Multi-view Factor Analysis

Abstract

Two-view factor analysis finds basis functions for two correlated random vectors using matrix factorization techniques and second order correlation statistics of the two random vectors. We show that the factorization technique can be motivated by a two view topic model under conditional independence assumptions, and the analysis can be used to find hidden topics.

1 Two view factor analysis

Consider two correlated high dimensional random vectors $X \in R^p$ and $Y \in R^q$ (two views). In two view factor analysis, we want to find a matrix $P \in R^{p\times s}$, $Q \in R^{q\times t}$ and $A \in R^{s\times t}$

$$EXY^T \approx PAQ^T.$$

That is, we may solve the following optimization problem:

$$\mathcal{L}_0(PAQ, E_{X,Y}XY^T) + R_0(P, A, Q),$$

where $\mathcal{L}_0$ is a loss function and R_0 is regularization.

Given such matrices (basis functions) P and Q, we can reduce dimension of $X \in R^p$ to $\bar{X} = P^TX \in R^t$. Similarly, we can reduce dimension of $Y \in R^q$ to $\bar{Y} = Q^TY \in R^s$. A more sophisticated dimension reduction technique is to find $\bar{X}$ such that

$$\bar{X} = \arg\min_{x\in\mathcal{X}}[\mathcal{L}_1(Px, X) + R_1(x)],$$

where $\mathcal{L}_1$ is a loss function, $\mathcal{X}$ a subset of R^s (e.g. sparsity), and $R_1(x)$ is a regularization parameter such as L_1 to encourage sparsity.

In factor analysis, a column j of P can be regarded as the representation of cluster (hidden state) j in view-X, and a column j' of Q can be regarded as the representation of cluster (hidden state) j' in view-Y. The matrix $A_{j,j'}$ indicates the correlation between states j and j'. In this interpretation, each component j of $\bar{X}$ is the strength of X in cluster j. A similar interpretation holds for $\bar{Y}$.

We are particularly interested in dimension reduction techniques so that P and Q are sparse. Sparsity implies that each cluster only covers a small number of features. If X (or Y) is sparse, then $\bar{X}$ (or $\bar{Y}$) is also sparse. This means that each observed X (or Y) belongs to a small number of clusters. One may also explicitly require that $\bar{X}$ and $\bar{Y}$ are sparse through $R()$.

2 Relation to two view topic model

We would like to motivate two view factor analysis from probability modeling, and in particular, a two-view analogy of topic modeling. In this model, we consider two additional hidden random variables (W, Z) that are unobserved. The following model is influenced by LDA, and we assume:

- $W \in R^s$ (topic mixture for the first view) is a probability vector with the meaning that X is generated from a mixture W over s topics. Let $p_\ell(X)$ be the probability of topic ℓ, then

$$p(X|W) = \sum_{\ell=1}^{s} W_\ell p_\ell(X)$$

- Conditioned on W, X is independent of Y and Z
- $Z \in R^t$ (topic mixture for the second view) is a probability vector with the meaning that Y is generated from a mixture Z over t topics. Let $q_\ell(Y)$ be the probability of topic ℓ, then

$$p(Y|Z) = \sum_{\ell=1}^{t} Z_\ell q_\ell(Y)$$

- Conditioned on Z, Y is independent of X and W
- W and Z are correlated.

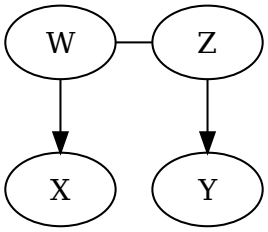

Figure 1: Two View Topic Model

A simplified graphical model representation is given in Figure 1. Under this model, we have

$$E_{X,Y} XY^T = \sum_{\ell,\ell'} E_{W,Z} W_\ell Z_{\ell'} \int X p_\ell(X) \int Y^T q_{\ell'}(Y).$$

Let $P \in R^{p \times s}$ be a matrix, where the ℓ-th column $P[\ell, :] = \int X p_\ell(X)$. Let $Q \in R^{q \times t}$ be a matrix, where the ℓ'-th column is $Q[\ell', :] = \int Y q_{\ell'}(Y)$. Let $A \in R^{s \times t}$ be the topic correlation matrix, where $A_{\ell,\ell'} = E_{W,Z} W_\ell Z_{\ell'}$. Then the above decomposition can be written in matrix notation as

$$E_{X,Y} XY^T = PAQ^T.$$

The matrix A satisfies $A_{i,j} \geq 0$ and $\sum_{i,j} A_{i,j} = 1$, measuring how topics are correlated. If the components of X are $\{0,1\}$ indicators of some events (e.g. whether a word occurs at some position), then $P[\ell,:]$ is the probability of X given cluster ℓ, and we require $P, Q \geq 0$. Moreover, if the events are disjoint and complete (as in the normal setting of topic models), we require $\|P[\ell,:]\|_1 = 1$, $\|Q[\ell',:]\|_1 = 1$.

