CS281A/Stat241A Lecture 16

Multivariate Gaussians and Factor Analysis

Peter Bartlett

Key ideas of this lecture

- Factorizing multivariate Gaussians
  - Motivation: factor analysis, Kalman filter.
  - Marginal and conditional Gaussians.
  - Schur complement.
  - Moment and natural parameterizations.
  - Sherman/Woodbury/Morrison formula.
- Factor Analysis.
  - Examples: stock prices. Netflix preference data.
  - Model: Gaussian factors, conditional Gaussian observations.

Factor analysis: modelling stock prices

Suppose that we want to model stock prices, perhaps to choose a portfolio whose value does not fluctuate excessively:

portfolio weights:	$w \in \Delta_n$	(n-simplex)
growth from $t-1$ to t:	$w' y_t$	($y_t =$returns)
variance of growth:	$w' \Sigma w$.	

Want to align w with a bet direction ('airline stocks will fall') while minimizing variance.
Need a model for covariance of prices. Can't hope to estimate an arbitrary Σ.

Modelling stock prices

Some observations about stock data:

1. Prices today tend to be close to what they were yesterday. It's the **change in price** that is interesting: $p_t - p_{t-1}$, where p_t is the price at time t.
2. The variance of the price increases as the price increases. So it's appropriate to consider a transformation, like the log of the price:

$$y_t = \log\left(\frac{p_t}{p_{t-1}}\right).$$

Modelling stock prices

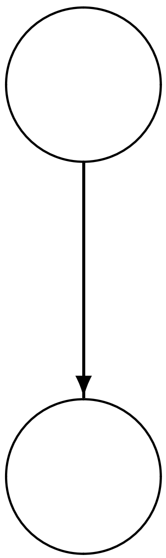
3. Stock prices tend to be strongly correlated:
   - Market moves.
   - Industry sectors (airlines, pharmaceuticals).

   We can think of the stock prices as affected by a (relatively small) set of **factors**:
   - The market as a whole
   - Technology versus not (NASDAQ vs NYSE)
   - Specific industry sectors
   - ...

   These factors have up and down days, and they affect different stocks differently.

Factor analysis

We can model a distribution like this using a directed graphical model:



$X_1 \in \mathbb{R}^p$, factors

$X_2 \in \mathbb{R}^q$, observations

Typically the number of factors is much smaller than the number of observations: $p \ll q$.

Factor analysis

We consider the local conditionals:

$$p(x_1) = \mathcal{N}(x_1 | 0, I),$$
$$p(x_2 | x_1) = \mathcal{N}(x_2 | \mu_2 + \Lambda x_1, \Sigma_{2|1}),$$

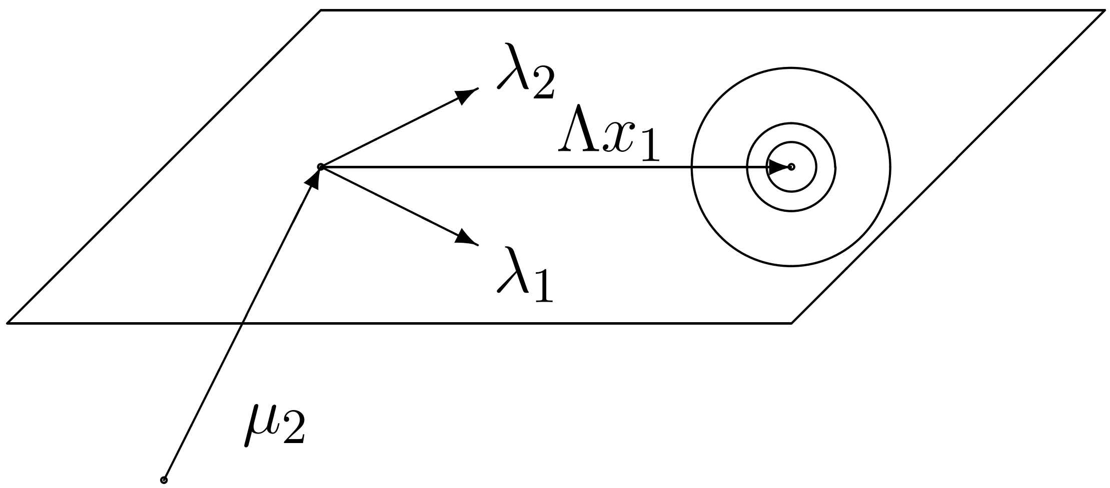
where the columns of $\Lambda \in \mathbb{R}^{q \times p}$ define the 'factors,' which form a p-dimensional subspace of $\mathbb{R}^q$. These are the directions in which X_2 varies the most (think of $\Sigma_{2|1}$ as not too large).

