Lab 2: Bayesian PCA

Machine Learning: Principles and Methods, November 2013

- The lab exercises should be made in groups of three people, or at least two people.
- The deadline is Wednesday, 11 December, 23:59.
- Assignment should be sent to T.S.Cohen at uva dot nl (Taco Cohen). The subject line of your email should be "[MLPM2013] lab#_lastname1_lastname2_lastname3".
- Put your and your teammates' names in the body of the email
- Attach the .IPYNB (IPython Notebook) file containing your code and answers. Naming of the file follows the same rule as the subject line. For example, if the subject line is "[MLPM2013] lab01_Kingma_Hu", the attached file should be "lab01_Kingma_Hu.ipynb". Only use underscores ("_") to connect names, otherwise the files cannot be parsed.

Notes on implementation:

- You should write your code and answers in an IPython Notebook: http://ipython.org/notebook.html (http://ipython.org/notebook.html). If you have problems, please contact us.
- Among the first lines of your notebook should be "%pylab inline". This imports all required modules, and your plots will appear inline.
- NOTE: test your code and make sure we can run your notebook / scripts!

Introduction

In this lab assignment, we will implement a variational algorithm for Bayesian PCA. Unlike regular PCA based on maximization of retained variance or minimization of projection error (see Bishop, 12.1.1 and 12.1.2), probabilistic PCA defines a proper density model over observed and latent variables. We will work with a fully Bayesian model this time, which is to say that we will put priors on our parameters and will be interested in learning the posterior over those parameters. Bayesian methods are very elegant, but require a shift in mindset: we are no longer looking for a point estimate of the parameters (as in maximum likelihood or MAP), but for a full posterior distribution.

The integrals involved in a Bayesian analysis are usually analytically intractable, so that we must resort to approximations. In this lab assignment, we will implement the variational method described in Bishop99. Chapter 12 of the PRML book contains additional material that may be useful when doing this exercise.

- [Bishop99] Variational Principal Components, C. Bishop, ICANN 1999

Below, you will find some code to get you started.

1. The Q-distribution (5 points)

In variational Bayes, we introduce a distribution $Q(\Theta)$ over parameters / latent variables in order to make inference tractable. We can think of Q as being an approximation of a certain distribution. What function does Q approximate, $p(D|\Theta)$, $p(\Theta|D)$, $p(D, \Theta)$, $p(\Theta)$, $p(D)$, and how do you see that?

A

It approximates the posterior $p(\Theta|D)$, which is seen from eq. 11 in [Bishop99]. The KL divergence is 0 when Q equals the posterior.

2. The mean-field approximation (15 points)

Equation 13 from [Bishop99] is a very powerful result: assuming only that $Q(\Theta)$ factorizes in a certain way (no assumptions on the functional form of the factors Q_i!), we get a set of coupled equations for the Q_i.

However, the expression given in eq. 13 for Q_i contains an error. Starting with the expression for the lower bound $\mathcal{L}(Q)$, derive the correct expression. You can proceed as follows: first, substitute the factorization of Q (eq. 12) into the definition of $\mathcal{L}(Q)$ and separate $\mathcal{L}(Q)$ into Q_i-dependent and Q_i-independent terms. At this point, you should be able to spot the expectations $< \cdot >_{k \neq i}$ over the other Q-distributions that appear in Bishop's solution (eq. 13). Now, keeping all $Q_k, k \neq i$ fixed, maximize the expression with respect to Q_i. You should be able to spot the form of the optimal lnQ_i, from which Q_i can easily be obtained.

A

Correct derivation can be found in PRML, section 10.1.1. If they have read this and see that it is what we are asking here, that's fine: they have understood the material.

3. The log-probability (10 points)

Write down the log-prob of data and parameters, $\ln p(\mathbf{X}, \mathbf{Z}, \mathbf{W}, \alpha, \tau, \mu)$, in full detail (where $\mathbf{X}$ are observed, $\mathbf{Z}$ is latent; this is different from [Bishop99] who uses $\mathbf{T}$ and $\mathbf{X}$ respectively, but $\mathbf{X}$ and $\mathbf{Z}$ are consistent with the PRML book and are more common nowadays). Could we use this to assess the convergence of the variational Bayesian PCA algorithm? If yes, how? If no, why not?

