x)Q_\theta(\theta)}{P(x,\theta|y)} &= KL(Q\|P) \end{aligned}$$

# Mixture of Factor Analysers

Goal: find the *number of clusters* and *their dimesions*.



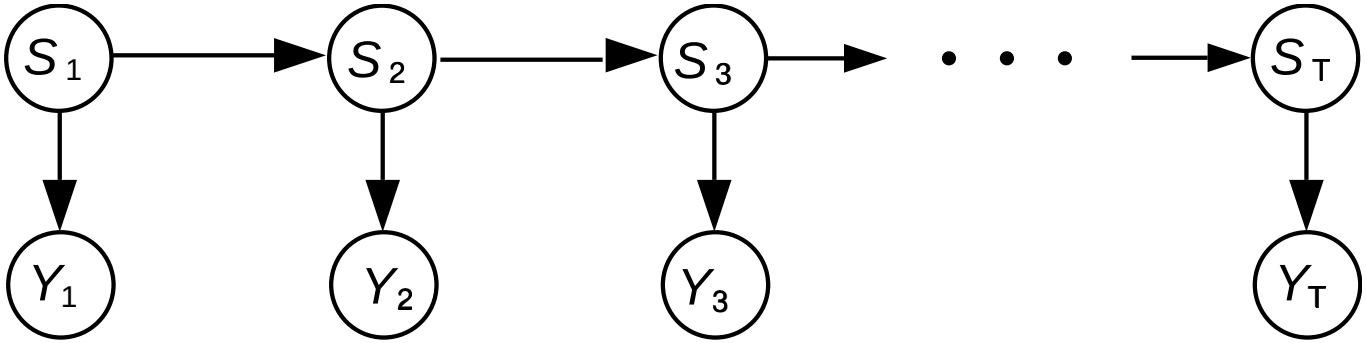
| number of points per cluster | intrinsic dimensionalities | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 7 | 4 | 3 | 2 | 2 |
| 8 | 2 | | | | 1 | |
| 8 | 1 | 2 | | | | |
| 16 | 1 | 4 | | | | 2 |
| 32 | 1 | 6 | 3 | 3 | 2 | 2 |
| 64 | 1 | 7 | 4 | 3 | 2 | 2 |
| 128 | 1 | 7 | 4 | 3 | 2 | 2 |

True data: 6 Gaussian clusters with dimensions: (1 7 4 3 2 2) embedded in 10-D

- Finds the clusters and dimensionalities efficiently.
- The model complexity reduces in line with the lack of data support.

`demos:` `embed_demo`, `run_simple` and `ueda_demo`

# Hidden Markov Models

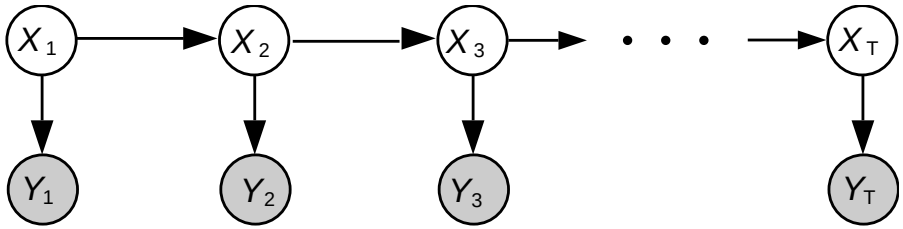



Discrete hidden states, $\mathbf{s}_t$.
Observations $\mathbf{y}_t$.

How many hidden states?
What structure state-transition matrix?

`demo: vbhmm_demo`

# Linear Dynamical Systems




- Assumes $\mathbf{y}_t$ generated from a sequence of Markov *hidden* state variables $\mathbf{x}_t$
- If transition and output functions are linear, time-invariant, and noise distributions are Gaussian, this is a linear-Gaussian state-space model:

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{w}_t, \qquad \mathbf{y}_t = C\mathbf{x}_t + \mathbf{v}_t$$

- Three levels of inference:
  - I Given data, structure and parameters, **Kalman smoothing** $\rightarrow$ hidden state;
  - II Given data and structure, **EM** $\rightarrow$ hidden state and parameter point estimates;
  - III Given data only, **VEM** $\rightarrow$ **model structure** *and* **distributions over parameters** *and* **hidden state.**

# Summary & Conclusions

- Bayesian learning avoids overfitting and can be used to do model comparison / selection.
- But we need approximations:
  - Laplace
  - BIC
  - Sampling
  - Variational