CS281A/Stat241A Lecture 17

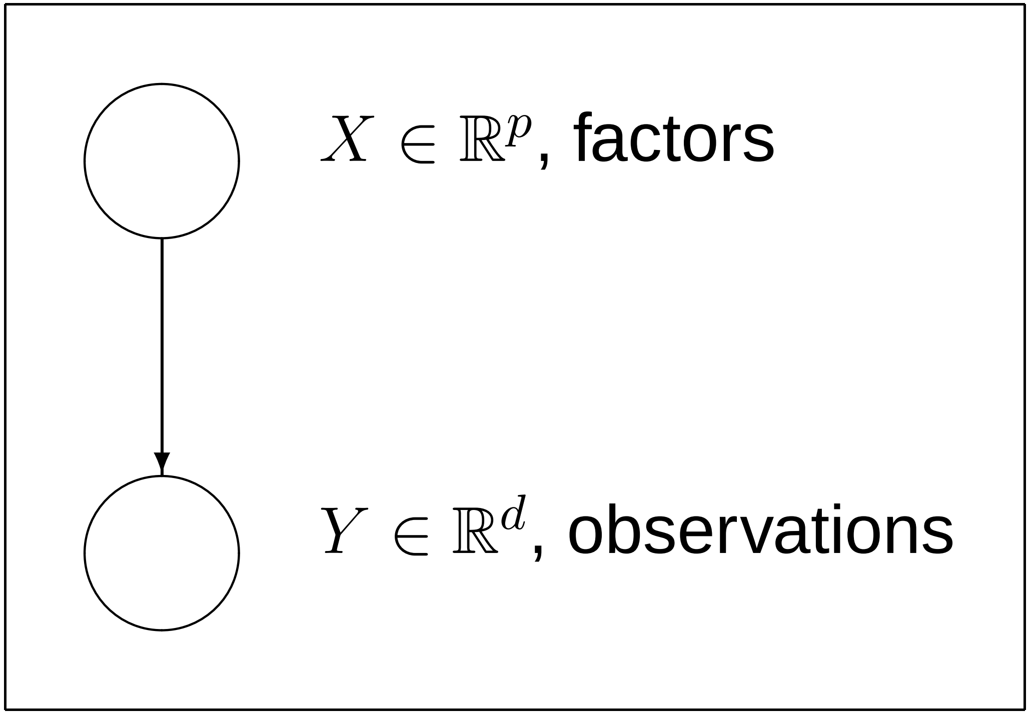
Factor Analysis and State Space Models

Peter Bartlett

Key ideas of this lecture

- Factor Analysis.
  - Recall: Gaussian factors plus observations.
  - Parameter estimation with EM.
- The vec operator.
  - Motivation: Natural parameters of Gaussian.
  - Linear functions of matrices as inner products. Kronecker product.
- State Space Models.
  - Linear dynamical systems, gaussian disturbances.
  - All distributions are Gaussian: parameters suffice.

