Tutorial on Factor Analysis

Javier R. Movellan
1 Generative Model

\[ O = aH + Z \]  

where \( O \) is an \( n \)-dimensional random vector of observations, \( H \) and \( d \)-dimensional vector of hidden variables and \( Z \) a noise vector. \( H \) is a standard Gaussian vector, i.e., \( H \sim N(0, I_d) \) and \( Z \) is a zero-mean Gaussian vector with diagonal covariance matrix \( \Psi \). The model parameters are the mixing matrix \( a \) and the covariance matrix \( \Psi \).

2 EM Learning

Given \( o^{(1)}, cdots, o^{(s)} \), a fair sample of observations from \( O \), our goal is to values of \( \lambda = \{a, \Psi\} \) that maximize \( \sum \log p(o^{(i)} \mid \lambda) \). To do so within the EM framework\(^1\) we form an auxiliary function

\[ Q(\bar{\lambda}, \lambda) \overset{\text{def}}{=} \sum_i \int p(h \mid o^{(i)} \bar{\lambda}) \log p(o^{(i)} h \mid \lambda) \, dh \]  

we note that

\[ p(o^{(i)} h \mid \lambda) = p(h)p(o^{(i)} \mid h \lambda) \]  

and since \( p(h) \) does not depend on \( \lambda \) we redefine \( Q \) as follows

\[ Q(\bar{\lambda}, \lambda) \overset{\text{def}}{=} \sum_i \int p(h \mid o^{(i)} \bar{\lambda}) \log p(o^{(i)} h \mid \lambda) \, dh \]  

From the definition of the multivariate Gaussian distribution it follows that

\[ Q(\bar{\lambda}, \lambda) = -\frac{1}{2} \sum_i E \left[ \log (2\pi \mid \Psi \mid) + (o^{(i)} - aH)^T \Psi^{-1} (o^{(i)} - aH) \mid o^{(i)} \bar{\lambda} \right] \]  

Taking the gradient with respect to \( a \) and setting it to zero\(^2\) we get

\[ \nabla_a Q(\bar{\lambda}, \lambda) = \Psi^{-1} \sum_i E \left[ (o^{(i)} - aH)H^T \mid o^{(i)} \bar{\lambda} \right] = 0 \]  

Thus

\[ \sum_i o^{(i)} \left(h^{(i)}\right)^T = a \sum_i E \left[ HH^T \mid o^{(i)} \bar{\lambda} \right] \]  

where

\[ h^{(i)} = E \left[ H \mid o^{(i)} \bar{\lambda} \right] = E(H \mid \bar{\lambda}) + \Sigma_{HO} \Sigma_{GO}^{-1} (o^{(i)} - E(O \mid \bar{\lambda})) = a^T (aa^T + \Psi)^{-1} o^{(i)} \]

\[ E \left[ HH^T \mid o^{(i)} \bar{\lambda} \right] = \Sigma_{HH} - \Sigma_{HO} \Sigma_{GO}^{-1} \Sigma_{OH} = I_d - a^T (aa^T + \Psi)^{-1} a + h^{(i)} (h^{(i)})^T \]

Thus,

\[ \hat{a} = \left( \sum_i o^{(i)} b^T (o^{(i)})^T \right) \left( sI_d - sba + \sum_i h^{(i)} (h^{(i)})^T \right)^{-1} \]

where

\[ b = a^T (aa^T + \Psi)^{-1} \]

\(^1\)See the tutorial on EM from the Kolmogorov project

\(^2\)See the tutorial on Matrix Calculus from the Kolmogorov project
which is independent of \( \Psi \). Note this solves the linear regression problem for least-squares prediction of \( o(i) \) based on \( h(i) \). Using the optimal value of \( a \) and gradient with respect to \( \Psi \) we get

\[
\nabla_{\Phi^{-1}} Q(\bar{\lambda}, \bar{a}, \Psi) = \text{diag} \nabla_{\Phi^{-1}} Q(\bar{\lambda}, \bar{a}, \Psi)
\]

\[
= \text{diag} \left[ -\frac{s}{2} \Psi - \frac{1}{2} \sum_{i=1}^{s} E \left[ (o(i) - \bar{a}H)^T (o(i) - \bar{a}H) \mid o(i) \bar{\lambda} \right] \right] = 0
\]

