National Technical University of Ukraine
“Igor Sikorsky Kyiv Polytechnic Institute”
Institute of Physics and Technology

Lecture 10
Generative Models for Discrete Data. Latent Linear Models

Dmytro Progonov,
PhD, Associate Professor,
Department Of Physics and Information Security Systems

Content

- Problem statement;
- Unidentifiability of FA parameters;
- EM for factor analysis models;
- Principle component analysis:
  - Classical PCA;
  - Probabilistic PCA;
- Singular value decomposition;
- Independent Component Analysis.

Problem statement (1/2)

One problem with *mixture models* is that they *only use a single latent variable to generate the observation*. In particular, each observation can only come from one of K prototypes.

One can think of a *mixture model as* using K hidden binary variables, representing a one-hot encoding of the cluster identity. But because *these variable are mutually exclusive*, the model is still limited in its representational power.

Problem statement (2/2)

An alternative is to use a vector of real-valued latent variables, $\mathbf{z}_i \in \mathbb{R}^L$. The simplest prior to use is a Gaussian:

$$p(\mathbf{z}_i) = \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$

If the observations a re also continuous, so $\mathbf{x}_i \in \mathbb{R}^D$, we may use a Gaussian for the likelihood. Just as in linear regression, we will assume the mean is a linear function of the (hidden) inputs:

$$p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \boldsymbol{\Psi})$$

where $\mathbf{W}$ – is a $D \times L$ ***factor loading matrix***, $\boldsymbol{\Psi}$ – is a $D \times D$ covariance matrix. We take $\boldsymbol{\Psi}$ to be diagonal, since the whole point of the model is to "force" $\mathbf{z}_i$ to explain the correlation, rather than "baking it in" to the observation's covariance.

This overall model is called ***factor analysis*** (***FA***).

Unidentifiability of FA parameters (1/2)

The parameters of an FA model are unidentifiable. To see this, suppose $\mathbf{R}$ is an arbitrary orthogonal rotation matrix, satisfying $\mathbf{R}\mathbf{R}^{-1} = \mathbf{I}$. Let us define $\widetilde{\mathbf{W}} = \mathbf{W}\mathbf{R}$; then the likelihood function of this modified matrix is the same as for the unmodified one, since

$$cov(\mathbf{x}) = \widetilde{\mathbf{W}}\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T]\widetilde{\mathbf{W}}^T + \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \mathbf{W}\mathbf{R}\mathbf{R}^T\mathbf{W}^T + \boldsymbol{\Psi} = \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi}$$

Geometrically, multiplying $\mathbf{W}$ by an orthogonal matrix is like rotating $\mathbf{z}$ before generating $\mathbf{x}$; but since $\mathbf{z}$ is drawn from an isotropic Gaussian, this make no difference to the likelihood.

Consequently, we cannot unique identify $\mathbf{W}$, and therefore cannot uniquely identify the latent factors, either.

Unidentifiability of FA parameters (2/2)

Since factor analysis is often used to uncover structure in data, the problem of uniquely identify the parameters needs to be addressed. There are some commonly used solutions:

1. Forcing $\mathbf{W}$ to be orthogonal;
2. Forcing $\mathbf{W}$ to be lower triangular;
3. Sparsity promotion priors on the weights;
4. Choosing an informative rotation matrix;
5. Use of non-Gaussian priors for the latent factors.

EM for factor analysis models (1/2)

In the *E step*, we compute the posterior responsibility of cluster c for data point i:

$$r_{ic} \triangleq p(q_i = c|\mathbf{x}_i, \boldsymbol{\theta}) \propto \pi_c \mathcal{N}\left(\mathbf{x}_i|\boldsymbol{\mu}_c, \mathbf{W}_c \mathbf{W}_c^T + \boldsymbol{\Psi}\right)$$

The conditional posterior for $\mathbf{z}_i$ is given by

$$p(\mathbf{z}_i|\mathbf{x}_i, q_i = c, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{z}_i|\mathbf{m}_{ic}, \boldsymbol{\Sigma}_{ic})$$

$$\boldsymbol{\Sigma}_{ic} \triangleq \left(\mathbf{I}_L + \mathbf{W}_c^T \boldsymbol{\Psi}_c^{-1} \mathbf{W}_c^T\right)^{-1}$$

$$\mathbf{m}_{ic} \triangleq \boldsymbol{\Sigma}_{ic}\left(\mathbf{W}_c^T \boldsymbol{\Psi}_c^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_c)\right)$$