A

$$
\begin{aligned}
\ln p(\mathbf{X}, \mathbf{Z}, \mathbf{W}, \alpha, \tau, \mu) &= \ln p(\mathbf{X}|\mathbf{Z}, \mathbf{W}, \mu, \tau) \\
&+ \ln p(\mathbf{Z}) \\
&+ \ln p(\mu) \\
&+ \ln p(\tau) \\
&+ \ln p(\mathbf{W}|\alpha) \\
&+ \ln p(\alpha) \\
&= \sum_n \ln \mathcal{N}(\mathbf{x}_n|\mathbf{W}\mathbf{z}_n + \mu, \tau) + \ln \mathcal{N}(\mathbf{z}_n|\mathbf{0}, \mathbf{I}) \\
&+ \ln \mathcal{N}(\mu|\mathbf{0}, \beta^{-1}\mathbf{I}) \\
&+ \ln \Gamma(\tau|a_\tau, b_\tau) \\
&+ \ln \prod_i \mathcal{N}(\|\mathbf{w}_i\||0, \alpha_i) \\
&+ \ln \prod_i \Gamma(\alpha_i|a_\alpha, b_\alpha) \\
&= \sum_n -\frac{\tau}{2}\|\mathbf{x}_n - \mathbf{W}\mathbf{z}_n - \mu\|^2 - \frac{d}{2}\ln 2\pi + \frac{d}{2}\ln \tau \\
&+ \sum_n -\frac{1}{2}\|\mathbf{z}_n\|^2 - \frac{d}{2}\ln 2\pi \\
&+ -\frac{\beta}{2}\|\mu\|^2 - \frac{d}{2}\ln 2\pi + \frac{d}{2}\ln \beta \\
&+ a_\tau \ln b_\tau + (a_\tau - 1)\ln \tau - b_\tau \tau - \ln \Gamma(a_\tau) \\
&+ \sum_i -\frac{\alpha_i}{2}\|\mathbf{w}_i\|^2 + \frac{d}{2}\ln \alpha_i - \frac{d}{2}\ln 2\pi \\
&+ \sum_i a_\alpha \ln b_\alpha + (a_\alpha - 1)\ln \alpha_i - b_\alpha \alpha_i - \ln \Gamma(a_\alpha)
\end{aligned}
$$

It is not sufficient for assessing convergence, because we don't have point estimates of the parameters as in ML/MAP. In variational Bayes, we're optimizing the lower bound on $p(X)$. It is part of computation of the lower bound though, which can be computed to assess convergence.

4. The lower bound $\mathcal{L}(Q)$ (25 points)

Derive an expression for the lower bound $\mathcal{L}(Q)$ of the log-prob $\ln p(X)$ for Bayesian PCA, making use of the factorization (eq. 12) and the form of the Q-distributions (eq. 16-20) as listed in [Bishop99]. Show your steps. Implement this function.

The following result may be useful:

For $x \sim \Gamma(a, b)$, we have $\langle \ln x \rangle = \psi(a) - \ln b$, where $\psi(a) = \frac{\Gamma'(a)}{\Gamma(a)}$ is the digamma function (which is implemented in numpy.special).

A

The definition:

$$\mathcal{L}(Q) = \int Q(\Theta) \ln \frac{p(D,\Theta)}{Q(\Theta)} d\Theta$$

$$= \int Q(\Theta) \ln p(D,\Theta) d\Theta - \int Q(\Theta) \ln Q(\Theta) d\Theta$$

Let's look at the first term first. Substitute the previous result for the log-prob in this equation, we find:

$$\begin{aligned}
\int Q(\Theta) \ln p(D,\Theta) d\Theta = &- \sum_n \int Q(\mathbf{W}) Q(\mu) Q(\mathbf{z}_n) Q(\tau) \frac{\tau}{2} \|\mathbf{x}_n - \mathbf{W}\mathbf{z}_n - \mu\|^2 d\mathbf{W} d\mu d\mathbf{z}_n d\tau \\
&- N\frac{d}{2} \ln 2\pi + N\frac{d}{2} \int Q(\tau) \ln \tau d\tau \\
&- \frac{1}{2} \sum_n \int Q(\mathbf{z}_n) \|\mathbf{z}_n\|^2 d\mathbf{z}_n \\
&- \frac{Nd}{2} \ln 2\pi \\
&- \frac{\beta}{2} \int Q(\mu) \|\mu\|^2 d\mu - \frac{d}{2} \ln 2\pi + \frac{d}{2} \ln \beta \\
&+ a_\tau \ln b_\tau - \ln \Gamma(a_\tau) + (a_\tau - 1) \int Q(\tau) \ln \tau d\tau - b_\tau \int Q(\tau) \tau d\tau \\
&+ \sum_i -\frac{1}{2} \int Q(\alpha_i) Q(\mathbf{w}_i) \alpha_i \|\mathbf{w}_i\|^2 d\alpha_i d\mathbf{w}_i + \frac{d}{2} \int Q(\alpha_i) \ln \alpha_i d\alpha_i - \frac{d}{2} \ln 2\pi \\
&+ \sum_i a_\alpha \ln b_\alpha + (a_\alpha - 1) \int Q(\alpha_i) \ln \alpha_i d\alpha_i - b_\alpha \int Q(\alpha_i) \alpha_i d\alpha_i - \ln \Gamma(a_\alpha)
\end{aligned}$$