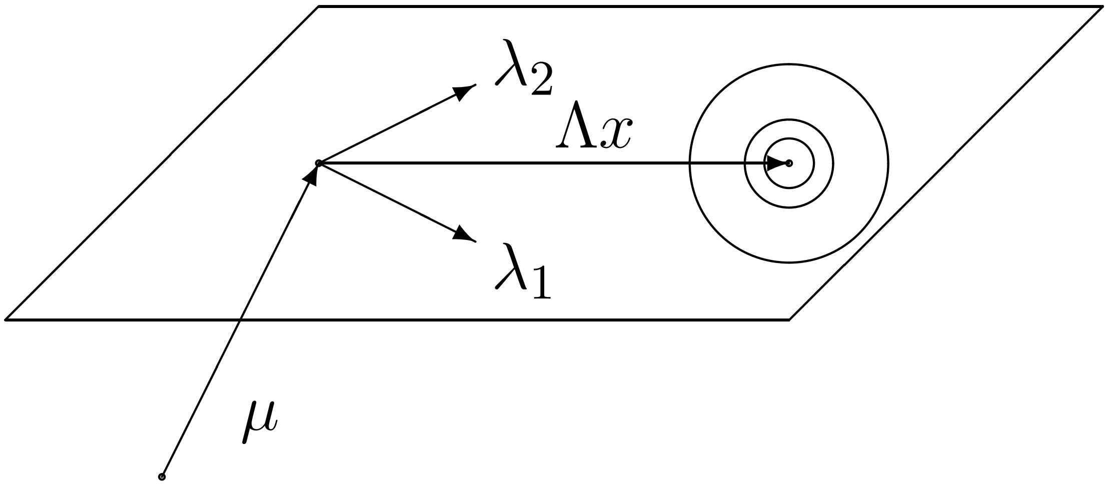
Factor Analysis: Definition



Local conditionals:

$$
\begin{aligned}
p(x) &= \mathcal{N}(x|0, I), \\
p(y|x) &= \mathcal{N}(y|\mu + \Lambda x, \Psi).
\end{aligned}
$$

Factor Analysis



Factor Analysis: Definition

Local conditionals:

$$p(x) = \mathcal{N}(x|0, I),$$
$$p(y|x) = \mathcal{N}(y|\mu + \Lambda x, \Psi).$$

- The mean of y is $\mu \in \mathbb{R}^d$.
- The matrix of factors is $\Lambda \in \mathbb{R}^{d \times p}$.
- The noise covariance $\Psi \in \mathbb{R}^{d \times d}$ is diagonal.
- Thus, there are $d + dp + d \sim dp \ll d^2$ parameters.

Factor Analysis: Marginals, Conditionals

Theorem

1. $Y \sim \mathcal{N}(\mu, \Lambda\Lambda' + \Psi)$.
2. $(X, Y) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \Sigma\right)$, with $\Sigma = \begin{pmatrix} I & \Lambda' \\ \Lambda & \Lambda\Lambda' + \Psi \end{pmatrix}$.
3. $p(x|y)$ is Gaussian, with
   mean $= \Lambda'(\Lambda\Lambda' + \Psi)^{-1}(y - \mu)$,
   covariance $I - \Lambda' \left(\Lambda\Lambda' + \Psi\right)^{-1} \Lambda$.

Factor Analysis: Parameter Estimation

- iid data $y = (y_1, \ldots, y_n)$.
- The log likelihood is

$$\begin{aligned}
\ell(\theta; y) &= \log p(y|\theta) \\
&= \text{const} - \frac{n}{2} \log \left| \Lambda\Lambda' + \Psi \right| \\
&\quad - \frac{1}{2} \sum_i (y_i - \mu)'(\Lambda\Lambda' + \Psi)^{-1}(y_i - \mu).
\end{aligned}$$

Factor Analysis: Parameter Estimation

Let's first consider estimation of μ:

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

as for the full covariance case.
From now on, let's assume $\mu = 0$, so we can ignore it:

$$\ell(\theta; y) = \text{const} - \frac{1}{2}\log\left|\Lambda\Lambda' + \Psi\right| - \frac{1}{2}\sum_i y_i'(\Lambda\Lambda' + \Psi)^{-1} y_i.$$

But how do we find a factorized covariance matrix,
$\Sigma = \Lambda\Lambda' + \Psi$?

Factor Analysis: EM

We follow the usual EM recipe:

1. Write out the complete log likelihood, ℓ_c.
2. **E step:** Calculate $\mathbb{E}[\ell_c|y]$.
   Typically, find $\mathbb{E}[\text{suff. stats}|y]$.
3. **M step:** Maximize $\mathbb{E}[\ell_c|y]$.