Therefore in this case we are interested in the matrix factorization problem

$$E_{X,Y} XY^T = PAQ^T, \quad A, P, Q \geq 0, \quad \|A\|_1 = 1, \tag{1}$$

If the events are not disjoint, then the 1-norm constraints on the columns of P and Q do not hold anymore. However, the interpretation of A as topic correlation still holds. In order to reliably identify topic clusters, it is natural to impose additional sparsity constraints on both P and Q. This condition means that a good topic cluster should only involve a small number of features. That is, each cluster is a small group of semantically meaningful and coherent features (words, phrases, etc).

Advantages of the approach

- One does not need to know the concrete form of probability model $P_\ell(X)$ and $P_{\ell'}(Y)$: the method works for arbitrary models without the need to know the form of probability distribution. In comparison, a Bayesian formulation of topic model requires knowing the specific probability distribution such as Dirichlet.

- (This is related to the previous point) One can use arbitrary features that do not have to be independent within each view (e.g. words, word groups, word features like suffix etc). This is unlike LDA-style topic model, which has to work with simple distributions such as Dirichlet.

- One does not need to know the prior over topic mixtures. This is unlike the Bayesian approach which often assumes it to have a certain form. It is possible to avoid assuming the prior using empirical Bayes approach, but computationally impossible to handle because the prior is a distribution on R^{s+t}.

- Consistency: Because the correlation matrix $E_{X,Y} XY^T$ can be estimated reliably when the sample size approaches infinity, as long as we can solve the matrix factorization problem accurately, the solution is consistent in that it recovers the correct factor/distribution.

  In comparison, approximate Bayesian inference for LDA won't be consistent, even if the optimization problem is solved exactly. The only way to be consistent in that approach is to do full Bayesian integration over the hidden variables W and Z and find the prior of W, Z with empirical Bayes. Since $W \in R^s$ and $Z \in R^t$ are distributions over topics, the correct Bayesian approach requires R^{s+t} dimensional integration over an unknown prior on W, Z for each document separately, and then optimize the prior across documents. This is not computationally manageable. The consistency problem should be severe when documents have small number of words, but is not significant if documents have many words. The latter is because W, Z can be reliably determined in such case from many words — the integration over W, Z becomes integration over a delta function, which can be handled with EM.

One View

Since in our approach, it is important that two views are independent given the hidden topic mixture, in order to solve a standard LDA like topic model using this approach, we should turn it into a two-view problem, and use factorization. E.g., we may require the assumption that two set of words are conditionally independent given the topic mixture.

One can expect this approach to work better than Bayesian inference for short documents (or sentence level topic modeling) due to consistency.

Two view factorization with constrained coding

One consider the following optimization:

$$\min_{P,A,Q}\left[\lambda_1 E_X \inf_x[\mathcal{L}_1(Px,X)+R_1(x)]+\lambda_2 E_Y \inf_y[\mathcal{L}_2(Qy,Y)+R_2(y)]+\mathcal{L}_0(PAQ,E_{X,Y}XY^T)+R_0(P,A,Q)\right],$$

where $R_1(x)$ may be used to encourage sparsity of $\bar{X}$, and $R_2(y)$ may be used to encourage sparsity of $\bar{Y}$.

3 Numerical Algorithm

From (1), we obtain for each $1 \le j \le p$ and $1 \le k \le q$ that

$$E_{X_j,Y_k} X_j Y_k = P_j^T A Q_k, \quad A, P_j, Q_k \ge 0,$$

where $P_j^T \in R$ is the j-th row of P, and Q_k^T is the k-th row of Q.

We do not need to use all (j,k), but only some important pairs (probably pairs such that either j is one of the most frequent features among $X_{(\cdot)}$ or k is one of the most frequent features among $Y_{(\cdot)}$). We can then form the following example regression problem

$$[P,A,Q] = \arg\min_{P,A,Q} \sum_{j,k} (P_j^T A Q_k - E X_j Y_k)^2, \quad A,P,Q \ge 0, \quad \|P\|_0 \le u, \|Q\|_0 \le v.$$

One may also use loss functions other than least squares.

Optimization: using alternating least squares.

- Fix P, A, solve for Q
- Fix P, Q, solve for A
- Fix A, Q, solve for P

Each one is an L_1 or L_0 regression problem with positive constraints. To simplify the problem, we may consider diagonal A at first.

4 Some Example Topic Modeling Applications

Here are some ideas for two view factor analysis in text modeling. One may also consider phrase based; or one view are phrases and the other view are words, etc.

- Standard topic modeling (LDA): we may either randomly select two words from each document as two views, or use the word-frequency vector directly for each document and consider self-correlation. The method solves the LDA type topic models, under the additional constraint that $P = Q$ and A is symmetric. In this setting, one may also consider one view as authors and the other view as documents to look at author-topic models, and so on.

  Because of the consistency of the factor analysis approach: this should work much better for short documents (e.g. topic modeling at sentence instead of document level).

- One view is document title and the other view is document body. This model examines how topic structure differs from title to body.

- One view is a sentence, and the other view is the next sentence. This model examines how topic structure changes from sentence to sentence. This can also be combined with dynamic modeling as in Learning HMM.

- One view is a word, and the other is the k-th word in the future. When k is large, this reveals the document-level topic structure. When k is small, this reveals local linguistic topic structure. By plotting across different k, we get different types of topics.