Next, we perform each integral:

$$\begin{aligned}
\int Q(\mathbf{W}) Q(\mu) Q(\mathbf{z}_n) Q(\tau) \frac{\tau}{2} \|\mathbf{x}_n - \mathbf{W}\mathbf{z}_n - \mu\|^2 d\mathbf{W} d\mu d\mathbf{z}_n d\tau &= \frac{\langle\tau\rangle}{2} \int Q(\mathbf{W}) Q(\mu) Q(\mathbf{z}_n) \|\mathbf{x}_n - \mathbf{W}\mathbf{z}_n - \mu\|^2 d\mathbf{W} d\mu d\mathbf{z}_n \\
&= \frac{\langle\tau\rangle}{2} \int Q(\mathbf{W}) Q(\mu) Q(\mathbf{z}_n) \left(\mathbf{x}_n^T\mathbf{x}_n - \mathbf{x}_n^T\mathbf{W}\mathbf{z}_n - \mathbf{x}_n^T\mu - \mathbf{z}_n^T\mathbf{W}^T\mathbf{x}_n + \mathbf{z}_n^T\mathbf{W}^T\mathbf{W}\mathbf{z}_n + \mathbf{z}_n^T\mathbf{W}^T\mu - \mu^T\mathbf{x}_n + \mu^T\mathbf{W}\mathbf{z}_n + \mu^T \right. \\
&= \frac{\langle\tau\rangle}{2} \int Q(\mu) Q(\mathbf{z}_n) \left(\mathbf{x}_n^T\mathbf{x}_n - \mathbf{x}_n^T\langle\mathbf{W}\rangle\mathbf{z}_n - \mathbf{x}_n^T\mu - \mathbf{z}_n^T\langle\mathbf{W}^T\rangle\mathbf{x}_n + \mathbf{z}_n^T\langle\mathbf{W}^T\mathbf{W}\rangle\mathbf{z}_n + \mathbf{z}_n^T\langle\mathbf{W}^T\rangle\mu - \mu^T\mathbf{x}_n + \mu^T\langle\mathbf{W}\rangle\mathbf{z}_n - \right. \\
&= \frac{\langle\tau\rangle}{2} \int Q(\mathbf{z}_n) \left(\mathbf{x}_n^T\mathbf{x}_n - \mathbf{x}_n^T\langle\mathbf{W}\rangle\mathbf{z}_n - \mathbf{x}_n^T\langle\mu\rangle - \mathbf{z}_n^T\langle\mathbf{W}^T\rangle\mathbf{x}_n + \mathbf{z}_n^T\langle\mathbf{W}^T\mathbf{W}\rangle\mathbf{z}_n + \mathbf{z}_n^T\langle\mathbf{W}^T\rangle\langle\mu\rangle - \langle\mu^T\rangle\mathbf{x}_n + \langle\mu^T\rangle\langle\mathbf{W}\rangle \right. \\
&= \frac{\langle\tau\rangle}{2} \left(\mathbf{x}_n^T\mathbf{x}_n - \mathbf{x}_n^T\langle\mathbf{W}\rangle\langle\mathbf{z}_n\rangle - \mathbf{x}_n^T\langle\mu\rangle - \langle\mathbf{z}_n^T\rangle\langle\mathbf{W}^T\rangle\mathbf{x}_n + \mathrm{Tr}(\langle\mathbf{W}^T\mathbf{W}\rangle\langle\mathbf{z}_n\mathbf{z}_n^T\rangle) + \langle\mathbf{z}_n^T\rangle\langle\mathbf{W}^T\rangle\langle\mu\rangle - \langle\mu^T\rangle\mathbf{x}_n + \langle\mu^T\rangle\langle\mathbf{V} \right. \\
&= \frac{\langle\tau\rangle}{2} \left(\mathbf{x}_n^T\mathbf{x}_n - 2\mathbf{x}_n^T\langle\mathbf{W}\rangle\langle\mathbf{z}_n\rangle - 2\mathbf{x}_n^T\langle\mu\rangle + \mathrm{Tr}(\langle\mathbf{W}^T\mathbf{W}\rangle\langle\mathbf{z}_n\mathbf{z}_n^T\rangle) + 2\langle\mu^T\rangle\langle\mathbf{W}\rangle\langle\mathbf{z}_n\rangle + \langle\mu^T\mu\rangle \right)
\end{aligned}$$

General result. For $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \Sigma)$, we seek $\langle\|\mathbf{x}\|^2\rangle$.

$$\begin{aligned}
\int \|\mathbf{x}\|^2 \mathcal{N}(\mathbf{x}|\mathbf{m},\Sigma) d\mathbf{x} &= \sum_i \int x_i^2 \mathcal{N}(\mathbf{x}|\mathbf{m},\Sigma) d\mathbf{x} \\
&= \sum_i \int x_i^2 \mathcal{N}(x_i|m_i,\sigma_i^2) dx_i \\
&= \sum_i E(x_i^2)
\end{aligned}$$

Where σ_i^2 is the i-th diagonal element of Σ. Using $\sigma^2 = E(x^2) - (E(x))^2$, we find

$$\langle\|\mathbf{x}\|^2\rangle = \sum_i m_i^2 + \sigma_i^2 = \|\mathbf{m}\|^2 + \mathrm{Tr}(\Sigma)$$

.