Factor Analysis: EM

1. Write out the complete log likelihood, ℓ_c.

$$\ell_c(\theta) = \log(p(x, \theta)p(y|x, \theta))$$
$$= \text{const} - \frac{n}{2}\log|\Psi| - \frac{1}{2}\sum_{i=1}^{n} x_i' x_i$$
$$- \frac{1}{2}\sum_{i=1}^{n}(y_i - \Lambda x_i)'\Psi^{-1}(y_i - \Lambda x_i).$$

Factor Analysis: EM

2. E step: Calculate $\mathbb{E}[\ell_c|y]$.
Typically, find $\mathbb{E}[\text{suff. stats}|y]$.

Claim: Sufficient statistics are x_i, $x_i x_i'$.

Indeed, $\mathbb{E}[\ell_c|y]$ is a constant plus $-\frac{n}{2}\log|\Psi|$ plus

$$
\begin{aligned}
&-\frac{1}{2}\mathbb{E}\left(\sum_{i=1}^{n} x_i' x_i + \sum_{i=1}^{n}(y_i - \Lambda x_i)'\Psi^{-1}(y_i - \Lambda x_i)|y\right) \\
&= -\frac{1}{2}\sum_{i=1}^{n}\left(\mathbb{E}(x_i' x_i|y_i) + \mathbb{E}\left(\text{tr}\left((y_i - \Lambda x_i)'\Psi^{-1}(y_i - \Lambda x_i)\right)|y_i\right)\right) \\
&= -\frac{1}{2}\sum_{i=1}^{n}\mathbb{E}(x_i' x_i|y_i) - \frac{n}{2}\mathbb{E}\left(\text{tr}\left(S\Psi^{-1}\right)|y_i\right),
\end{aligned}
$$

Factor Analysis: EM. E-step

$$\mathbb{E}[\ell_c|y] = -\frac{n}{2}\log|\Psi| - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}(x_i'x_i|y_i) - \frac{n}{2}\mathbb{E}\left(\mathrm{tr}\left(S\Psi^{-1}\right)|y_i\right),$$

with $S = \frac{1}{n}\sum_{i=1}^{n}(y_i - \Lambda x_i)(y_i - \Lambda x_i)'$,

We used the fact that the trace (sum of diagonal elements) of a matrix satisfies

$$\mathrm{tr}(ABC) = \mathrm{tr}(CAB),$$

as long as the products ABC and CAB are square.

Factor Analysis: EM. E-step

Now, we can calculate

$$\begin{aligned}\mathbb{E}(S|y) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[y_i y_i' - 2\Lambda x_i y_i' + \Lambda x_i x_i' \Lambda' \middle| y_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} y_i y_i' - 2\Lambda\mathbb{E}[x_i|y_i]y_i' + \Lambda\mathbb{E}[x_i x_i'|y_i]\Lambda',\end{aligned}$$

and from this it is clear that the expected sufficient statistics are $\mathbb{E}(x_i|y_i)$, $\mathbb{E}(x_i x_i'|y_i)$ (and its trace, $\mathbb{E}(x_i' x_i|y_i)$).

Factor Analysis: EM. E-step

We calculated these conditional expectations earlier:

$$\mathbb{E}[x_i|y_i] = \Lambda' \left(\Lambda\Lambda' + \Psi\right)^{-1} (y_i - \mu)$$

$$\text{Var}[x_i|y_i] = I - \Lambda' \left(\Lambda\Lambda' + \Psi\right)^{-1} \Lambda$$

$$\mathbb{E}[x_i x_i'|y_i] = \text{Var}[x_i|y_i] + \mathbb{E}[x_i|y_i]\mathbb{E}[x_i'|y_i].$$

So we can plug them in to calculate the expected complete log likelihood.

