Non-parametric Exploratory Clustering using TURN-RES (nectr)

May 23, 2015

1 MAP Estimation of Gaussian Mixture Model

This document is not intended to be a tutorial on deriving the update equations for a Gaussian Mixture Model (GMM). However, we aim to be transparent with regard to the content of the package. While the algorithm for TURN-RES is available in (Foss, 2002 [1]), fitting the GMM is not done using the standard Maximum Likelihood update equations, but using a MAP estimate. The Bayesian machinery available in MAP estimates is particularly useful for this package, since the clusters discovered using TURN-RES can be used as priors in a sort of Empirical Bayes approach. As one might suspect, the MAP update equations are very similar to the ML equations, but the derivation may be helpful to some. Those not familiar with fitting latent variable models are invited to read Bishop [2], chapter 9, while the free energy formulation can be found in Neal & Hinton, 1998 [3].

(a) The Free Energy for MAP estimation

Free energy is defined below for observed x and latent z as a lower bound on Maximum Likelihood:

$$\mathcal{F}(q,\theta) := \int q(z) \log \frac{p(x,z|\theta)}{q(z)} dz$$

As a lower bound on MAP, up to a constant of proportionality:

$$\mathcal{F}_{\text{MAP}}(q,\theta) := \int q(z) \log \frac{p(x,z|\theta)p(\theta)}{q(z)} dz$$

The E and M steps are quickly derived and shown to be similar to ML:

$$\begin{aligned}
\mathcal{F}_{\text{MAP}}(q,\theta) &= \int q(z) \log \frac{p(z|x,\theta)p(x|\theta)p(\theta)}{q(z)} dz \\
&= \int q(z) \log \frac{p(z|x,\theta)}{q(z)} dz + \int q(z) p(x|\theta) p(\theta) dz \\
&= KL\left(q(z)||p(z|x,\theta)\right) + \ell_{\text{MAP}}(\theta)
\end{aligned}$$

Thus the E-step is maximised when we take $q(z)$ to be the posterior over the latents and the M-step is a maximisation over $\mathbb{E}_{q(z)}\left[p(x,z|\theta)p(\theta)\right]$

(b) GMM Model

Likelihood term:

$$p(x|\pi_{1:K}, \mu_{1:K}, \Sigma_{1:K}) = \sum_z \prod_{k=1}^{K} \pi_k^{z_k} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k}$$

z is coded as a 1-of-K vector. We use conjugate priors to ensure tractable inference:

$$\begin{aligned} &p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\alpha}) && \text{choose:} \quad \alpha_k = \alpha_0 \tilde{\pi}_k \\ &p(\mu_k|\Sigma_k, \beta_k) = \mathcal{N}(\mu_k|m_k, \beta_k^{-1}\Sigma_k) && \quad\quad\; m_k = \tilde{\mu}_k \\ &p(\Sigma_k|\nu_k, \Psi_k) = \text{InvWish}(\Sigma_k|\Psi_k, \nu_k) && \quad\quad\; \Psi_k = \tilde{\Sigma}_k \end{aligned}$$

where $\tilde{\pi}_k, \tilde{\mu}_k, \tilde{\Sigma}_k$ are the empirical estimates of the mixing proportions, means and covariances of each cluster. $\alpha_0, \beta_{1:K}, \nu_{1:K}$ are tuning parameters to adjust the strength of the prior. The joint distribution is:

$$p(x_{1:N}, z_{1:N}, \pi_{1:K}, \mu_{1:K}, \Sigma_{1:K}) = \left(\prod_{k=1}^{K} \left(\prod_{n=1}^{N} \pi_k^{z_{nk}} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{nk}} \right) \mathcal{N}(\mu_k|m_k, \beta_k^{-1}\Sigma_k) \text{InvWish}(\Sigma_k|\Psi_k, \nu_k) \right) \text{Dir}(\boldsymbol{\alpha})$$

$$\begin{aligned} &\propto \prod_{k=1}^{K} \left(\prod_{n=1}^{N} \pi_k^{z_{nk}} \left(\log |2\pi\Sigma_k|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(x_n - \mu_k)^\top \Sigma_k^{-1}(x_n - mu_k) \right\} \right)^{z_{nk}} \right) \\ &\quad \times \log |2\pi\beta_k^{-1}\Sigma_k|^{-\frac{1}{2}} \exp\left\{ -\frac{\beta_k}{2}(\mu_k - m_k)^\top \Sigma_k^{-1}(mu_k - m_k) \right\} \\ &\quad \times |\Sigma_k|^{-\frac{1}{2}(\nu_k + d + 1)} \exp\left\{ -\frac{1}{2}\text{Tr}\left((\nu_k \Psi_k) \Sigma_k^{-1} \right) \right\} \quad \pi_k^{\alpha_k - 1} \end{aligned}$$

