Distinguishing independent factors from clusters

December 3, 2018

Abstract

It can be difficult to discern whether variation in a population is due to independent interacting factors, or because there are multiple subtypes within the population with their own patterns of variation. We propose using probabilistic generative models to evaluate whether it is more likely that any given dataset is generated by independent factors, as modeled by the Independent Factor Analysis model, or a cluster-based model, in the form of a Gaussian mixture model.

1 Factor and Mixture Models

Factor models are a very common choice for modeling structured variation, and in particular for identifying orthogonal or independent factors describing different modes of variation. In this section we will use the probabilistic interpretation of this class of models.

In the most general setting, we assume N observations $\boldsymbol{y}_n \in \mathbb{R}^D$ are generated by the projection of a latent source vector $\boldsymbol{s}_n \in \mathbb{R}^S$ via a mixing matrix $\boldsymbol{A} \in \mathbb{R}^{D\times S}$ and the addition of Gaussian noise with covariance matrix $\boldsymbol{\Lambda} \in \mathcal{S}_+^{S\times S}$ as,

$$\boldsymbol{y}_n|\boldsymbol{s}_n, \boldsymbol{A}, \boldsymbol{\Lambda} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{s}_n, \boldsymbol{\Lambda}). \tag{1}$$

By imposing different constraints and priors on $\boldsymbol{s}_n$, $\boldsymbol{A}$ and $\boldsymbol{\Lambda}$, we sweep out a number of well-known models. If $\boldsymbol{s}_n \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, then we can marginalize out the source analytically, resulting in a Gaussian marginal density for $\boldsymbol{y}_n$ of form,

$$\boldsymbol{y}_n|\boldsymbol{A}, \boldsymbol{\Lambda} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Lambda}). \tag{2}$$

If we assume that $\boldsymbol{\Lambda}$ is diagonal, we get what is known as the probabilistic factor analysis model (PFA) in the literature. If we assume that it is not just diagonal, but a scaled identity like $\boldsymbol{\Lambda} = \lambda^2\boldsymbol{I}$, we retrieve the probabilistic principal component analysis model (PPCA).

Models with this type of covariance structure describe densities with ellipsoidal contours, with the factors U of their eigen-decomposition $UVU^\top = \boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Lambda}$ describing the axes of the ellipsoid, and the eigenvalues V determining their respective lengths. If $\boldsymbol{\Lambda}$ goes to 0, the density concentrates in the low-dimensional hyperplane spanned by the column vectors of $\boldsymbol{A}$.

Since these densities remain ellipsoidal, it might be necessary to assume a more complicated density than a standard normal on the source vector to adequately model the data. Since we can marginalize over Gaussians, a convenient choice is a mixture of Gaussians defined as,

$$\boldsymbol{s}_n|\{w_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}_{c=1}^C \sim \sum_{c=1}^C w_c \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \tag{3}$$

where in a slight abuse of notation, we have used the sum over distributions to describe a distribution with the mixture density,

$$p(\boldsymbol{s}_n|\{w_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}_{c=1}^C) = \sum_{c=1}^C w_c \frac{|\boldsymbol{\Sigma}_c|^{-1/2}}{(2\pi)^{S/2}} e^{-\frac{1}{2}(\boldsymbol{s}_n - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1}(\boldsymbol{s}_n - \boldsymbol{\mu}_c)}. \tag{4}$$

Using this density on the source space leads to a model we will call the projected mixture of Gaussians (projMoG), which is related to the so-called mixture of factor analyzers (MoFA), but is comparatively weaker as it does not allow the mixing matrix $\boldsymbol{A}$ to differ between the clusters. Instead, it describes a standard Gaussian mixture that is restricted to the hyperspace spanned by $\boldsymbol{A}$,

$$\boldsymbol{y}_n|\boldsymbol{A}, \boldsymbol{\Lambda}, \{w_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}_{c=1}^C \sim \sum_{c=1}^C w_c \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu}_c, \boldsymbol{A}\boldsymbol{\Sigma}_c\boldsymbol{A}^\top + \boldsymbol{\Lambda}). \tag{5}$$

The projected mixture is a good fit for modeling cluster-based variation, but does not model independent variation along the factors specified by $\boldsymbol{A}$ since its Gaussian components span over all the dimensions of the subspace. To constrain the model further, we can assume that each element of the source vector is drawn from its own univariate mixture,

$$s_{ns}|\{\hat{w}_{sq}, \hat{\mu}_{sq}, \hat{\sigma}_{sq}^2\}_{q=1}^Q \sim \sum_{q=1}^Q \hat{w}_{sq} \mathcal{N}(\hat{\mu}_{sq}, \hat{\sigma}_{sq}^2). \tag{6}$$

By multiplying together these univariate densities, we can find the corresponding multivariate mixture of a form like in equation (3) but with parameters,

$$w_c = \prod_{s=1}^S \prod_{q=1}^Q \hat{w}_{sq}^{\xi_c^{sq}}, \quad \mu_{cs} = \sum_{q=1}^Q \xi_c^{sq} \hat{\mu}_{sq}, \quad \sigma_{cs}^2 = \sum_{q=1}^Q \xi_c^{sq} \hat{\sigma}_{sq}^2. \tag{7}$$

Here, the clusters indexed by c are associated with the $C = Q^S$ combinatorial combinations of the components in the sub-mixtures, with ξ_c^{sq} equal to 1 if the q'th component of the mixture governing factor s is involved in cluster c, or 0 otherwise.

