More PCA; and, Factor Analysis

36-350, Data Mining

26 September 2008

Reading: *Principles of Data Mining*, section 14.3.3 on latent semantic indexing.

1 Latent Semantic Analysis: Yet More PCA and Yet More Information Retrieval

Back when I was talking about abstraction, I mentioned that dimension reduction is something that can be layered in between our original representations (like bags of words) and techniques that work in feature space (like similarity searching or nearest-neighbors). That is, rather than looking at the original features, we can apply the same procedures to the reduced, synthetic features we get from doing dimension reduction. This can have advantages in terms of speed (the vectors are smaller), memory (ditto) and even accuracy (since there are fewer parameters, explicit or implicit, to learn).

One particularly nice application of this idea is to combine information retrieval with the use of principal components in dimension reduction. This is called **latent semantic analysis** or **latent semantic indexing**. Remember from last time that the principal components we get from a collection of vectors depend on the covariance across the features. Features which are strongly correlated with each other will have projections on to the principal components which are very close to each other, while features which are weakly correlated or not at all will have nearly or exactly orthogonal projections — they'll project on to different principal components.

Now suppose that the features are words in a big matrix of bag of word vectors. Two words being correlated means that they tend to appear together in the documents, or not at all. But this tendency needn't be absolute — it can be partial because the words mean slightly different things, or because of stylistic differences, etc. But the projections of those features on to the principal components will generally be more similar than the original features are.

To see how this can be useful, imagine we have a collection of documents (a **corpus**), which we want to search for documents about agriculture. It's entirely possible that many documents on this topic don't actually contain the *word* "agriculture", just closely related words like "farming". A simple feature-vector search on "agriculture" will miss them. But it's very likely that the occurrence

of these related words is well-correlated with the occurrence of "agriculture". This means that all these words will have similar projections on to the principal components, so if we do similarity searching on the *images* of the query and the corpus, a search for "agriculture" will turn up documents that use "farming" a lot.

To see why this is latent semantic *indexing*, think about what goes into coming up with an index for a book by hand. Someone draws up a list of topics and then goes through the book noting all the passages which refer to the topic, and maybe a little bit of what they say there. For example, here's the start of the entry for "Agriculture" in the index to Adam Smith's *The Wealth of Nations*:

> Agriculture, the labour of, does not admit of such subdivisions as manufactures, 6; this impossibility of separation, prevents agriculture from improving equally with manufactures, 6; natural state of, in a new colony, 92; requires more knowledge and experience than most mechanical professions, and yet is carried on without any restrictions, 127; the terms of rent, how adjusted between landlord and tenant, 144; is extended by good roads and navigable canals, 147; under what circumstances pasture land is more valuable than arable, 149; gardening not a very gainful employment, 152–3; vines the most profitable article of culture, 154; estimates of profit from projects, very fallacious, *ib.*; cattle and tillage mutually improve each other, 220; ...

and so on. (Agriculture is an important topic in *The Wealth of Nations*.) It's asking a lot to hope for a computer to be able to do something like this, but we could at least hope for a list of pages like "$6, 92, 126, 144, 147, 152 - -3, 154, 220, \ldots$". One could imagine doing this by treating each page as its own document, forming its bag-of-words vector, and then returning the list of pages with a non-zero entry for the feature "agriculture". This will fail: only two of those nine pages actually contains that word, and this is pretty typical. On the other hand, they are full of words strongly correlated with "agriculture", so asking for the pages which are most similar in their principal components projection to that word will work great.[1]

At first glance, and maybe even second, this seems like a wonderful trick for extracting meaning ("semantics") from pure correlations. Of course there are also all sorts of ways it can fail, not least from spurious correlations. If our training corpus happens to contain lots of documents which mention "farming" and "Kansas", as well as "farming" and "agriculture", latent semantic indexing will not make a big distinction between the relationship between "agriculture" and "farming" (which is genuinely semantic) and that between "Kansas" and "farming" (which is accidental, and probably wouldn't show up in, say, a corpus collected from Europe).

[1] Or it should anyway; I haven't actually done the experiment with this book.

Despite this susceptibility to spurious correlations, latent semantic indexing is an *extremely* useful technique in practice, and the foundational papers (Deerwester *et al.*, 1990; Landauer and Dumais, 1997) are worth reading; you can find them on Blackboard or the course website.

2 Factor Analysis

There are two ways to go from principal components analysis to factor analysis — two motivating stories.

Measurement error Suppose that the numbers we write down as our observations aren't precisely accurate — that our numbers are the real variables plus some measurement noise. (Or, if we're not making the measurements ourselves but just taking numbers from some database, that whoever created the database wasn't able to measure things perfectly.) PCA doesn't care about this — it will try to reproduce true-value-plus-noise from a small number of components. But that's kind of weird — why try to reproduce the noise?[2] Can we do something like PCA, where we reduce a large number of features to additive combinations of a smaller number of variables, but which allows for noise?