Note the particular form of the Inverse Wishart, where the parameter ν_k also multiplies the prior matrix. This is chosen to result in a more interpretable choice of the pair (ν_k, Ψ_k)

(c) E-step for MAP GMM

This is identical to the E-step for maximum likelihood,

$$r_{nk} := q(z_{nk}) \;\; \propto \;\; \pi_k^{z_{nk}} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{nk}} \quad \Rightarrow \quad r_{nk} = \frac{\pi_k^{z_{nk}} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{nk}}}{\sum_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{nk}}}$$

(d) M-step for MAP GMM

Estimate for $\hat{\pi}_k$:

$$\begin{aligned} \log \mathbb{E}_z \left[p(x, z, \pi) \right] &\propto \left(\sum_{n=1}^{N} r_{nk} \log \pi_k \right) + (\alpha_k - 1) \log \pi_k \\ \mathcal{L}(\boldsymbol{\pi}, \lambda) : &= \sum_k (\alpha_k - 1 + \sum_n r_{nk}) \log \pi_k \quad - \lambda (\sum_k \pi_k - 1) \\ \frac{\partial \mathcal{L}}{\partial \pi_k} &= \frac{N_k + \alpha_k - 1}{\pi_k} - \lambda \quad \stackrel{!}{=} 0 \end{aligned}$$

solving for λ in the usual way gives:

$$\hat{\pi}_k = \frac{N_k + \alpha_k - 1}{\sum_k (N_k + \alpha_k - 1)} = \frac{N_k + \alpha_k - 1}{N - K + \sum_k \alpha_k} \tag{1}$$

where $N_k = \sum_n r_{nk}$.

Estimate for $\hat{\mu}_k$:

$$\log \mathbb{E}_z \left[p(x, z, \mu_k)\right] \propto -\frac{1}{2}\left(\left(\sum_{n=1}^{N} r_{nk}(x_n - \mu_k)^\top \Sigma_k^{-1}(x_n - \mu_k)\right) + \beta_k(\mu_k - m_k)^\top \Sigma_k^{-1}(\mu_k - m_k)\right)$$

$$\frac{\partial}{\partial \mu_k} = -\left(\sum_{n=1}^{N} r_{nk}\Sigma_k^{-1}\mu_k - r_{nk}\Sigma_k^{-1}x_n\right) - \beta_k\left(\Sigma_k^{-1}\mu_k - \Sigma_k^{-1}m_k\right) \stackrel{!}{=} 0$$

$$\Rightarrow \quad \hat{\mu}_k = \frac{\sum_n r_{nk}x_n + \beta_k m_k}{N_k + \beta_k} \tag{2}$$

Estimate for $\hat{\Sigma}_k$:

$$\log \mathbb{E}_z \left[p(x, z, \mu_k, \Sigma_k)\right] \propto -\frac{1}{2}\left(\left(\sum_{n=1}^{N} r_{nk}\log|2\pi\Sigma_k| + r_{nk}(x_n - \mu_k)^\top \Sigma_k^{-1}(x_n - \mu_k)\right) + \right.$$
$$\left. \log|2\pi\beta_k^{-1}\Sigma_k| + \beta_k(\mu_k - m_k)^\top \Sigma_k^{-1}(\mu_k - m_k) + (\nu_k + d + 1)\log|\Sigma_k| + \operatorname{Tr}\left(\nu_k\Psi_k\Sigma_k^{-1}\right)\right)$$
$$\propto (N_k + \nu_k + d + 2)\log|\Sigma_k| + \operatorname{Tr}\left(\sum_n r_{nk}(x_n - \mu_k)(x_n - \mu_k)^\top \Sigma_k^{-1}\right) +$$
$$\beta_k \operatorname{Tr}\left((\mu_k - m_k)(\mu_k - m_k)^\top \Sigma_k^{-1}\right) + \operatorname{Tr}\left(\nu_k\Psi_k\Sigma_k^{-1}\right)$$

$$\frac{\partial}{\partial \Sigma_k^{-1}} = -(N_k + \nu_k + d + 2)\,\Sigma^\top + \left(\sum_n r_{nk}(x_n - \mu_k)(x_n - \mu_k)^\top + (\mu_k - m_k)(\mu_k - m_k)^\top + \nu_k\Psi_k\right)$$

$$\Rightarrow \quad \hat{\Sigma}_k = \frac{\sum_n r_{nk}(x_n - \mu_k)(x_n - \mu_k)^\top + \beta_k(\mu_k - m_k)(\mu_k - m_k)^\top + \nu_k\Psi_k}{N_k + \nu_k + d + 2} \tag{3}$$

