Lecture 10 — April 30

Lecturer: Lester Mackey *Scribe: Joey Arthur, Rakesh Achanta*

10.1 Factor Analysis

10.1.1 Recap

Recall the factor analysis (FA) model for linear dimensionality reduction of continuous data. In this model, our observations $x_i \in \mathbb{R}^p$ are related to latent factors $z_i \in \mathbb{R}^q$ in the following manner:

$$z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_{q\times q}), \quad x_i|z_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu + \Lambda z_i; \Psi),$$

where we assume $\Psi \in \mathbb{R}^{p\times p}$ is diagonal. Given the observations, we would like to infer the latent factors, which provide a lower dimensional (approximate) representation of our data. Last time we computed the conditional distribution of z_i given x_i, but this distribution of course depends on the unknown parameters $\theta = (\mu, \Lambda, \Psi)$. We derived the maximum likelihood estimator for μ, which is the sample mean. However, Λ and Ψ are coupled together in the likelihood by a determinant and a matrix inverse, and there is no closed-form MLE for these parameters. We will instead estimate Λ and Ψ using an EM algorithm.

10.1.2 EM Parameter Estimation

Since the MLE for μ is known, we will assume w.l.o.g. that the data have been mean-centered as $x_i \leftarrow x_i - \hat{\mu}_{\text{MLE}}$ and remove the parameter μ from the model. In order to derive an EM algorithm, we begin as usual with the complete log-likelihood of our data together with the latent variables:

$$\log p(z_{1:n}, x_{1:n}; \theta) = -\frac{1}{2}\sum_{i=1}^{n} z_i^T z_i - \frac{n}{2}\log|\Psi| - \frac{1}{2}\sum_{i=1}^{n}(x_i - \Lambda z_i)^T\Psi^{-1}(x_i - \Lambda z_i) + C_1, \quad (10.1)$$

where C_1 is a parameter free term including normalizing constants. Observe that in this complete log-likelihood, Λ and Ψ are no longer coupled together as they were in the marginal likelihood of the observed data. Noting that the $z_i^T z_i$ term above does not involve the

parameters and making several other simplifications, we have

$$\begin{aligned}
&\log p(z_{1:n},\, x_{1:n};\theta) \\
&\quad = -\frac{n}{2}\log|\Psi| - \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}\left((x_i-\Lambda z_i)^T\Psi^{-1}(x_i-\Lambda z_i)\right) + C_2 \\
&\quad = -\frac{n}{2}\log|\Psi| - \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}\left((x_i-\Lambda z_i)(x_i-\Lambda z_i)^T\Psi^{-1}\right) + C_2 \\
&\quad = -\frac{n}{2}\log|\Psi| - \frac{n}{2}\sum_{i=1}^{n}\operatorname{tr}(S\Psi^{-1}) + C_2,
\end{aligned}$$

where S is defined as the empirical conditional covariance

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\Lambda z_i)(x_i-\Lambda z_i)^T = \frac{1}{n}\sum_{i=1}^{n}\left[x_ix_i^T + \Lambda z_i z_i^T\Lambda^T - \Lambda z_i x_i^T - x_i z_i^T\Lambda^T\right].$$

In the second line above we used the fact that a scalar is equal to its trace. In the third line we used the cyclic property of the trace $\operatorname{tr}(ABC) = \operatorname{tr}(CAB)$, which can be applied whenever the matrix/vector multiplications are all well-defined.

We now derive the E-step by computing the expected complete log-likelihood (ECLL) under $q_t(z_{1:n}) = p(z_{1:n}|x_{1:n};\theta^{(t)})$, where $\theta^{(t)}$ is our estimate from the previous EM iteration. Recall that this conditional distribution is Gaussian and that we derived its mean and covariance last time. The ECLL is

$$\mathbb{E}_{q_t}\log p(z_{1:n}, x_{1:n};\theta) = -\frac{n}{2}\log|\Psi| - \frac{n}{2}\operatorname{tr}\left(\mathbb{E}_{q_t}[S]\Psi^{-1}\right) + C_2$$

after interchanging the trace and expectation. We must therefore compute

$$\mathbb{E}_{q_t}[S] = \frac{1}{n}\sum_{i=1}^{n}\left[x_ix_i^T + \Lambda\mathbb{E}_{q_t}[z_iz_i^T]\Lambda^T - \Lambda\mathbb{E}_{q_t}[z_i]x_i^T - x_i\mathbb{E}_{q_t}[z_i^T]\Lambda^T\right],$$

where $\mathbb{E}_{q_t}[z_i] = \mathbb{E}[z_i|x_i]$ was computed last time and

$$\mathbb{E}_{q_t}[z_iz_i^T] = \operatorname{Cov}[z_i|x_i] + \mathbb{E}[z_i|x_i]\mathbb{E}[z_i|x_i]^T$$

is similarly easy to compute.