The simplest model, starting from PCA, would be something like this. Each object or record has p features, so X_{ij} is the value of feature j for object i. As before, we'll center all the observations (subtract off their mean), and for simplicity we'll also standardize them (divide by the standard deviation), so each feature has mean 0 and variance 1 across the data-set. We now postulate that there are k **factor** variables, and each observation is a linear combination of **factor scores** F_{ir} plus noise:

$$X_{ij} = \epsilon_{ij} + \sum_{r=1}^{k} F_{ir} w_{rj} \tag{1}$$

The weights w_{rj} are called the **factor loadings** of the observable features; they say how much feature j changes, on average, in response to a one-unit change in factor score r. Notice that we are allowing each feature to go along with more than one factor (for a given j, w_{rj} can be non-zero for multiple r). This would correspond to our measurements running together what are really distinct variables.

Here ϵ_{ij} is as usual the noise term for feature j on object i. We'll assume this has mean zero and variance ψ_j — i.e., different features has differently-sized noise terms. The ψ_j are known as the **specific variances**, because they're specific to individual features. We'll further assume that $\epsilon_{ij} \perp \epsilon_{lm}$, unless $i = l$, $j = m$ — that is, each object and each feature has independent noise.

We can also re-write the model in vector form,

$$\vec{X_i} = \vec{\epsilon_i} + \vec{F_i}\mathbf{w} \tag{2}$$

[2] One reason would be if we're not sure what's noise, or if what seems to be noise for one purpose is signal for something else. But let's press onward.

with $\mathbf{w}$ being a $k \times p$ matrix. If we stack the vectors into a matrix, we get

$$\mathbf{X} = \epsilon + \mathbf{F}\mathbf{w} \tag{3}$$

This is the factor analysis model. The only (!) tasks are to estimate the factor loadings $\mathbf{w}$, the factor scores $\mathbf{F}$, and the specific variances ψ_j.

A common question at this point is, or should be, where does the model (1) come from? The answer is, we *make it up*. More formally, we *posit* it, and all the stuff about the distribution of the noise, etc., as a *hypothesis*. All the rest of our reasoning is conditional, premised on the assumption that the posited hypothesis is in fact true. It is unfortunately too common to find people who just state the hypothesis in a semi-ritual manner and go on. What we should really do is try to *test* the hypothesis, i.e., to check whether it's actually right. We will come back to this.

Preserving correlations PCA aims to preserve variance, or (what comes to the same thing) minimize mean-squared residuals (reconstruction error). But it doesn't preserve correlations. That is, the correlations of the features of the image vectors are not the same as the correlations among the features of the original vectors (unless $k = p$, and we're not really doing any data reduction). We might value those correlations, however, and want to preserve them, rather than the variance.[3] That is, we might ask for a set of vectors whose image in the feature space will have the same correlation matrix as the original vectors, or as close to the same correlation matrix as possible while still reducing the number of dimensions.

This *also* leads to the factor analysis model, as we'll see.

2.1 Roots of Factor Analysis in Causal Discovery

The roots of factor analysis go back to work by Charles Spearman just over a century ago (Spearman, 1904); he was trying to discover the hidden structure of human intelligence. His observation was that schoolchildren's grades in different subjects were all correlated with each other. He went beyond this to observe a particular *pattern* of correlations, which he thought he could explain as follows: the reason grades in math, English, history, etc., are all correlated is performance in these subjects is all correlated with *something else*, a general or **common** factor, which he named "general intelligence", for which the natural symbol was of course g or G.

Put in a form like Eq. 1, Spearman's model becomes

$$X_{ij} = \epsilon_{ij} + G_i w_j \tag{4}$$

[3] Why? Well, originally the answer was that the correlation coefficient had just been invented, and was about the only way people had of measuring relationships between variables. Since then it's been propagated by statistics courses where it is the only way people are *taught* to measure relationships. The great statistician John Tukey once wrote "Does anyone know when the correlation coefficient is useful? If so, why don't they tell us?"

(Since there's only one common factor, the factor loadings w_j need only one subscript index.) If we assume that the features and common factor are all standardized to have mean 0 and variance 1, and that there is no correlation between ϵ_{ij} and G_i for any j, then the correlation between the j^{th} feature, $X_{\cdot j}$, and G is just w_j. (EXERCISE: Show this.)

Now we can begin to see how factor analysis reproduces correlations. Under these assumptions, it follows that the correlation between the j^{th} feature and the l^{th} feature, call that ρ_{jl}, is just the product of the factor loadings:

$$\rho_{jl} = w_j w_l \tag{5}$$

(EXERCISE: show this.)