Factor analysis

$$p(x_1) = \mathcal{N}(x_1|0, I),$$

$$p(x_2|x_1) = \mathcal{N}(x_2|\mu_2 + \Lambda x_1, \Sigma_{2|1}),$$

$$\Lambda = [\lambda_1 \lambda_2 \cdots \lambda_p] \qquad \text{factors}$$



Factor analysis

This implies that the joint distribution is Gaussian:

$$(X_1, X_2) \sim \mathcal{N}(\mu, \Sigma).$$

- What is the relationship between the parameters of the joint distribution and those of the local conditionals?
- The same question arises when studying linear dynamical systems with Gaussian noise.

Factorizing Multivariate Gaussians

Notation:

$$x_1 \in \mathbb{R}^p,$$

$$x_2 \in \mathbb{R}^q$$

$$p(x_1, x_2) = (2\pi)^{-(p+q)/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu) \right)$$

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Factorizing Multivariate Gaussians

Theorem: [Marginal and conditional Gaussian]

$$p(x_1) = \mathcal{N}(x_1 | \mu_1, \Sigma_{11})$$

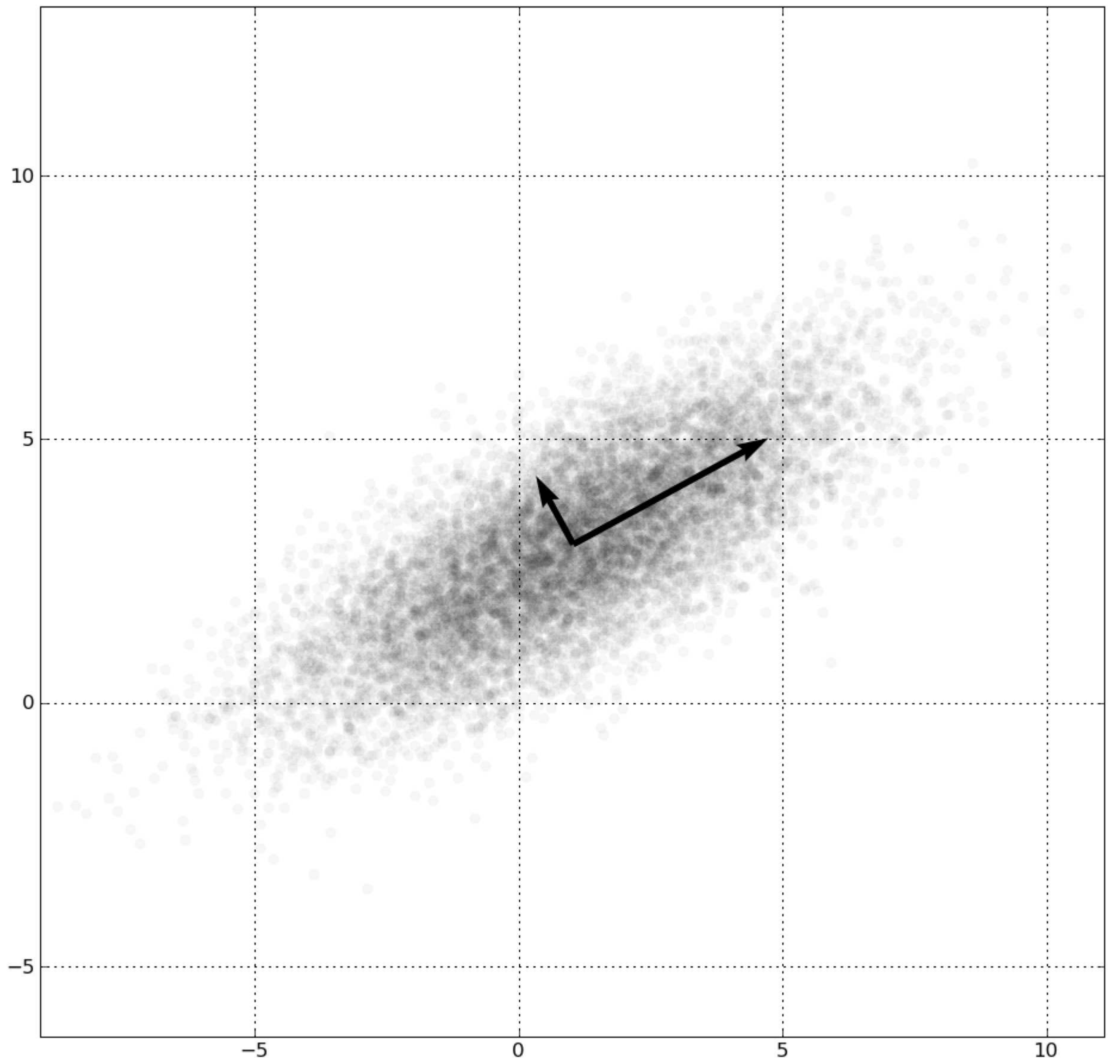
$$p(x_2 | x_1) = \mathcal{N}(x_2 | \mu_{2|1}(x_1), \Sigma_{2|1})$$