This model is known as independent factor analysis (IFA) and we further define the centered IFA (cIFA) as being the variant where $\mu_{sq} = 0$ for all subcomponents. These models are weaker than the projected Gaussian mixture

as they are limited to a factorial mixture on the sources. On the other hand, they are closely related to models like the (probabilistic) independent component analysis model (ICA) which nominally employs independent non-Gaussian source distributions, such as the Laplace distribution. Many such non-Gaussian densities can be modeled or approximated by a univariate mixture of Gaussians like in equation (6), making IFA a tractable alternative to such models (although the $C = Q^S$ scaling means that factors and sub-components have to be limited in number). While Gaussian source distributions induce a Gaussian density in the observation space, ICA is appropriate for modeling data where only a few modes of variation are expressed for any one observation, such as data that concentrates along multiple one-dimensional subspaces in a star-like configuration.

A Appendix

A.1 Calibration

Let use first define the means and covariances of each induced cluster component on the observation space following the mixture in equation (5),

$$\bar{\boldsymbol{\mu}}_c = \boldsymbol{A}\boldsymbol{\mu}_c, \quad \bar{\boldsymbol{\Sigma}}_c = \boldsymbol{A}\boldsymbol{\Sigma}_c\boldsymbol{A}^\top + \boldsymbol{\Lambda} \tag{8}$$

We can calculate the first two moments of all the models we have discussed so far, which can be described by the general form of equation (5),

$$\mathbb{E}\big[\boldsymbol{y}_n|\boldsymbol{A}, \boldsymbol{\Lambda}, \{w_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}_{c=1}^C\big] = \sum_{c=1}^C w_c \bar{\boldsymbol{\mu}}_c$$

$$\mathbb{E}\big[\boldsymbol{y}_n\boldsymbol{y}_n^\top|\boldsymbol{A}, \boldsymbol{\Lambda}, \{w_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}_{c=1}^C\big] = \sum_{c=1}^C w_c \left(\bar{\boldsymbol{\Sigma}}_c + \bar{\boldsymbol{\mu}}_c\bar{\boldsymbol{\mu}}_c^\top\right)$$

Using the tower property, we can then get the first and second moment of the marginal distribution over $\boldsymbol{y}$ by taking the expectation with respect to the remaining parameters, which are all independent.

$$\mathbb{E}[\boldsymbol{y}_n] = \sum_{c=1}^C \mathbb{E}[w_c]\,\mathbb{E}[\boldsymbol{A}]\,\mathbb{E}[\boldsymbol{\mu}_c] \tag{9}$$

$$\mathbb{E}\big[\boldsymbol{y}_n\boldsymbol{y}_n^\top\big] = \sum_{c=1}^C \mathbb{E}[w_c]\left(\mathbb{E}\big[\boldsymbol{A}(\mathbb{E}[\boldsymbol{\Sigma}_c] + \mathbb{E}\big[\boldsymbol{\mu}_c\boldsymbol{\mu}_c^\top\big])\boldsymbol{A}^\top\big] + \mathbb{E}[\boldsymbol{\Lambda}]\right) \tag{10}$$

These expectations reduce to means and second moments of atomic random variables, except for the expectation involving the two mixing matrices $\boldsymbol{A}$ which is confounded by the interspersed factor.

We simplify the above expression considerably by first assuming that $\mathbb{E}[\boldsymbol{\mu}_c] = 0$, which is reasonable for centered data and sets $\mathbb{E}[\boldsymbol{y}_n] = \boldsymbol{0}$, which in turn identifies the covariance of $\boldsymbol{y}_n$ with its second moment. We then set the covariances

to be scaled identities, so $\mathbb{E}[\boldsymbol{\Sigma}_c] = m_{\sigma^2}\boldsymbol{I}$, $\mathbb{E}\left[\boldsymbol{\mu}_c\boldsymbol{\mu}_c^\top\right] = m_\mu\boldsymbol{I}$, and $\mathbb{E}[\boldsymbol{\Lambda}] = m_{\lambda^2}\boldsymbol{I}$, which is good if the data is whitened and appropriate if the correlation structure is unknown a priori. Finally, we take $\mathbb{E}[w_c] = 1/C$ which assumes that no preference is given to any cluster. Then,

$$\mathrm{Cov}(\boldsymbol{y}_n) = (m_{\sigma^2} + m_\mu)\,\mathbb{E}\left[\boldsymbol{A}\boldsymbol{A}^\top\right] + m_{\lambda^2}\boldsymbol{I} \tag{11}$$

All of the factor-based models we consider default to a $[\boldsymbol{A}]_{ij} \sim \mathcal{N}(0,1)$ prior, which results in the covariance

$$\mathrm{Cov}(\boldsymbol{y}_n) = \left(S(m_{\sigma^2} + m_\mu) + m_{\lambda^2}\right)\boldsymbol{I} \tag{12}$$