Factor Analysis: EM. M-step

3. M step: Maximize $\mathbb{E}[\ell_c|y]$.
For Λ**,** this is equivalent to minimizing

$$n\text{tr}\left(\mathbb{E}(S|y)\Psi^{-1}\right) = \text{tr}\left((Y' - \Lambda X')(Y' - \Lambda X')'\Psi^{-1}\right),$$

where $Y \in \mathbb{R}^{n\times d}$, with rows y_i, $X \in \mathbb{R}^{n\times p}$, with rows x_i. This is a matrix version of linear regression, with the d separate components of the squared error weighted by one of the diagonal entries in Ψ^{-1}. Thus, the Ψ matrix plays no role, and the solution satisfies the normal equations.

Factor Analysis: EM. M-step

Normal Equations:

$$\hat{\Lambda}' = \left(\sum_{i=1}^{n} \mathbb{E}[x_i x_i' | y_i] \right)^{-1} \sum_{i=1}^{n} \left(\mathbb{E}[x_i | y_i] y_i' \right)$$

(They are the same sufficient statistics as in linear regression; here we need to compute the expectation of the suff. stats given the observations.)

Factor Analysis: EM. M-step

For Ψ**,** we need to find a diagonal Ψ to minimize

$$\log|\Psi| + \text{tr}\left(\mathbb{E}(S|y)\Psi^{-1}\right) .= \sum_{j=1}^{d} \left(\log \psi_j + s_j/\psi_j\right),$$

It's easy to check that this is minimized for $\psi_j = s_j$, the diagonal entries of $\mathbb{E}[S|y]$.

Factor Analysis: EM. Summary.

E step: Calculate the expected suff. stats:

$$\mathbb{E}[x_i|y_i] = \Lambda' \left(\Lambda\Lambda' + \Psi\right)^{-1} (y_i - \mu)$$

$$\text{Var}[x_i|y_i] = I - \Lambda' \left(\Lambda\Lambda' + \Psi\right)^{-1} \Lambda$$

$$\mathbb{E}[x_i x_i'|y_i] = \text{Var}[x_i|y_i] + \mathbb{E}[x_i|y_i]\mathbb{E}[x_i'|y_i].$$

And use these to compute the (diagonal entries of the) matrix

$$\mathbb{E}(S|y) = \frac{1}{n}\sum_{i=1}^{n} y_i y_i' - 2\Lambda\mathbb{E}[x_i|y_i]y_i' + \Lambda\mathbb{E}[x_i x_i'|y_i]\Lambda'.$$

Factor Analysis: EM. Summary.

M step: Maximize $\mathbb{E}[\ell_c|y]$:

$$\hat{\Lambda}' = \left(\sum_{i=1}^{n} \mathbb{E}[x_i x_i'|y_i]\right)^{-1} \sum_{i=1}^{n} \left(\mathbb{E}[x_i|y_i]y_i'\right)$$

$$\hat{\Psi} = \text{diag}(\mathbb{E}(S|y)),$$

and recall: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i.$

Key ideas of this lecture

- Factor Analysis.
  - Recall: Gaussian factors plus observations.
  - Parameter estimation with EM.
- The vec operator.
  - Motivation: Natural parameters of Gaussian.
  - Linear functions of matrices as inner products. Kronecker product.
- State Space Models.
  - Linear dynamical systems, gaussian disturbances.
  - All distributions are Gaussian: parameters suffice.
  - Inference: Kalman filter and smoother.
  - Parameter estimation with EM.

The vec operator: Motivation

Consider the multivariate Gaussian:

$$p(x) = (2\pi)^{-(d)/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu) \right).$$

What is the natural parameterization?

$$p(x) = h(x) \exp\left(\theta' T(x) - A(\theta) \right).$$

The vec operator: Motivation

If we define

$$\Lambda = \Sigma^{-1} \qquad \eta = \Sigma^{-1}\mu,$$

then we can write

$$\begin{aligned}(x-\mu)'\Sigma^{-1}(x-\mu) &= \mu'\Sigma^{-1}\mu - 2\mu'\Sigma^{-1}x + x'\Sigma^{-1}x \\ &= \eta'\Lambda^{-1}\eta - 2\eta'x + x'\Lambda x.\end{aligned}$$

Why is this of the form $\theta' T(x) - A(\theta)$?

