STA613/CBB540: Statistical methods in computational biology

Factor Analysis (2/21/13)

Lecturer: Barbara Engelhardt *Scribe: Peter Tonner*

1 Probabilistic PCA

This is the probabilistic PCA (PPCA) model presented in [Roweis, 97], [Tipping & Bishop 99]. We have as input $X = \{x_1, x_2, \ldots, x_n\}$ and attempt to model X with the formula:

$$X = \Lambda F + \epsilon$$
$$X_{i,j} \sim \mathcal{N}\left(\sum_{k=1}^{K} \lambda_{i,k} F_{k,j}, \sigma^2\right)$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

With variables representing:

Variable	represntation	dimensions
X	data	$n \times p$
Λ	loadings	$n \times k$
F	factors	$k \times p$
ϵ	noise	$n \times p$

The loadings and factors in combination determine the mean of a specific element of X, $X_{i,j}$. This method is essentially a matrix decomposition and, when k is small, dimensionality reduction, since the high-dimensional matrix X has been approximated in k dimensions.

2 Identifiability

Identifiablity: it is theoretically possible to learn the true parameters of a model with infinite number of observations.

The probabilisitic PCA is unidentifiable. There is not a unique solution for Λ and F with an infinite number of observations in X. This is because multiple factor and factor loading matrices can produce the same data likelihood. This is true for both PPCA and factor analysis. This is caused by two effects: scaling and rotation.

Scaling: $\Lambda F = \frac{\alpha}{\alpha}\Lambda F = (\frac{1}{\alpha}\Lambda)(\alpha F)$
This leads to equivalent solutions to the equation with different values of Λ and F.

Rotation: $\Lambda F = \Lambda A A^{-1} F = \Lambda' F'$
For some $k \times k$ rotation matrix A we have equivalent, but different, Λ' and F'.

The maximum likelihood (ML) estimate will be the same for ΛF scaled & rotated.

3 Factor Analysis

In factor analysis (FA), the input data is modeled with a set of latent variables, F_i $(1 \leq i \leq K)$. Each latent variable is distributed with the standard normal:

$$F \sim N(0, I)$$

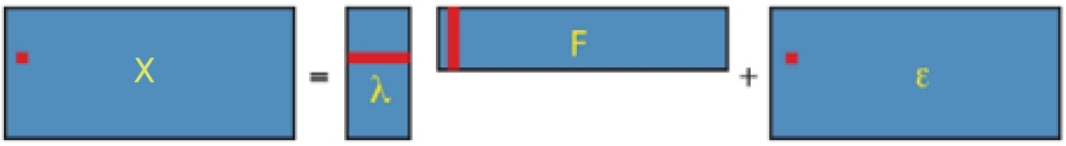
The goal is to capture all the covariance between data points using F, so that

$$Cov(x_i, x_j|F) = 0.$$

We can define a loading matrix Λ $(N \times K)$, where $\Lambda_{i,k}$ determines the weight of each latent variable k in terms of capturing, or being predictive of, an observation x_i. After adding a variance matrix ϵ where $\epsilon_i \sim N(0, \Psi_i)$, we have the familiar model:

$$X = \Lambda F + \epsilon$$

Which can be seen visually:



In this way FA is similar to both PPCA and mixture models. Like mixture models, we model the data with multiple (weighted) Gaussian distributions. However, mixture models make a hard assignment to a specific factor, while FA allows for a data point to come from multiple mixture components with varying weights. Like PPCA, we have factors and a loading matrix, but FA places a distribution on F, and the variance terms in ϵ are not the same for all columns of the original matrix X.

The importance of the model is that $K < \min(n, p)$, meaning there are less dimensions used to describe each factor than there are for the observations. We can approximate the full set of observations by a smaller set of latent (hidden) variables, which reduces computational load and improves interpretability.

We can also think about the likelihood and the posterior probability with respect to the model:

$$P(X|F, \Lambda, \Psi) \sim N(\Lambda F, \Psi)$$

$$P(F|X, \Lambda, \Psi) \sim N(\Lambda F, \Psi)N(0, 1)$$

Because F models the covariance of X, we can integrate F out of the likelihood to get the following:

$$XX^T = \Lambda\Lambda^T + \Psi,$$

which means $\Lambda\Lambda^T + \Psi$ estimates the covariance matrix of X with a low-dimensional representation (specifically, Λ and Ψ, which is a diagonal matrix). We can think of factor analysis, then, as modeling the correlations (or scaled covariance) in the data. In contrast, PCA models directions of maximal variation in the data.

We can also see that as Ψ approaches I, FA becomes PCA (all error variances become the same) up to rotation and scaling.