where $\mu_{2|1}(x_1) = \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)$

$$\Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

$\Sigma_{2|1}$ is Σ/Σ_{11}, the **Schur complement** of Σ wrt Σ_{11}.

Conditional Gaussians



Factorizing Multivariate Gaussians

Conditional Gaussian:

$$\mu_{2|1}(x_1) = \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)$$

$$\Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

- $\mu_{2|1} \neq \mu_2$.
- $\Sigma_{2|1} \leq \Sigma_{22}$.
- $x_1 \perp\!\!\!\perp x_2 \Rightarrow \Sigma_{2|1} = \Sigma_{22}$.

Factorizing Multivariate Gaussians

- Marginal parameters are simple for moment parameterization.
- Conditional parameters are simple for natural parameterization.
- Natural parameterization:

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

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

Factorizing Multivariate Gaussians

Corollary: [Marginal and conditional in natural parameters]

$$
\begin{aligned}
p(x_1) &= \mathcal{N}(x_1 | \eta_1^m, \Lambda_1^m) \\
p(x_2 | x_1) &= \mathcal{N}(x_2 | \eta_{2|1}^c(x_1), \Lambda_{2|1}^c) \\
\text{where } \eta_1^m &= \eta_1 - \Lambda_{12}\Lambda_{22}^{-1}\eta_2 \\
\Lambda_1^m &= \Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21} (= \Lambda/\Lambda_{22}) \\
\eta_{2|1}^c(x_1) &= \eta_2 - \Lambda_{21}x_1 \\
\Lambda_{2|1}^c &= \Lambda_{22}.
\end{aligned}
$$

Factorizing Multivariate Gaussians

Proof Idea:

- To split $p(x)$ into $p(x_1)p(x_2|x_1)$, we need to express

$$(x-\mu)'\Sigma^{-1}(x-\mu)$$

  as a sum of similar quadratic forms involving x_1 and x_2. For this, we need to decompose Σ^{-1}.
- We consider the block LDU decomposition of Σ^{-1}. LDU is lower triangular-diagonal-upper triangular. This relies on the idea of a **Schur complement** of a block matrix.