Up to this point, this is all so much positing and assertion and hypothesis. What Spearman did next, though, was to observe that this hypothesis carried a very strong implication about the *ratios* of correlation coefficients. Pick any four features, j, l, r, s. Then, if the model (4) is true,

$$\frac{\rho_{jr}/\rho_{lr}}{\rho_{js}/\rho_{ls}} = \frac{w_j w_r / w_l w_r}{w_j w_s / w_l w_s} \tag{6}$$

$$= \frac{w_j/w_l}{w_j/w_l} \tag{7}$$

$$= 1 \tag{8}$$

The relationship

$$\rho_{jr}\rho_{ls} = \rho_{js}\rho_{lr} \tag{9}$$

is called the "tetrad equation", and we will meet it again later when we consider methods for causal discovery.

Spearman found that the tetrad equation held in his data on school grades (to a good approximation), and concluded that a single general factor of intelligence must exist. This was, of course, logically fallacious. (EXERCISE: Why?)

Later work, using large batteries of different kinds of intelligence tests, showed that the tetrad equation does not hold in general, or more exactly that departures from it are too big to explain away as sampling noise. (Recall that the equations are about the true correlations between the variables, but we only get to see sample correlations, which are always a little off.) The response, done in an *ad hoc* way by Spearman and his followers, and then more systematically by Thurstone, was to introduce *multiple* factors. This breaks the tetrad equation, but still accounts for the correlations among features by saying that features are really directly correlated with factors, and uncorrelated conditional on the factor scores.[4] Thurstone's form of factor analysis is basically the one people still use — there have been refinements, of course, but it's mostly still his method.

[4] You can (and should!) read the classic "The Vectors of Mind" paper (Thurstone, 1934) online.

2.2 Preliminaries to Factor Estimation

Assume all the factor scores are uncorrelated with each other and have variance 1; also that they are uncorrelated with the noise terms. We'll solve the estimation problem for factor analysis by reducing it to an eigenvalue problem again.

Start from the matrix form of the model, Eq. 3, which you'll recall was

$$\mathbf{X} = \epsilon + \mathbf{F}\mathbf{w}$$

We know that $\mathbf{X}^T\mathbf{X}$ is a $p \times p$ matrix, in fact it's n times the sample covariance matrix $\mathbf{V}$. So

$$\begin{aligned}
n\mathbf{V} &= \mathbf{X}^T\mathbf{X} && (10)\\
&= (\epsilon + \mathbf{F}\mathbf{w})^T (\epsilon + \mathbf{F}\mathbf{w}) && (11)\\
&= \left(\epsilon^T + \mathbf{w}^T\mathbf{F}^T\right)(\epsilon + \mathbf{F}\mathbf{w}) && (12)\\
&= \epsilon^T\epsilon + \epsilon^T\mathbf{F}\mathbf{w} + \mathbf{w}^T\mathbf{F}^T\epsilon + \mathbf{w}^T\mathbf{F}^T\mathbf{F}\mathbf{w} && (13)\\
&= n\Psi + 0 + 0 + n\mathbf{w}^T I\mathbf{w} && (14)\\
&= n\Psi + n\mathbf{w}^T\mathbf{w} && (15)\\
\mathbf{V} &= \Psi + \mathbf{w}^T\mathbf{w} && (16)
\end{aligned}$$

where Ψ is the diagonal matrix whose entries are the ψ_j. The cross-terms cancel because the factor scores are uncorrelated with the noise, and the $\mathbf{F}^T\mathbf{F}$ term is just n times the covariance matrix of the factor scores, which by assumption is the identity matrix.

At this point, the actual factor scores have dropped out of the problem, and all we are left with are the more "structural" parameters, namely the factor loadings $\mathbf{w}$ and the specific variances ψ_j. We know, or rather can easily estimate, the covariance matrix $\mathbf{V}$, so we want to solve Eq. 16 for these unknown parameters.

The problem is that we want $k < p$, but on its face (16) gives us p^2 equations, one for each entry of $\mathbf{V}$, and only $p + pk$ unknowns, which generally spells trouble. We'll see how to get around this next time, through the wonders of eigenvectors.[5]

References

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer and Richard Harshman (1990). "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science*, **41**: 391–407. URL `http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf`. doi:10.1002/(SICI)1097-4571(199009)41:6¡391::AID-ASI1¿3.0.CO;2-9.

[5] I admit this is not much of a cliff-hanger, but it's the best I could do with the material at hand.

Landauer, Thomas K. and Susan T. Dumais (1997). "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge." *Psychological Review*, **104**: 211–240. URL `http://lsa.colorado.edu/papers/plato/plato.annote.html`.

Spearman, Charles (1904). ""General Intelligence," Objectively Determined and Measured." *American Journal of Psychology*, **15**: 201–293. URL `http://psychclassics.yorku.ca/Spearman/`.

Thurstone, L. L. (1934). "The Vectors of Mind." *Psychological Review*, **41**: 1–32. URL `http://psychclassics.yorku.ca/Thurstone/`.