For the M-step, one can show that the ECLL is maximized by taking

$$\Lambda^{(t+1)} = \left(\sum_i x_i\mathbb{E}_{q_t}[z_i^T]\right)\left(\sum_i \mathbb{E}_{q_t}[z_iz_i^T]\right)^{-1}$$

and

$$\Psi^{(t+1)} = \operatorname{diag}(\mathbb{E}_{q_t}[S]) = \frac{1}{n}\operatorname{diag}(\sum_i x_ix_i^T - \Lambda^{(t+1)}\sum_i \mathbb{E}_{q_t}[z_i]x_i^T).$$

Note the similarity of the Λ update to the normal equations solved during linear regression. Also, notice that the Ψ update involves the updated $\Lambda^{(t+1)}$ and not $\Lambda^{(t)}$.

10.1.3 Observations

There are several connections between FA and previous models/algorithms we have considered. We might consider FA as similar to Gaussian mixture modeling but with the latent variables z_i continuous rather than discrete. We can also draw similarities between FA and PCA. Both methods describe data using a lower dimensional linear representation. However, factor analysis allows for more general covariance structure than PCA does, and so the loadings and factors derived from factor analysis do not in general correspond to the results of PCA. In the case that Ψ is restricted to be isotropic (i.e., $\Psi = \sigma^2 I$ for unknown σ^2) we recover the probabilistic PCA (PPCA) model (Tipping & Bishop '95). In this restricted case there are closed form MLEs. If U is the matrix whose columns are the top q eigenvectors of the empirical covariance $\frac{X^T X}{n}$, and $\lambda_1, \ldots, \lambda_p$ are the eigenvalues, then we have

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{p-q} \sum_{j=q+1}^{p} \lambda_j,$$
$$\hat{\Lambda}_{\text{MLE}} = U(\text{diag}(\lambda_1, \ldots, \lambda_q) - \hat{\sigma}^2_{\text{MLE}})^{1/2}.$$

In this restricted setup, the factor analysis loadings (columns of $\hat{\Lambda}$) span the same subspace as the PCA loadings U. Moreover, if we consider σ^2 as known then as $\sigma^2 \to 0$, PPCA actually recovers the PCA algorithm. This is another example of *small variance asymptotics*, like we have seen before.

We should also mention a few caveats to using factor analysis. First, the FA parameters are in general *not* identifiable. For example, given an orthogonal matrix O (such that $OO^T = O^T O = I$), the parameters Λ and ΛO will give rise to the same distribution of x_i. Hence, interpretation of the learned values of Λ and z_i must be done with care.

Even apart from these interpretability issues, factor analysis treats datapoints as independent draws. What if our data has known, e.g., sequential, dependence structure? Such structure arises in a variety of settings:

- Tracking 3D object movement given radar or video
- Autopilot, in which we would like to estimate the state of a vehicle over time from internal and external sensors
- The inference of evolving market factors from financial time series
- Character recognition based on touch screen contact over time
- GPS navigation
- Recommender systems, in which we aim to estimate users' preferences over time

We will next investigate a probabilistic model designed for data with such a known sequential dependence structure.

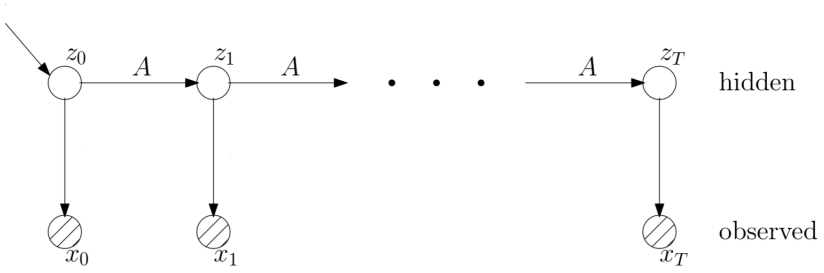

Figure 10.1. Graphical model for LGSSM

10.2 Linear Gaussian State Space Model

The **linear Gaussian state space model** is a generalization of factor analysis to the settting of *sequential* continuous data. Under this model, we view our data sequence $x_0, x_1, \ldots x_T \in \mathbb{R}^p$ as a random draw from the following generative process:

(0) $z_0 \sim \mathcal{N}(0, \Sigma_0)$: sample the initial state in $\mathbb{R}^q$

(1) $z_t = Az_{t-1} + w_{t-1}$, where, $w_{t-1} \overset{\text{ind}}{\sim} \mathcal{N}(0, Q)$ or alternately, $z_t|z_{t-1} \overset{\text{ind}}{\sim} \mathcal{N}(Az_{t-1}, Q)$. i.e. z_t is sampled from linear gaussian dynamics given the prior state z_{t-1} via the unknown transition matrix $A \in \mathbb{R}^{q\times q}$ and unknown covariance matrix $Q \in \mathbb{R}^{q\times q}$.