chur complements and LDU decompositio

Definition: [Schur complement]

For $$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \quad \text{where } |A|, |D| \neq 0,$$

define
$$M/A = D - CA^{-1}B,$$
$$M/D = A - BD^{-1}C.$$

chur complements and LDU decompositio

Lemma: [UDL decomposition]

$$\begin{bmatrix} A & 0 \\ 0 & M/A \end{bmatrix} = \begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I & -A^{-1}B \\ 0 & I \end{bmatrix}$$

$$M^{-1} = \begin{bmatrix} I & -A^{-1}B \\ 0 & I \end{bmatrix} \begin{bmatrix} A^{-1} & 0 \\ 0 & (M/A)^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix}$$

$$|M| = |M/A||A|$$

chur complements and LDU decompositio

Lemma: [LDU decomposition]

$$\begin{bmatrix} M/D & 0 \\ 0 & D \end{bmatrix} = \begin{bmatrix} I & -BD^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I & 0 \\ -D^{-1}C & I \end{bmatrix}$$

$$M^{-1} = \begin{bmatrix} I & 0 \\ -D^{-1}C & I \end{bmatrix} \begin{bmatrix} (M/D)^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix} \begin{bmatrix} I & -BD^{-1} \\ 0 & I \end{bmatrix}$$

$$|M| = |M/D||D|.$$

Schur complements/LDU decompositions

Proofs: The two formulations have identical proofs:

1. Easy to check: do the multiplication.
2. $(EFG)^{-1} = G^{-1}F^{-1}E^{-1}$, so $F^{-1} = G(EFG)^{-1}E$. Plug into 1.
3. Take determinants of 1.

An aside: S/W/M

Sherman/Woodbury/Morrison matrix inversion lemma
Corollary of LDU decomposition: For any (compatible) A, B, C, D,
if A, D are invertible,

$$(A - BDC)^{-1} = A^{-1} + A^{-1}B\left(D^{-1} - CA^{-1}B\right)^{-1}CA^{-1}.$$

Proof: Use the two expressions (LDU and UDL) for the top left block of M^{-1}:

$$\begin{aligned}\begin{pmatrix} I & 0 \end{pmatrix} M^{-1} \begin{pmatrix} I \\ 0 \end{pmatrix} &= A^{-1} + A^{-1}B(M/A)^{-1}CA^{-1} \\ &= (M/D)^{-1}.\end{aligned}$$

An aside: S/W/M

Corollary of LDU decomposition: For any (compatible) A, B, C, D,
if A, D are invertible,

$$(A - BDC)^{-1} = A^{-1} + A^{-1}B\left(D^{-1} - CA^{-1}B\right)^{-1}CA^{-1}.$$

Useful for incrementally updating the inverse of a matrix.
e.g., $S = X'X$ and its inverse S^{-1}.
Add a new observation x, inverse becomes

$$(S + xx')^{-1} = S^{-1} - S^{-1}x(1 + x'S^{-1}x)^{-1}x'S^{-1}.$$

This involves only matrix-vector multiplications: $O(d^2)$.
Versus matrix inversion: $O(d^3)$.

Gaussian Marginals and Conditionals

Now we can come back to the question of expressing a joint Gaussian as a marginal plus a conditional.
We can use the UDL decomposition to write

$$
\begin{aligned}
&\begin{pmatrix} x_1' & x_2' \end{pmatrix} \Sigma^{-1} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\
&= \begin{pmatrix} x_1' & x_2' \end{pmatrix} \begin{pmatrix} I & -\Sigma_{11}^{-1}\Sigma_{12} \\ 0 & I \end{pmatrix} \\
&\qquad \times \begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & (\Sigma/\Sigma_{11})^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -\Sigma_{21}\Sigma_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\
&= x_1'\Sigma_{11}^{-1}x_1 + \left(x_2 - \Sigma_{21}\Sigma_{11}^{-1}x_1\right)' (\Sigma/\Sigma_{11})^{-1} \left(x_2 - \Sigma_{21}\Sigma_{11}^{-1}x_1\right)
\end{aligned}
$$