The vec operator: Example

$$\begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$= \lambda_{11} x_1^2 + \lambda_{21} x_1 x_2 + \lambda_{12} x_1 x_2 + \lambda_{22} x_2^2$$

$$= \begin{pmatrix} \lambda_{11} & \lambda_{21} & \lambda_{12} & \lambda_{22} \end{pmatrix} \begin{pmatrix} x_1^2 \\ x_1 x_2 \\ x_2 x_1 \\ x_2^2 \end{pmatrix}$$

$$= \mathsf{vec}(\Lambda)' \mathsf{vec}(xx').$$

The vec operator

Definition [vec]: For a matrix A,

$$\text{vec}(A) = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

where the a_i are the column vectors of A:

$$A = \begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix}$$

The vec operator

Theorem:

$$\text{tr}(A'B) = \text{vec}(A)'\text{vec}(B).$$

(Trace of the product is the sum of the corresponding row$\times$column inner products.)
In the example above,

$$\begin{aligned} x'\Lambda x &= \text{tr}\left(x'\Lambda x\right) \\ &= \text{tr}\left(\Lambda x x'\right) \\ &= \underbrace{\text{vec}(\Lambda')'}_{\text{nat. param.}} \underbrace{\text{vec}(xx')}_{\text{suff. stat.}} \end{aligned}$$

The Kronecker product

We also use the vec operator for matrix equations like

$$X\Theta Y = Z,$$

or, for instance, for least squares minimization of $X\Theta Y - Z$. Then we can write

$$\text{vec}(Z) = \text{vec}(X\Theta Y) = (Y' \otimes X)\text{vec}(\Theta),$$

where $(Y' \otimes X)$ is the **Kronecker product** of Y' and X:

The Kronecker product

Definition [Kronecker product] The Kronecker product of A and B is

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}$$

The Kronecker product

Theorem [Kronecker product and vec operator]