(2) $x_t = Cz_t + v_t$ for $v_t \overset{\text{ind}}{\sim} \mathcal{N}(0, R)$ or, $x_t|z_t \overset{\text{ind}}{\sim} \mathcal{N}(Cz_t, R)$. i.e. x_t are the sample observations given the state z_t, normally distributed with mean Cz_t, where $C \in \mathbb{R}^{p\times q}$ is the unknown **loadings matrix**, and unknown covariance $R \in \mathbb{R}^{p\times p}$

Notice that this is similar to the emission model from factor analysis but with a more general covariance matrix R and with dependent states.

10.2.1 Graphical Model

The LGSSM graphical model is the same as the Hidden Markov Model graphical model, since the two models have identical conditional independence structures. However, in the present setting, we have Gaussian as opposed to discrete hidden variables z_i.

10.2.2 Unsupervised Learning Goal

Our unsupervised learning goal is to draw inferences about the hidden states $z_0, z_1, \ldots, z_T$. Here are three of the most common inferential tasks:

(1) *Filtering.* Infer the current state given history of observations $P(z_t|x_0, \ldots x_t)$. e.g:- What is the current state of the missile given its position over some past time?

(2) *Smoothing.* Infer a past state given observations $P(z_s|x_0, \ldots x_t)$ where $s < t$ e.g:- Where did the missile originate given we observed it over some time?

(3) *Prediction.* Predict a future state given observations $P(z_u|x_0, \dots x_t)$ where $u > t$ e.g:- Where would we expect the missile to be in some time from now?

In this lecture and the next, we will detail recursive algorithms for carrying out filtering and smoothing, assuming all model parameters are known.

10.2.3 Kalman Filter

The **Kalman Filter** is an algorithm for filtering in an LGSSM where the parameters are known. We wish to find the probability distribution of the current state given history of observations. Since the states and the observations are jointly Gaussian, it suffices to find the mean and variance of the conditional distribution which is also going to be Gaussian. We introduce the following shorthand notation for filtered means and covariances

$$\begin{aligned}\hat{z}_{t|t} &= \mathbb{E}[z_t|x_{0:t}] \\ P_{t|t} &= \mathbb{E}[(z_t - \hat{z}_{t|t})(z_t - \hat{z}_{t|t})^T|x_{0:t}]\end{aligned}$$

We will also be interested in computing the one-step prediction means and covariances $\hat{z}_{t|t-1}, P_{t|t-1}$ of $Pr(z_t|x_{0:t-1})$.

Filtering Strategy

We will derive our filtering algorithm via a two step recursion:

(1) Time update: Compute the prediction distribution $P(z_{t+1}|x_{0:t})$ given the last filtered distribution $P(z_t|x_{0:t})$

(2) Measurement update: Compute the new filtered distribution $P(z_{t+1}|x_{0:t+1})$ given the prediction distribution $P(z_{t+1}|x_{0:t})$

Time Update

We will use the fact that $z_{t+1} = Az_t + w_t$ to compute the mean and covariance of the prediction distribution from the filtered distribution:

$$\begin{aligned}\hat{z}_{t+1|t} &= \mathbb{E}[Az_t + w_t|x_{0:t}] \\ &= A\mathbb{E}[z_t|x_{0:t}] + 0 \\ &= A\hat{z}_{t|t}\end{aligned}$$

$$\begin{aligned}P_{t+1|t} &= \mathbb{E}[(z_{t+1} - \hat{z}_{t+1|t})(z_{t+1} - \hat{z}_{t+1|t})^T|x_{0:t}] \\ &= \mathbb{E}[(Az_t + w_t - A\hat{z}_{t|t})(Az_t + w_t - A\hat{z}_{t|t})^T|x_{0:t}] \\ &= A\mathbb{E}[(z_t - \hat{z}_{t|t})(z_t - \hat{z}_{t|t})^T|x_{0:t}]A^T + \mathbb{E}[w_t w_t^T|x_{0:t}] + 0 \\ &= AP_{t|t}A^T + Q\end{aligned}$$