Gaussian Marginals and Conditionals

Using this, we have

$$
\begin{aligned}
&p(x_1, x_2) \\
&= (2\pi)^{-(p+q)/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right) \\
&= (2\pi)^{-p/2}|\Sigma_{11}|^{-1/2} \exp\left(-\frac{1}{2}(x_1-\mu_1)'\Sigma_{11}^{-1}(x_1-\mu_1)\right) \\
&\qquad \times (2\pi)^{-q/2}|\Sigma/\Sigma_{11}|^{-1/2} \\
&\qquad \times \exp\left(-\frac{1}{2}(x_2-\mu_{2|1}(x_1))'(\Sigma/\Sigma_{11})^{-1}(x_2-\mu_{2|1}(x_1))\right) \\
&= \mathcal{N}(x_1|\mu_1, \Sigma_{11})\mathcal{N}\left(x_2|\underbrace{\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1-\mu_1)}_{\mu_{2|1}(x_1)}, \Sigma/\Sigma_{11}\right).
\end{aligned}
$$

Key ideas of this lecture

- Factorizing multivariate Gaussians
  - Motivation: factor analysis, Kalman filter.
  - Marginal and conditional Gaussians.
  - Schur complement.
  - Moment and natural parameterizations.
  - Sherman/Woodbury/Morrison formula.
- Factor Analysis.
  - Examples: stock prices. Netflix preference data.
  - Model: Gaussian factors, conditional Gaussian observations.
  - Parameter estimation with EM.

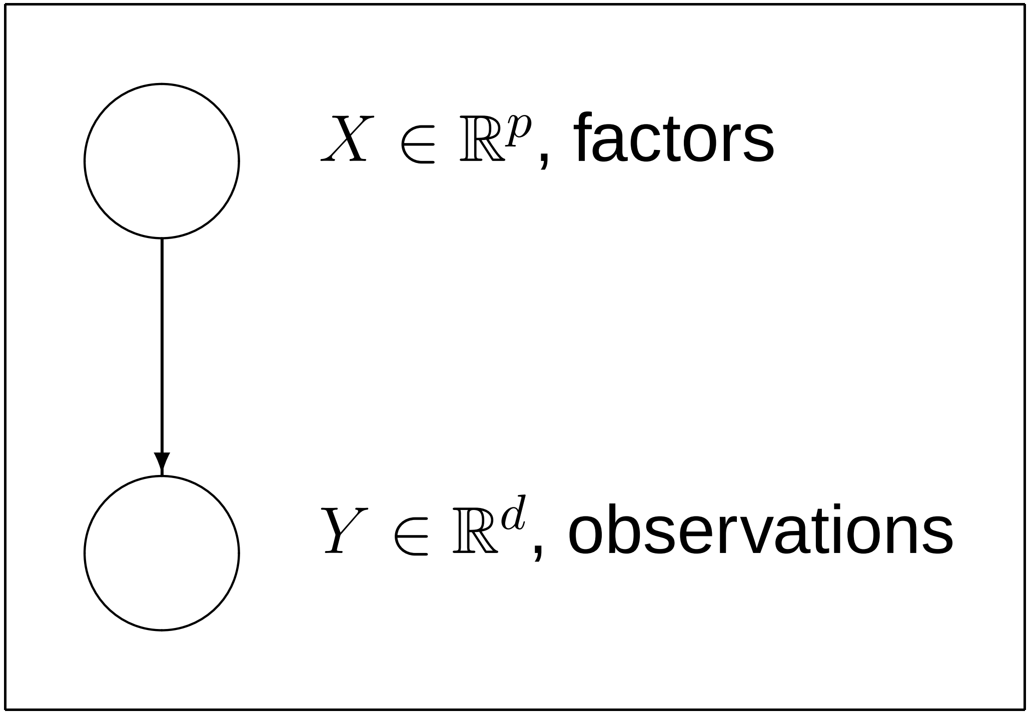
Factor Analysis: Motivation

Netflix movie ratings The data, for each individual, is a vector of their ratings (on the scale $[0, 5]$) of many tens of thousands of movies.