Using these results we find

$$\begin{aligned}
\int Q(\Theta)\ln p(D,\Theta)d\Theta = &-\sum_n \frac{\langle\tau\rangle}{2}\left(\mathbf{x}_n^T\mathbf{x}_n - 2\mathbf{x}_n^T\langle\mathbf{W}\rangle\langle\mathbf{z}_n\rangle - 2\mathbf{x}_n^T\langle\mu\rangle + \mathrm{Tr}(\langle\mathbf{W}^T\mathbf{W}\rangle\langle\mathbf{z}_n\mathbf{z}_n^T\rangle) + 2\langle\mu^T\rangle\langle\mathbf{W}\rangle\langle\mathbf{z}_n\rangle + \langle\mu^T\mu\rangle\right)\\
&- N\frac{d}{2}\ln 2\pi + N\frac{d}{2}(\psi(\tilde{a}_\tau) - \ln\tilde{b}_\tau)\\
&- \frac{1}{2}\sum_n \|\mathbf{m}_\mathbf{z}{}^n\|^2\\
&- \frac{1}{2}N\mathrm{Tr}(\mathbf{\Sigma_z}) - Nd\ln 2\pi\\
&- \frac{\beta}{2}(\|\mathbf{m}_\mu\|^2 + \mathrm{Tr}(\mathbf{\Sigma}_\mu))\\
&- \frac{d}{2}\ln 2\pi + \frac{d}{2}\ln\beta\\
&+ a_\tau\ln b_\tau - \ln\Gamma(a_\tau) + (a_\tau - 1)(\psi(\tilde{a}_\tau) - \ln\tilde{b}_\tau) - b_\tau\frac{\tilde{a}_\tau}{\tilde{b}_\tau}\\
&+ \sum_i -\frac{\tilde{a}_\alpha}{2\tilde{b}_{\alpha_i}}(\|\mathbf{m}_\mathbf{w}{}^i\|^2 + \mathrm{Tr}(\mathbf{\Sigma_w})) + \frac{d}{2}(\psi(\tilde{a}_\alpha) - \ln\tilde{b}_{\alpha_i}) - \frac{d}{2}\ln 2\pi\\
&+ \sum_i a_\alpha\ln b_\alpha + (a_\alpha - 1)(\psi(\tilde{a}_\alpha) - \ln\tilde{b}_{\alpha_i}) - b_\alpha\frac{\tilde{a}_\alpha}{\tilde{b}_{\alpha_i}} - \ln\Gamma(a_\alpha)
\end{aligned}$$

The second term of $\mathcal{L}(Q)$ is the sum of entropies of the Q distributions. From wikipedia:

$$-\int Q(\mathbf{z}_n)\ln Q(\mathbf{z}_n) = \frac{d}{2}(1+\ln 2\pi) + \frac{1}{2}\ln|\mathbf{\Sigma_z}|$$

Since this does not depend on n, we can simply multiply by N to get the entropy of all latent variables $\mathbf{Z}$.

$$\begin{aligned}
-\int Q(\mu)\ln Q(\mu) &= \frac{d}{2}(1+\ln 2\pi) + \frac{1}{2}\ln|\mathbf{\Sigma}_\mu|\\
-\int Q(\mathbf{w}_i)\ln Q(\mathbf{w}_i) &= \frac{d}{2}(1+\ln 2\pi) + \frac{1}{2}\ln|\mathbf{\Sigma_w}|
\end{aligned}$$

For the gammas:

$$\begin{aligned}
-\int Q(\alpha_i)\ln Q(\alpha_i) &= \tilde{a}_\alpha - \ln\tilde{b}_{\alpha_i} + \ln\Gamma(\tilde{a}_\alpha) + (1-\tilde{a}_\alpha)\psi(\tilde{a}_\alpha)\\
-\int Q(\tau)\ln Q(\tau) &= \tilde{a}_\tau - \ln\tilde{b}_\tau + \ln\Gamma(\tilde{a}_\tau) + (1-\tilde{a}_\tau)\psi(\tilde{a}_\tau)
\end{aligned}$$

The final result is then obtained by adding the $\int Q(\Theta)\ln p(D,\Theta)d\Theta$ term derived above to the sum over the entropies of all variables.