3.1 Expectation Maximization

We can derive the EM updates by modeling X and F as jointly Gaussian:

$$\begin{bmatrix} X \\ F \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Lambda\Lambda^T + \epsilon & \Lambda \\ \Lambda^T & I \end{bmatrix} \right)$$

In the E step we estimate the factors using $E[F]$ and $E[FF^T]$.

$$E[FF^T] = I - \beta\Lambda + \beta XX^T\beta^T$$

where

$$\beta = \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}$$

The M step maximizes Λ and Ψ with respect to the first and second moments of the factors. This derivation is described nicely in [Ghahramani & Hinton, 1996].

4 *Analysis of Population Structure: A Unifying Framework and Novel Methods Based on Sparse Factor Analysis* (Engelhardt and Stephens, 2010)

In this paper, the relation between sparse factor analysis (SFA), PCA, and admixture models (AM) is shown. Both PCA and AM are generalized by SFA with constraints on the factors and loadings. For the general factor analysis model, we have:

$$X = \Lambda F + \epsilon)$$

4.1 PCA

For PCA, the following constraints are:

$$FF^T = I$$
$$\Lambda\Lambda^T = \gamma I = diag(\gamma)$$
$$\psi_i = \psi$$

In other words, F is orthonormal, Λ is orthogonal, and all variances are the same. From the SVD of X,

$$X = U\Sigma V^T$$

Λ can be constructed from the first k columns of U, and F from the first k rows of ΛV^T

4.2 Admixture Models

The admixture model is:

$$X_{i,j} \sim Bin(2, r_{i,j})$$
$$r_{i,j} = \sum_{k=1}^{K} \Lambda_{i,k} 2 P_{k,j}$$

Where $\Lambda_{i,k}$ is the admixture proportion for individual i in population k and $P_{k,j}$ is the allele frequency in population k for locus j.

The constraints are:

- Λ is non-negative, and rows sum to one
- F is defined by $P_{k,j}$, $0 \leq P_{k,j} \leq 1$.

4.3 Automatic Relevance Determination Prior

The automatic relevance determination prior (ARD) is used to make SFA identifiable by enforcing sparsity in the Λ matrix.

We define:

$$\Lambda_{i,k} \sim \mathcal{N}(0, \sigma^2_{i,k})$$
$$\sigma^2_{i,k} \sim IG(\alpha, \beta)$$

Where IG is inverse gamma. This is equivalent to a student's t distribution. For small $\sigma^2_{i,k}$, the distribution of $\Lambda_{i,k}$ is approximately a point mass around zero; for larger $\sigma^2_{i,k}$ the value of $\Lambda_{i,k}$ is also regularized. In the paper, they modified this prior slightly, but it had the same effect.