(12)

Thus,

\[
\hat{\Psi} = \frac{1}{s} \sum_{i=1}^{s} \text{diag} \left[ o(i)(o(i))^T - 2\bar{a}(i) \left( \bar{a} E \left[ H^T \mid o(i) \bar{\lambda} \right] \right)^T + \bar{a} E \left[ H H^T \mid o(i) \bar{\lambda} \right] \bar{a}^T \right]
\]

(13)

### 3 Independent Factor Analysis

In ICA one uses superGaussian source distributions, as opposed to Gaussians. A convenient way to get a super-Gaussian sources is to use a mixture of 2 Gaussians with zero mean and variances \( 1, \alpha \). We will define a new random vector \( M = (M_1, \cdots, M_d)^T \) where \( M_i \) takes values in \( \{0, 1\} \). The distribution of \( S \) given \( M = m \) is zero mean Gaussian with variance

\[
\sigma_m = \alpha^2 \text{diag}(m) + (I_m - \text{diag}(m))
\]

(14)

Here we address the case in which \( \lambda = \{a, \Psi\} \). To use the EM algorithm we now need to treat \( H, M \) as hidden variables. We have

\[
p(ohm \mid \lambda) = p(m)p(h \mid m)p(o \mid hm \lambda)
\]

(15)

Thus,

\[
Q(\bar{\lambda}, \lambda) \overset{\text{def}}{=} \sum_i \sum_d \int p(hm \mid o(i) \bar{\lambda}) \left( \log p(m) + \log p(h \mid m) + \log p(o(i) \mid hm \lambda) \right) dh dm
\]

(16)

We can redefine \( Q \) by eliminating the terms that do not depend on \( \lambda \). Thus

\[
Q(\bar{\lambda}, \lambda) \overset{\text{def}}{=} \sum_m \sum_i p(m \mid o(i) \bar{\lambda}) \int p(h \mid o(i) m \bar{\lambda}) \log p(o(i) \mid hm \lambda) dh
\]

(17)

\[
= \sum_m p(m \mid o(i) \bar{\lambda}) Q_m(\bar{\lambda}, \lambda)
\]

(18)

Taking derivatives with respect to \( a \) we get

\[
\nabla_a Q(\bar{\lambda}, \lambda) = \Psi^{-1} \sum_m \bar{p}(m \mid o(i) \bar{\lambda}) \sum_{i=1}^{s} E \left[ (o(i) - aH)^T \mid o(i) \bar{\lambda} \right] = 0
\]

(19)

Thus

\[
\sum_m p(m \mid o(i) \bar{\lambda}) \sum_{i=1}^{s} o(i) \left( E \left[ H \mid o(i) m \bar{\lambda} \right] \right)^T = a \sum_m p(m \mid o(i) \bar{\lambda}) \sum_{i=1}^{s} E \left[ H H^T \mid o(i) m \bar{\lambda} \right]
\]

(20)
where

\[
E [H \mid o^{(i)} m \lambda] = E(H \mid m \lambda) + \Sigma_{HO}^m (\Sigma_{OO}^m)^{-1} (o^{(i)} - E(O \mid m \lambda)) = \sigma_m a^T (a \sigma_m a^T + \Psi)^{-1} o^{(i)} \tag{21}
\]

\[
E [HH^T \mid o^{(i)} m \lambda] = \Sigma_{HH} - \Sigma_{HO}^m (\Sigma_{OO}^m)^{-1} \Sigma_{OH}^m + E [H \mid o^{(i)} m \lambda] \left( E [H \mid o^{(i)} m \lambda] \right)^T
= \sigma_m - \sigma_m a^T (a \sigma_m a^T + \Psi)^{-1} a \sigma_m + E [H \mid o^{(i)} m \lambda] \left( E [H \mid o^{(i)} m \lambda] \right)^T \tag{22}
\]

and

\[
p(m \mid o^{(i)} \lambda) = \frac{p(m)p(o^{(i)} \mid m \lambda)}{\sum_{m'} p(m')p(o^{(i)} \mid m \lambda)} \tag{23}
\]
4 History

- The first version of this document was written by Javier R. Movellan in January 2004, as part of the Kolmogorov project.