Again, the covariance of these variables is very structured: people tend to like movies of particular genres, and with particular stars. So the ratings of similar movies tend to be similar.

Again, we could hypothesize a factor model with a (relatively) small set of factors.

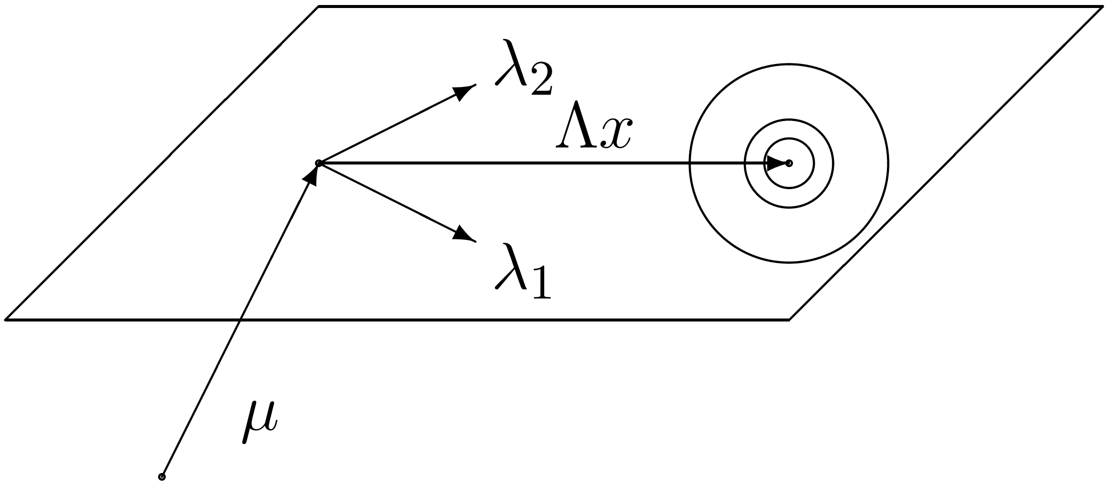
Factor Analysis: Definition



Local conditionals:

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

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 $d + dp + d \sim dp$ parameters.
- A full covariance matrix has d^2 parameters.
  Here, with only p factors (and $p \ll d$), the covariance for a factor model has far fewer parameters to estimate.

ctor Analysis: Joint, Marginals, Condition

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 $I - \Lambda' \left(\Lambda\Lambda' + \Psi\right)^{-1} \Lambda$.

ctor Analysis: Joint, Marginals, Condition

1. Shows that the marginal distribution for Y is centered at μ, and has covariance that is Ψ plus the low rank (rank $\leq p$) factored matrix $\Lambda\Lambda'$. If $p \ll d$, this corresponds to pd parameters, rather than d^2 for a full covariance matrix. It's an easy calculation (once we decompose y as $y = \mu + \Lambda x + w$).
2. Shows how the joint covariance depends on Λ. Again, it's an easy calculation using $y = \mu + \Lambda x + w$.
3. Shows how we can invert the conditional distribution. We'll rely on this for EM; x is the hidden variable. Its proof uses the theorem: take the joint and calculate the conditional.

Key ideas of this lecture

- Factorizing multivariate Gaussians
  - Motivation: factor analysis, Kalman filter.
  - Marginal and conditional Gaussians.
  - Schur complement.
  - Moment and natural parameterizations.
  - Sherman/Woodbury/Morrison formula.
- Factor Analysis.
  - Examples: stock prices. Netflix preference data.
  - Model: Gaussian factors, conditional Gaussian observations.
  - Parameter estimation with EM.