We will use the scalar *average total variance*

$$\frac{1}{D}\,\mathrm{Tr}\{\mathrm{Cov}(\boldsymbol{y}_n)\} = S(m_{\sigma^2} + m_\mu) + m_{\lambda^2}, \tag{13}$$

as a calibration measure, with it roughly corresponding to the average variance along a dimension. As a standard Gaussian will have an average total variance of 1, we use this as our calibration target, with the noise-level as an independent tuning parameter, resulting in the calibration relations

$$m_\mu + m_{\sigma^2} = \frac{1 - m_{\lambda^2}}{S}. \tag{14}$$

A.1.1 Centered Independent Factor Analysis

Since the cIFA is zero-mean, we have that $m_\mu = 0$, which through the calibration relation gives an immediate rule for setting the mean of $\boldsymbol{\Sigma}$. If we put identical priors on all the $\hat{\sigma}^2$ then we should set the prior's mean $m_{\hat{\sigma}^2}$ as per the rule,

$$m_{\hat{\sigma}^2} = \frac{1 - m_{\lambda^2}}{S}. \tag{15}$$

A consequence of this would be that many of the univariate mixtures would have very similar components, resulting in roughly Gaussian source distributions. To avoid this, we impose a different prior on the first cluster of each univariate source mixture with mean $\rho m_{\hat{\sigma}^2}$ for $\rho \in (0, 1]$, and then $m_{\hat{\sigma}^2}$ for the remaining clusters.

Since the clusters in this modified cIFA no longer have identical covariance, we recognize the alternative definition $m_{\sigma^2} = \frac{1}{C}\sum_{c=1}^{C}\mathbb{E}[\boldsymbol{\Sigma}_c]$ in the calibration formula, and then calculate $m_{\sigma^2} = (1 - \frac{1-\rho}{Q})m_{\hat{\sigma}^2}\boldsymbol{I}$, leading to the calibration relation,

$$m_{\hat{\sigma}^2} = \frac{1 - m_{\lambda^2}}{S(1 - Q^{-1}(1-\rho))}. \tag{16}$$

A.1.2 Projected Gaussian Mixture

To calibrate the projected Gaussian mixture, we first introduce a tuning parameter ρ that controls how much of the remaining variance should be explained by

the cluster means, as opposed to the component variance, yielding the calibration relations,

$$m_\mu = \rho \frac{1 - m_{\lambda^2}}{S}, \quad m_{\sigma^2} = (1 - \rho) \frac{1 - m_{\lambda^2}}{S}, \tag{17}$$

which jointly observe the original calibration relation of equation (14).

In the conjugate setting, $\boldsymbol{\Sigma}_c$ follows an inverse Wishart $\mathcal{IW}(\nu, \kappa \boldsymbol{I})$ with degrees of freedom ν and scale matrix $\kappa \boldsymbol{I}$, and mean,

$$m_{\sigma^2} = \frac{\kappa}{\nu - S - 1} \boldsymbol{I}. \tag{18}$$

Since this leaves us with one constraint and two variables, we further note that $[\boldsymbol{\Sigma}_q]_{ii}$ is distributed as an inverse gamma,

$$[\boldsymbol{\Sigma}_q]_{ii} \sim \text{IG}\left(\frac{\nu - K + 1}{2}, \frac{\kappa}{2}\right) \tag{19}$$

with mean $\frac{\kappa}{\nu - K - 1}$. Using this relationship, we can find parameters that lead to an appropriate marginal distribution by tuning a new inverse Gamma $\text{IG}(\alpha, \beta)$ to have mean $\frac{1 - m_{\lambda^2}}{2K}$ and appropriate variance following section A.1.3, and then setting

$$\nu = K + 2\alpha - 1, \quad \kappa = 2\beta, \tag{20}$$

to ensure marginals with the same properties.

A.1.3 Calibrating variances of inverse Gamma distributions

For an inverse Gamma distribution $\text{InvGamma}(\alpha, \beta)$ the mean and variance does not exist unless $\alpha > 2$, so we consider a random variable $X \sim \text{InvGamma}(2 + \alpha, \beta)$. X then has mean and variance equal to

$$\mathbb{E}[X] = \frac{\beta}{\alpha + 1} \tag{21}$$

$$\text{Var}(X) = \frac{\beta^2}{\alpha(\alpha + 1)^2} \tag{22}$$

If we set $\mathbb{E}[X] = m$ and $\text{Var}(X) = v$, then we can isolate the alpha and beta parameters in terms of the mean and variance as,

$$\beta = m\left(1 + \frac{m^2}{v}\right), \quad \alpha = \frac{m^2}{v} \tag{23}$$

Employing Markov's inequality, we can bound the tail probability using both the mean and the variance as

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}, \quad \mathbb{P}[X \geq a] \leq \frac{\text{Var}(X)}{(a - \mathbb{E}[X])^2}. \tag{24}$$

Assume that the mean is set equal to m. Then if we want less than t mass in the tails, we can select the variance to be

$$\text{Var}(X) = t(a - m)^2 \tag{25}$$