(e) Notes on Parameter Choices

We still have a number of hyperparameters to choose even though TURN-RES can give us estimates of the sufficient statistics. We will treat each in turn:

1. **Dirichlet distribution over $\boldsymbol{\pi}$.** We can match the first moments of the dirichlet with the ML estimates from TURN-RES. This gives us:
$$\alpha_k = \frac{\sum_n \delta\left(x_n \in \text{cls}(k)\right)}{N}$$
ie the empirical proportions discovered in the TURN-RES stage. However, this gives us only the mean of the prior - we must identify the *strength* (or covariance) of the prior. One option would be to set a baseline minimum variance, set to avoid sparse distributions given by $\alpha_k < 1$ for some k. Thus we choose $a = (\min_k \alpha_k)^{-1}$ and $\alpha_k = aN_k/N$. Another option might be to match the second moments and solve for a. Unfortunately this is non-trivial, not least because the covariance matrix of a Dirichlet is singular by construction. We implement the first option since it is not obvious that we always wish to reduce variance in π with increasing dataset size, for instance due to multiscale issues with TURN-RES.

2. **Gaussian distribution over $\boldsymbol{\mu_k}$.** We are using a Normal Inverse Wishart prior, and thus the covariance of the means is the covariance of the component, Σ_k. However, while we cannot change the covariance

per se, there is a hyperparameter β_k that can be used to scale the covariance in line with our beliefs about μ_k. Note the dependence on β_k in (2) and (3). In (2), we see that the mean estimator can be understood as a convex combination of the ML estimate and the prior. Taking $\beta_k \to 0$, it reduces to the ML estimate, and taking $\beta_k \to \infty$, only the prior influence remains.

We must be careful about the influence upon the covariance estimate however. It is clear from (3) that the covariance can be increased arbitrarily with the parameter β_k and thus it is inadvisable to make β_k large. This is particularly unfortunate for us, since we want the facility to set means as given by TURN-RES, which is clearly not feasible by use of β_k.

3. **Inverse Wishart distribution over $\boldsymbol{\Sigma_k}$**: In an exponential family conjugate prior, we usually have two parameters to choose, the counts of pseudo-observations and the sufficient statistics of pseudo-observations. It is easy to see that these correspond to the parameters ν_k and Ψ_k. But since from the conjugate derivation, $\Psi_k = \tilde{n}\tilde{S}$, where $\tilde{n}$ and $\tilde{S}$ are the pseudo counts and pseudo scatter matrix respectively, we likewise must multiply the empirical scatter matrix $\tilde{\Sigma}_k = \sum_{n \in \text{cl}(k)} (x_n - \mu_k^{t-1})(x_n - \mu_k^{t-1})^\top / N_k$ by the count. Hence the density given above has $\nu_k \Psi_k$ instead of simply Ψ_k inside the Trace.

   Since both β_k and ν_k are both measure of pseudo counts, or the strength of our beliefs regarding the priors of μ_k and Σ_k, it makes sense to couple them - set $\nu_k = \beta_k$. This also solves the problem associated with large values of β_k. Since this value is now on the denominator of the $\hat{\Sigma}_k$ estimate, we can take β_k arbitrarily large. This does restrict the degrees of freedom in the model, and because we require $\nu_k \geq d + 1$, where d is the dimension of the model, we are restricted in our choice of β_k. Finally, the effective pseudo-sufficient statistic for $p(\Sigma_k)$ is now $(\mu_k - m_k)(\mu_k - m_k)^\top + \Psi_k$. We believe this is a tolerable restriction since the additional term simply reflects the uncertainty in μ_k in the covariance estimate.

We now have only 2 types of free paramaters for the user to choose, a (`alpha` in the `fitGMM` documentation), and β_k. Both reflect the strength of belief in our prior, that is the clusters found during the TURN-RES phase. And we can even choose $\beta_k \propto \alpha_k$, the empirical proportions from TURN-RES with an appropriate multiplier b, (`beta` in the `fitGMM` documentation). For ease, the user is only introduced to these *strength* parameters, as a prior belief in the mixing proportions and a prior belief in the cluster moments respectively, but we leave access to the underlying parameters for the more adventurous users.

References

[1] A. Foss, O.R. Zaiane, *A parameterless method for efficiently discovering clusters of arbitrary shape in large datasets*. IEEE, 2002

[2] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006

[3] R.M. Neal, G.E. Hinton, *A view of the EM algorithm that justifies incremental, sparse, and other variants*. Learning in Graphical Models, MIT Press, 1998