T-61.5140 Machine Learning: Advanced Probabilistic Methods
Pekka Marttinen, Sami Remes (Spring 2015)
Exercise solutions, round 8, Tuesday, 17th March, 2015
https://noppa.aalto.fi/noppa/kurssi/t-61.5140

Problem 1. *"Factor analysis on exam grades."*

First go through the Matlab Statistics toolbox demo on factor analysis[1] *(write* `showdemo factorandemo` *in Matlab). The demo consists of fitting an FA model to data consisting of five different exam grades of some students, specifically two mathematics scores, two on literature and one comprehensive exam.*

What if you replaced the variable containing the grade from the comprehensive exam with one containing some randomly assigned scores, and redid the analysis? Does the interpretation of the two common factors change?

Solution. Rerun the example code with replacing the one variable with random scores.

Problem 2. *"VB for a factor analysis model."*

Consider the factor analysis model

$$\begin{aligned} \mathbf{x}_n &\sim \mathcal{N}_D(\mathbf{W}\mathbf{z}_n, \mathrm{diag}(\boldsymbol{\psi})^{-1}) \quad \forall n \in \{1,\ldots,N\} \\ \psi_d &\sim \mathrm{Gamma}(a,b) \quad \forall d \in \{1,\ldots,D\} \\ \mathbf{W}_k &\sim \mathcal{N}_D(\mathbf{0}, \alpha\mathbf{I}) \quad \forall k \in \{1,\ldots,K\} \\ \mathbf{z}_n &\sim \mathcal{N}_K(\mathbf{0},\mathbf{I}) \quad \forall n \in \{1,\ldots,N\}, \end{aligned}$$

where $\mathbf{W}_k$ denotes the loadings of the kth factor and ψ_d^{-1} the specific variance of the dth observed variable. Furthermore D denotes the number of observed variables (i.e. $\mathbf{x}_n \in R^D$), N the number of data points, and K the number of factors in the model. $\mathrm{diag}(\boldsymbol{\psi})$ *is a diagonal matrix with elements of $\boldsymbol{\psi} = (\psi_1,\ldots,\psi_D)^T$ on its diagonal.*

Using the variational Bayes approach to approximate the posterior distribution with a factorized form

$$q(\Theta) = \prod_{d=1}^{D} q(\mathbf{W}_d) \prod_{n=1}^{N} q(\mathbf{z}_n) \prod_{d=1}^{D} q(\psi_d),$$

find the VB update for the factor $q(\mathbf{W}_d)$. Here $\mathbf{W}_d$ denotes the dth row of the loading matrix $\mathbf{W}$ as a column vector.

[1] Also available at http://se.mathworks.com/help/stats/examples/factor-analysis.html.

Solution.

$$\begin{aligned}
\log q(\mathbf{w}_d) &= \langle \log p(\mathbf{x}|\mathbf{W}, \mathbf{z}, \boldsymbol{\psi}) + \log p(\mathbf{W}) \rangle_{q(\mathbf{z},\boldsymbol{\psi})} \\
&= \left\langle -\frac{1}{2} \sum_n (\mathbf{x}_n - \mathbf{W}\mathbf{z}_n)^T \operatorname{diag}(\boldsymbol{\psi})(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n) - \frac{1}{2}\alpha^{-1}\mathbf{w}_d^T\mathbf{w}_d \right\rangle_q \\
&= \left\langle -\frac{1}{2} \sum_n (\mathbf{z}_n^T\mathbf{W}^T \operatorname{diag}(\boldsymbol{\psi})\mathbf{W}\mathbf{z}_n - 2\mathbf{x}_n^T \operatorname{diag}(\boldsymbol{\psi})\mathbf{W}\mathbf{z}_n) \right\rangle_q - \frac{1}{2}\alpha^{-1}\mathbf{w}_d^T\mathbf{w}_d \\
&= \left\langle -\frac{1}{2} \sum_n \operatorname{tr}(\operatorname{diag}(\boldsymbol{\psi})\mathbf{W}\mathbf{z}_n\mathbf{z}_n^T\mathbf{W}^T) - 2\mathbf{x}_n^T \operatorname{diag}(\boldsymbol{\psi})\mathbf{W}\mathbf{z}_n) \right\rangle_q - \frac{1}{2}\alpha^{-1}\mathbf{w}_d^T\mathbf{w}_d \\
&= \left\langle -\frac{1}{2} \sum_n (\psi_d\mathbf{w}_d^T\mathbf{z}_n\mathbf{z}_n^T\mathbf{w}_d - 2\mathbf{x}_{nd}\psi_d\mathbf{z}_n^T\mathbf{w}_d) \right\rangle_q - \frac{1}{2}\alpha^{-1}\mathbf{w}_d^T\mathbf{w}_d \\
&= -\frac{1}{2}\mathbf{w}_d^T \underbrace{\left(\langle\psi_d\rangle \sum_n \left\langle \mathbf{z}_n\mathbf{z}_n^T \right\rangle + \alpha^{-1}\mathbf{I} \right)}_{\boldsymbol{\Sigma}_d^{-1}} \mathbf{w}_d + \underbrace{\sum_n x_{nd} \langle\psi_d\rangle \left\langle \mathbf{z}_n^T \right\rangle}_{\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\mu}_d} \mathbf{w}_d
\end{aligned}$$

where we see by completing the square that this is a Gaussian with the following parameters

$$\begin{aligned}
q(\mathbf{w}_d) &= \mathcal{N}(\mathbf{w}_d|\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d) \\
\boldsymbol{\Sigma}_d^{-1} &= \langle\psi_d\rangle \sum_n \left\langle \mathbf{z}_n\mathbf{z}_n^T \right\rangle + \alpha^{-1}\mathbf{I} \\
\boldsymbol{\mu}_d &= \boldsymbol{\Sigma}_d \langle\psi_d\rangle \sum_n x_{nd} \langle\mathbf{z}_n\rangle
\end{aligned}$$

Problem 3. *"Factor analysis for digits data."*

Fit a factor analysis model to each of digits '3' and '8' in the digits data set separately, and use the two models to predict the digit labels in test data. Do this by implementing the missing parts in the `ex8_digit_template.m`. Comment on the accuracy of the solution. What ways to improve the accuracy could you use in practice (you should be able to come up with at least a few different ways, but do not need to implement them here)?

Note: pixels without any variation result in a warning due to numerical problems, when inverting the estimated covariance matrix, although the results seem reasonable. What could be done to remove the numerical problem (not required for solution)?

Comment on data: the figures are examples of handwritten digits, which consist of 28×28 *real-valued pixels.*

Solution. See code for solution.

Some ideas to improve our modelling:

- Select number of factors properly (BIC, CV, etc.)

- Normalize/preprocess the data better, detect outliers
- For some purposes consider optimizing precision instead; classify only digits that you are very sure of
- Classify digits written only by one person without trying to generalize
- Consider a different factor model (e.g. introduce sparse loadings)

Non-varying pixels could either be removed or the estimated covariance matrix regularized by adding some (small) constant diagonal matrix $\alpha\mathbf{I}$ to it.