In the M step, it is easier to estimate $\boldsymbol{\mu}_c$ and $\mathbf{W}_c$ at the same time, by defining $\widetilde{\mathbf{W}}_c = (\mathbf{W}_c, \boldsymbol{\mu}_c), \tilde{\mathbf{z}} = (\mathbf{z}, 1)$. Also, define:

$$\mathbf{b}_{ic} \triangleq \mathbb{E}[\tilde{\mathbf{z}}|\mathbf{x}_i, q_i = c] = [\mathbf{m}_{ic}, 1]$$

$$\mathbf{C}_{ic} \triangleq \mathbb{E}[\tilde{\mathbf{z}}\tilde{\mathbf{z}}^T|\mathbf{x}_i, q_i = c] = \begin{pmatrix} \mathbb{E}[\mathbf{z}\mathbf{z}^T|\mathbf{x}_i, q_i = c] & \mathbb{E}[\mathbf{z}|\mathbf{x}_i, q_i = c] \\ \mathbb{E}[\mathbf{z}|\mathbf{x}_i, q_i = c]^T & 1 \end{pmatrix}$$

EM for factor analysis models (2/2)

Then the *M step* is as follows

$$\widetilde{\mathbf{W}}_c = \left[\sum_i r_{ic}\mathbf{x}_i\mathbf{b}_{ic}{}^T\right]\left[\sum_i r_{ic}\mathbf{C}_{ic}\right]^{-1}$$

$$\boldsymbol{\Psi} = \frac{1}{N} diag\left\{\sum_{ic} r_{ic}\left(\mathbf{x}_i - \widetilde{\mathbf{W}}_c\mathbf{b}_{ic}\right)\mathbf{x}_i{}^T\right\}$$

$$\pi_c = \frac{1}{N}\sum_{i=1}^{N} r_{ic}$$

Principle component analysis (PCA)

Consider the FA model where we constrain $\boldsymbol{\Psi} = \sigma^2 \mathbf{I}$, and $\mathbf{W}$ to be orthonormal. It can be shown, that as $\sigma^2 \to 0$, this model reduces to classical (non-probabilistic) ***principle component analysis*** (***PCA***), also known as the ***Karhunen-Loeve transform***. The version where $\sigma^2 > 0$ is known as ***probabilistic PCA*** (***PPCA***).

Classical PCA (1/2)

Theorem: Suppose we want to find an orthogonal set of L linear basis vectors $\mathbf{w}_j \in \mathbb{R}^D$, and the corresponding scores $\mathbf{z}_i \in \mathbb{R}^L$, such that we minimize the average reconstruction error

$$J(\mathbf{W}, \mathbf{Z}) = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2$$

where $\hat{\mathbf{x}}_i = \mathbf{W}\mathbf{z}_i$, subject to the constrain that $\mathbf{W}$ is orthonormal. Equivalently, we can write this objective as follow:

$$J(\mathbf{W}, \mathbf{Z}) = \|\mathbf{X} - \mathbf{W}\mathbf{Z}^T\|_F{}^2$$

where $\mathbf{Z}$ is an $N \times L$ matrix with the $\mathbf{z}_i$ in its rows, and $\|\mathbf{A}\|_F$ is the Frobenius norm of matrix $\mathbf{A}$, defined by

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}{}^2} = \sqrt{tr\,[\mathbf{A}^T\mathbf{A}]} = \|\mathbf{A}(:)\|_2$$

Classical PCA (2/2)

The optimal solution is obtained by setting $\widehat{\mathbf{W}} = \mathbf{V}_L$, where $\mathbf{V}_L$ contains the L eigenvectors with the largest eigenvalues of empirical covariance matrix, $\widehat{\mathbf{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i{}^T$ (we assume the $\mathbf{x}_i$ have zeros mean, for notation simplicity). Furthermore, the optimal low-dimensional encoding of data is given by $\hat{\mathbf{z}}_i = \mathbf{W}^T \mathbf{x}_i$, which is an orthogonal projection of the data onto the column space spanned by the eigenvectors.

Probabilistic PCA (1/2)

Theorem: Consider a factor analysis model in which $\mathbf{\Psi} = \sigma^2\mathbf{I}$. The observed data log likelihood is given by

$$\log[p(\mathbf{X}|\mathbf{W},\sigma^2)] = \left(-\frac{N}{2}\right)\ln[\mathbf{C}] - \frac{1}{2}\sum_{i=1}^{N}\mathbf{x}_i{}^T\mathbf{C}^{-1}\mathbf{x}_i = \left(-\frac{N}{2}\right)\ln[\mathbf{C}] + tr\left[\mathbf{C}^{-1}\widehat{\mathbf{\Sigma}}\right]$$

where $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$ and $\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i{}^T = \frac{1}{N}\mathbf{X}^T\mathbf{X}$ (we are assuming centered data, for notational simplicity). The maxima of the log-likelihood are given by

$$\widehat{\mathbf{W}} = \mathbf{V}(\mathbf{\Lambda} - \sigma^2\mathbf{I})^{\frac{1}{2}}\mathbf{R}$$

where $\mathbf{R}$ is an arbitrary $L \times L$ orthogonal matrix, $\mathbf{V}$ is the $D \times L$ matrix whose columns are the first L eigenvectors of the $\mathbf{S}$, and $\mathbf{\Lambda}$ is the corresponding diagonal matrix of eigenvalues.