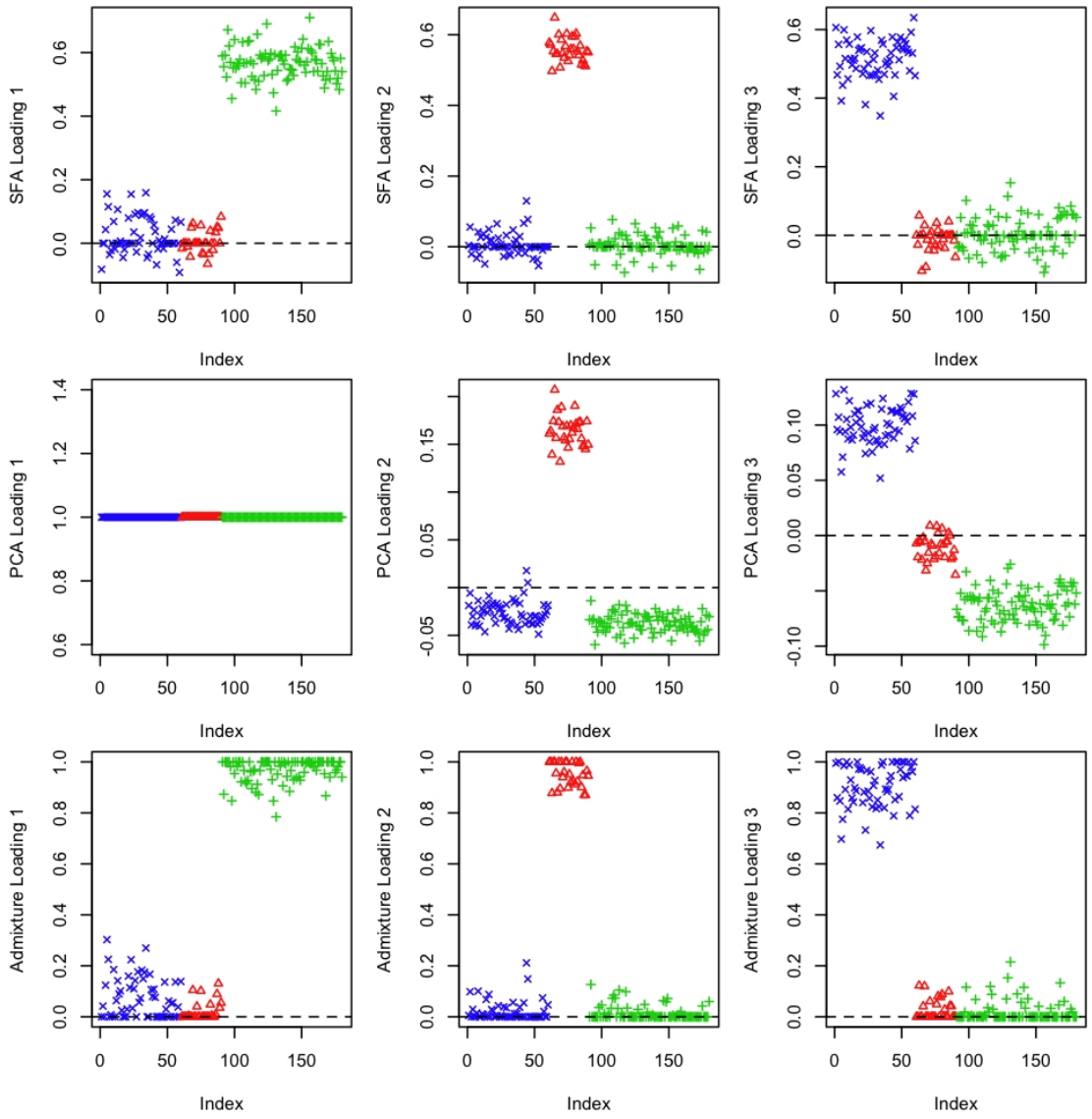
4.4 Expectation Conditional Maximization Either

The model is fit using the Expectation Conditional Maximization Either (ECME) algorithm. Briefly, this is similar to EM but parameters are updated with either the expected log likelihood or the marginal log likelihood. In this paper, μ,F, and Ψ are updated using the expected log likelihood, and Σ is updated using the log marginal likelihood.

4.5 Application to data

4.5.1 HapMap

Application of the three models (PCA, SFA, and AM) to HapMap data for 210 Europeans, Africans, and Asians is shown below. This is an example of discrete population structure, so that we expect three distinct clusters of the individuals based on ancestral populations.

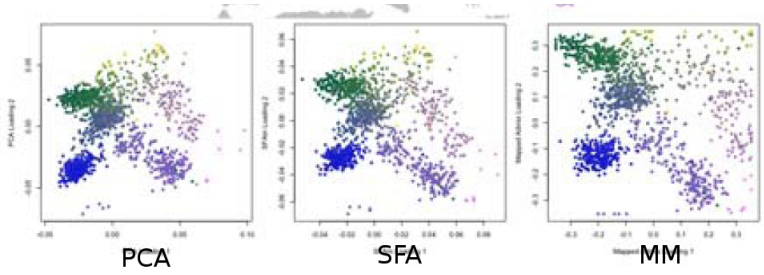


In the figure, SFA is in the first row, PCA in the second, and AM in the third. The results show for each model similar clustering of the three genetic lineages. Also note that each model type has slightly different structure to the output. AM can only model probabilities for each ancestral population between zero and one. Also note that SFA and PCA do not necessarily range from zero to one, and for this model the relative distance between points is more important than absolute distance. Plotting PC1 versus PC2 shows us three distinct clusters; however the corresponding F matrix does not contain allele frequencies for a single ancestral population, but a linear combination of ancestral populations.

4.5.2 Europe

All three modeling techniques (PCA, SFA, and AM) were applied to genotype data of 197,146 SNPs for 1,387 individuals from Europe.

The results can be seen for the three models:



The PCA model is shown only with rotation of the first two PCs to better recognize Europe's shape. SFA first subtracted the mean of the data to center the display at a specific point. AM requires 4 factors to appropriately display the 2d layout. This can be seen by recognizing that in order to describe admixture between two ancestral populations (in this multinomial setting), we need to model two ancestral populations, where the admixed individual falls somewhere on the line connecting the two ancestral populations. Adding a third factor creates a triangular manifold, and then adding a fourth allows us to consider individuals on a square (the corners of the convex space).

5 References

1. Figures taken from Engelhardt and Stephens, 2010
2. *Machine Learning - A Probabilistic Perspective*, Kevin P. Murphy 2012