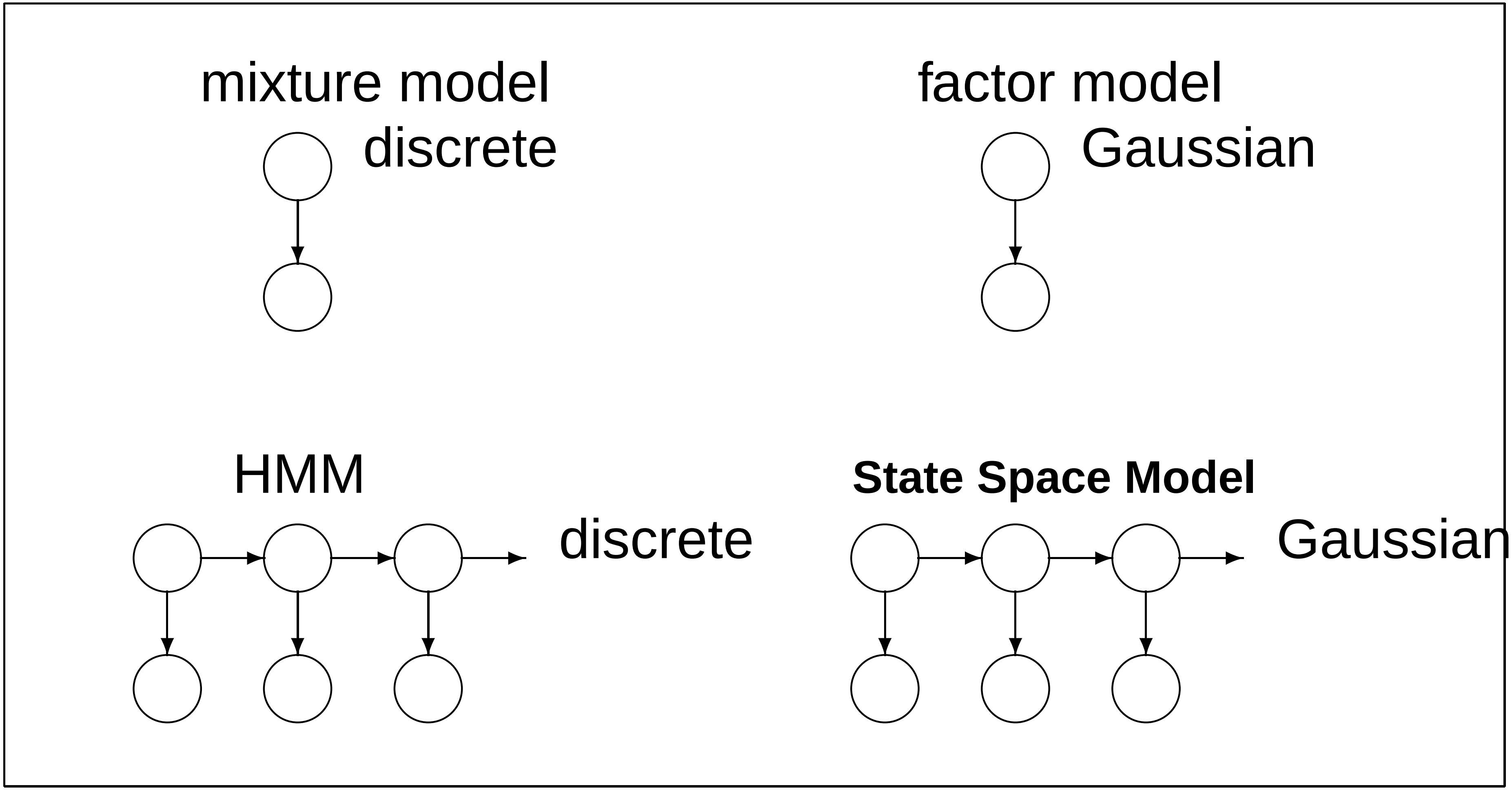
$$\text{vec}(ABC) = (C' \otimes A)\text{vec}(B).$$

The Kronecker product and vec operator are used in matrix algebra. They are convenient for differentiation of a function of a matrix, and they arise: in statistical models involving products of features; in systems theory; and in stability theory.

Key ideas of this lecture

- Factor Analysis.
  - Recall: Gaussian factors plus observations.
  - Parameter estimation with EM.
- The vec operator.
  - Motivation: Natural parameters of Gaussian.
  - Linear functions of matrices as inner products. Kronecker product.
- State Space Models.
  - Linear dynamical systems, gaussian disturbances.
  - All distributions are Gaussian: parameters suffice.
  - Inference: Kalman filter and smoother.
  - Parameter estimation with EM.