Probabilistic PCA (2/2)

Without loss of generality, we can set $\mathbf{R} = \mathbf{I}$. Furthermore, the MLE of the noise variance s given by

$$\hat{\sigma}^2 = \frac{1}{D-L} \sum_{j=L+1}^{D} \lambda_j$$

which is the average variance associated with the discarded dimensions.

Singular value decomposition (SVD)

We have defined the solution to PCA in terms of eigenvectors of the covariance matrix. However, there is another way to obtain the solution, based on the ***singular value decomposition*** (***SVD***). This basically generalizes the notion of eigenvectors from square matrices to any kind of matrix.

In particular, any (real) $N \times D$ matrix $\mathbf{X}$ can be decomposed as follows

$$\mathbf{X}_{N\times D} = \mathbf{U}_{N\times N}\mathbf{S}_{N\times D}\mathbf{V}_{D\times D}{}^{T}$$

where $\mathbf{U}$ and $\mathbf{V}$ are an matrices whose columns are orthonormal, $\mathbf{S}$ is matrix containing the $r = \min[N, D]$ ***singular values*** $\sigma_i > 0$ on the main diagonal, with 0s filling the rest of matrix.

The columns of $\mathbf{U}$ are the left singular vectors, and the columns of $\mathbf{V}$ is the right singular vectors.

EM for PCA

E step: $\tilde{\mathbf{Z}} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\tilde{\mathbf{X}}$

M step: $\mathbf{W} = \tilde{\mathbf{X}}\tilde{\mathbf{Z}}^T\left(\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^T\right)^{-1}$

where $\tilde{\mathbf{Z}}$ is a $L \times N$ matrix storing the posterior means (low-dimensional representation) along its columns; $\tilde{\mathbf{X}} = \mathbf{X}^T$ store the original data along its columns.

EM for PCA has the following advantages over eigenvector method:

- EM can be faster;
- EM can be implemented in an online fashion, i.e. we can update our estimate of $\mathbf{W}$ as the data streams in;
- EM can handle missing data in a simple way;
- EM can be extended to handle mixtures of PPCA/FA models;
- EM can be modified to variational EM or to variational Bayes EM to fit more complex models;

Independent Component Analysis (ICA)

Our goal is to deconvolve the mixed signals into their constituent parts. This is known as the example of ***blind signal separation*** (***BSS***) or ***blind source separation***, where "blind" means we known "nothing" about the source of the signals.

We can formalize the problem as follows. Let $\mathbf{x}_t \in \mathbb{R}^D$ be the observed signal at the sensor at "time" t, and $\mathbf{z}_t \in \mathbb{R}^L$ be the vector of source signals. We assume that

$$\mathbf{x}_t = \mathbf{W}\mathbf{z}_t + \boldsymbol{\epsilon}_t$$

where $\mathbf{W}$ is an $D \times L$ ***mixing matrix***, and $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$. Without loss of generality, we can constrain the variance of the source distribution to be 1, because any other variance can be modelled by scaling the rows of $\mathbf{W}$ appropriatery.

Criteria for estimate the ICA parameters

Maximum likelihood estimation:

$$NLL(\mathbf{V}) = \sum_{j=1}^{L} \mathbb{E}\left[G_j(z_j)\right] \rightarrow min$$

where $\mathbf{V} = \mathbf{W}^{-1}$ is recognition weights; $z_j \triangleq \mathbf{v}_j{}^T \mathbf{x}$ and $G_j(z) \triangleq \left(-\log\left[p_j(z)\right]\right)$.

Maximizing non-Gaussianity – kurtosis or negentropy:

$$negentropy(z) \triangleq \mathbb{H}[\mathcal{N}(\mu, \sigma^2)] - \mathbb{H}[z]$$

where $\mu = \mathbb{E}[z]$ and $\sigma^2 = var\ [z]$.

Minimizing mutual information:

$$I(\mathbf{z}) \triangleq \mathbb{KL}\left[p(\mathbf{z}) || \prod_j p(z_j)\right] = \sum_j \mathbb{H}\left[z_j\right] - \mathbb{H}[\mathbf{z}]$$

Conclusion

- Problem of unidentifiability of FA parameters was considered;
- EM for factor analysis models was presented;
- Classical and Probabilistic Principle component analysis was described;
- Special types of Factor Analysis, such as Independent Component Analysis, were presented.