State Space Models



State Space Models

In **linear dynamic systems**,

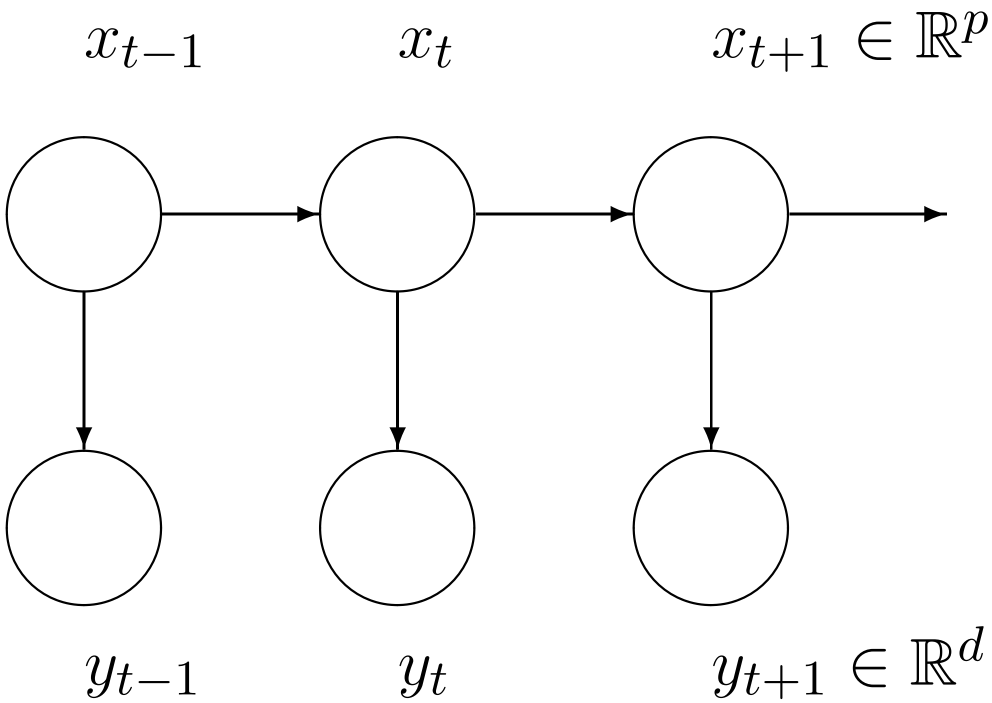
- The evolution of the state x_t, and
- The relationship between the state x_t and the observation y_t

are **linear** with **Gaussian noise.**

State Space Models

- State space models revolutionized control theory in the late 50s and early 60s. Prior to these models, classical control theory could cope with decoupled low order systems. State space models allowed the effective control of complex systems like spacecraft and fast aircraft.
- The directed graph is identical to an HMM, so the conditional independencies are identical: Given the current state (*not* observation), the past and the future are independent.

Linear System: Definition



Linear System: Definition

State	$x_t \in \mathbb{R}^p$
Observation	$y_t \in \mathbb{R}^d$
Initial state	$x_0 \sim \mathcal{N}(0, P_0)$
Dynamics	$x_{t+1} = Ax_t + Gw_t, \qquad w_t \sim \mathcal{N}(0, Q)$
Observation	$y_t = Cx_t + v_t, \qquad v_t \sim \mathcal{N}(0, R).$

Linear System: Observations

1. All the distributions are Gaussian (joints, marginals, conditionals), so they can be described by their means and variances.
2. The conditional distribution of the next state, $x_{t+1}|x_t$, is

$$\mathcal{N}(Ax_t, GQG').$$

To see this:

$$\mathbb{E}(x_{t+1}|x_t) = Ax_t + GE(w_{t+1}|x_t) = Ax_t.$$

$$\begin{aligned}\text{Var}(x_{t+1}|x_t) &= \mathbb{E}(Gw_t(Gw_t)') \\ &= GQG'.\end{aligned}$$

Linear System: Observations

3. The marginal distribution of x_t is $\mathcal{N}(0, P_t)$, where P_0 is given and, for $t \geq 0$,

$$P_{t+1} = AP_tA' + GQG'.$$

To see this:

$$\begin{aligned}
\mathbb{E}x_{t+1} &= \mathbb{E}\mathbb{E}(x_{t+1}|x_t) = AE(x_t) = 0. \\
P_{t+1} &= \mathbb{E}(x_{t+1}x'_{t+1}) \\
&= \mathbb{E}((Ax_t + Gw_t)(Ax_t + Gw_t)') \\
&= AP_tA' + GQG'.
\end{aligned}$$

Inference in SSMs

Filtering: $p(x_t|y_0,\ldots,y_t)$.

Smoothing: $p(x_t|y_0,\ldots,y_T)$.

For inference, it suffices to calculate the appropriate conditional means and covariances.

Inference in SSMs: Notation

$$\hat{x}_{t|s} = \mathbb{E}(x_t | y_0, \ldots, y_s),$$

$$\hat{P}_{t|s} = \mathbb{E}((x_t - x_{t|s})(x_t - x_{t|s})' | y_0, \ldots, y_s).$$

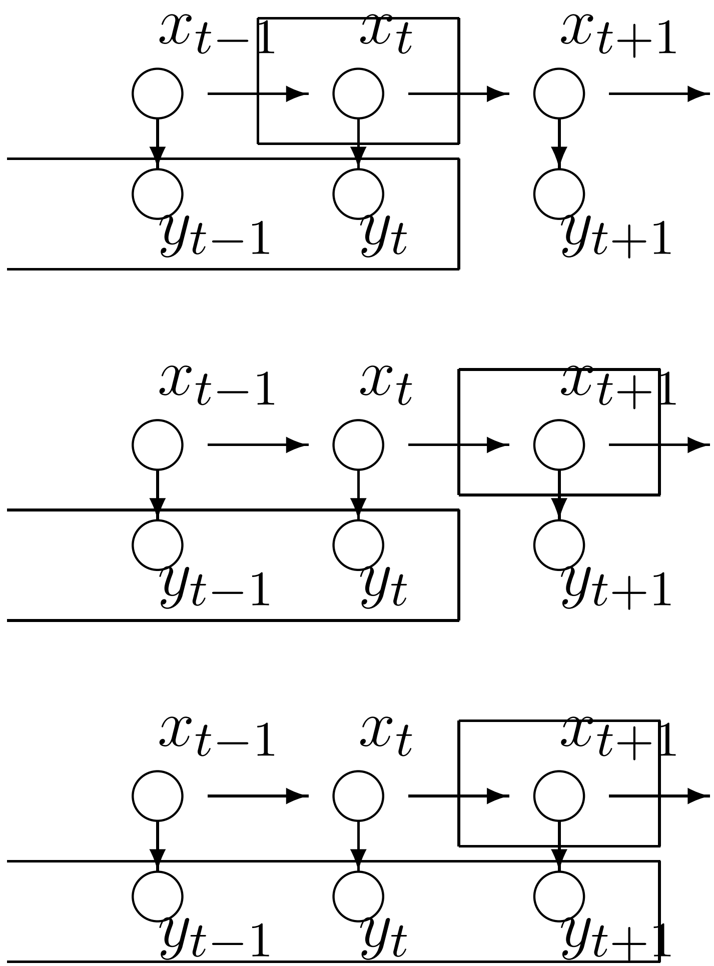
Filtering: $x_{t|t} \sim \mathcal{N}(\hat{x}_{t|t}, P_{t|t}),$

Smoothing: $x_{t|T} \sim \mathcal{N}(\hat{x}_{t|T}, P_{t|T}).$

The **Kalman Filter** is an inference algorithm for $\hat{x}_{t|t}$, $P_{t|t}$.

The **Kalman Smoother** is an inference algorithm for $\hat{x}_{t|T}$, $P_{t|T}$.

The Kalman Filter



$p(x_t|y_0,\ldots,y_t) \qquad \hat{x}_{t|t}, P_{t|t}$

time update

$p(x_{t+1}|y_0,\ldots,y_t) \qquad \hat{x}_{t+1|t}, P_{t+1|t}$

measurement update

$p(x_{t+1}|y_0,\ldots,y_{t+1}) \; \hat{x}_{t+1|t+1}, P_{t+1|t+1}$

Key ideas of this lecture

- Factor Analysis.
  - Recall: Gaussian factors plus observations.
  - Parameter estimation with EM.
- The vec operator.
  - Motivation: Natural parameters of Gaussian.
  - Linear functions of matrices as inner products. Kronecker product.
- State Space Models.
  - Linear dynamical systems, gaussian disturbances.
  - All distributions are Gaussian: parameters suffice.
  - Inference: Kalman filter and smoother.
  - Parameter estimation with EM.