EM for Mixtures of Robust PCA Model

Evgenii Egorov, egorov.evgenyy@ya.ru

October 2, 2020

The goal of this technical note is to show the machinery of EM-algorithm. The model of Mixtures of Robust PCA is good choice for that, because it has both discrete and continuous latent variables. I tried to formulate some hints, that make the derivation as simple as possible:

- Write the objective for M-step at first, without taking the expectations. From it you can see what expectations you will need to take and then decide what posterior's you need to find. Keep in mind the power of nested expectations.
- We could take the expectation of the gradient, not the objective.

Also it's useful to reduce the notation as much as possible at the each step.

Let's write joint model:

$$p(\{y_n, x_n, u_n, z_n\}_{n=1}^N) = \prod_{n,k=1}^{N,K} \left[\pi_k \mathcal{N}(y_n | W_k x_n + \mu_k, (u\tau_k)^{-1} I_D) \mathcal{N}(x | 0, u^{-1} I_d) \mathrm{G}(u | \nu_k, \nu_k)\right]^{z_{nk}}.$$

Our latent variables are $\{x_n, u_n, z_n\}_{n=1}^N$ and our parameters of optimization at the M-step $\theta = \{W_k, \mu_k, \nu_k, \pi_k\}$. Let's write log of joint model, keeping related to the θ terms:

$$\log p(\{y_n, x_n, u_n, z_n\}_{n=1}^N) \propto^+ \sum_{n,k=1}^{N,K} z_{nk} \left[\log \pi_k - \tfrac{u_n \tau_k}{2} \|y_n - \mu_k - W_k x_n\|_2^2 - \tfrac{u_n}{2} \|x_n\|_2^2 - \right.$$
$$\left. - \log Z(\nu_k) + (\nu_k - 1) \log u_n - \nu_k u_n + \tfrac{d}{2} \log u_n + \tfrac{D}{2} [\log u_n + \log \tau_k]\right].$$

We need to take the expectation of the following: z_{nk} and for $u_n x_n x_n^T, u_n, \log u_n$ given the mixture component $z_{nk} = 1$. Hence, we consider the following factorization of the posterior:

$$p(x_n, u_n | y_n, z_{nk} = 1) = p(x_n | u_n, y_n) p(u_n | y_n) = \dots \; ?,$$
$$p(z_{nk} = 1 | y_n) \propto p(y_n | z_{nk}) p(z_{nk}) = \dots \; ?$$

Let's find the distributions, that we mention above. Starting with simple steps:

1. $p(y|u) = \int dx \; \mathcal{N}(y | Wx + \mu, (u\tau)^{-1} I_D) \mathcal{N}(x | 0, u^{-1} I_d)$,
2. $p(y) = \int du \; \mathrm{G}(u | \nu, \nu) p(y | u)$.

We could note, that first is a Gaussian distribution. Hence, instead of the integration we could just find the moments using nested expectation:

$$E_y y = E_x E_{y|x} y = E_x W x + \mu = \mu,$$
$$E_y y y^T = E_x E_{y|x} (Wx + \mu)(Wx + \mu)^T = (u\tau)^{-1} I + u^{-1} W W^T + \mu \mu^T,$$
$$\mathrm{Cov}(y) = u^{-1}(\tau^{-1} I + W W^T) \Longrightarrow p(y|u) = \mathcal{N}(y | \mu, u^{-1}[\tau^{-1} I + W W^T]).$$

And for the second step:

$$p(y) = \int du \; \mathrm{G}(u | \nu, \nu) p(y|u) = \int du \; \mathrm{G}(u | \nu, \nu) \mathcal{N}(y | \mu, u^{-1}[\tau^{-1} I + W W^T]) =$$
$$= \int du \; [Z(\nu, \nu)]^{-1} \left|2\pi(\tau^{-1} I + W W^T)\right|^{-1/2} \exp\left[-u \left(\nu + \tfrac{1}{2}(y - \mu)^T (\tau^{-1} I + W W^T)^{-1} (y - \mu)\right)\right] u^{\nu + D/2} =$$
$$= [Z(\nu, \nu)]^{-1} \, Z\left[\nu + \tfrac{D}{2}, \nu + \tfrac{1}{2}(y - \mu)^T (\tau^{-1} I + W W^T)^{-1} (y - \mu)\right] \left|2\pi(\tau^{-1} I + W W^T)\right|^{-1/2} =$$
$$= \frac{\Gamma(\nu + \frac{D}{2})}{\Gamma(\nu)} \left|2\pi\nu(\tau^{-1} I + W W^T)\right|^{-1/2} \left[1 + \tfrac{1}{2}(y - \mu)^T (\tau^{-1} I + W W^T)^{-1} (y - \mu)\right]^{-(\nu + D/2)},$$

where we used $Z(a,b)$ to denote the normalization constant of the gamma distribution. Also we could denote our result as student distribution:

$$\mathrm{St}(y;\nu,\mu,\tau,W)=\frac{\Gamma(\nu+\frac{D}{2})}{\Gamma(\nu)}\left|2\pi\nu(\tau^{-1}I+WW^T)\right|^{-1/2}\left[1+\tfrac{1}{2}(y-\mu)^T(\tau^{-1}I+WW^T)^{-1}(y-\mu)\right]^{-(\nu+D/2)}.$$

Finally, we could get posterior for discrete latent variables and then continue only with particular components:

$$\boxed{p(z_{nk}=1|y_n)=\frac{p(z_{nk}=1)p(y_n|z_{nk}=1)}{\sum_l p(z_{nl}=1)p(y_n|z_{nl}=1)}=\frac{\pi_k\ \mathrm{St}(y;\nu_k,\mu_k,\tau_k,W_k)}{\sum_l \pi_l\ \mathrm{St}(y_n;\nu_l,\mu_l,\tau_l,W_l)}}.$$

For shortness, we denote this with $r_{nk}=p(z_{nk}=1|y_n)$. Next we can find $p(u_n|y_n,z_{nk}=1)$:

$$p(u_n|y_n,z_{nk}=1)\propto p(y_n|u_n,z_{nk}=1)p(u_n|z_{nk}=1).$$

Next all densities are conditioned with the $z_{nk}=1$ to the particular component, so I omit this to clear notation. Also I add the index n only when I use summation over objects. Let's derive the $p(u|y)$:

$$\begin{aligned}\log p(u|y)&\propto^{+}\log p(y|u)+\log p(u)=\log\mathcal{N}(y|\mu,u^{-1}[WW^T+\tau^{-1}I])+\log\mathrm{G}(u;\nu,\nu)=\\&=u^{D/2+\nu-1}\exp\left\{-u(\nu+\tfrac{1}{2}(y-\mu)^T(WW^T+\tau^{-1}I)^{-1}(y-\mu)\right\}\\&\Longrightarrow\boxed{p(u|y)=\mathrm{G}(u;D/2+\nu,\nu+\tfrac{1}{2}(y-\mu)^T(WW^T+\tau^{-1}I)^{-1}(y-\mu)}.\end{aligned}$$

As next step, we need to find $p(x|y,u)$:

$\log p(x|y,u)\propto^{+}\log p(y|x,u)+\log p(x|u)=-\frac{u\tau}{2}\|y-\mu-Wx\|_2^2-\frac{u}{2}\|x\|_2^2\Longrightarrow$ Gaussian.

Let's find moments as mean is MAP and covariance is inverse minus hessian:

$\nabla_x: u\tau W^T(y-\mu-Wx)-ux=0\Longrightarrow E_{x|y,u}x=\tau(I+\tau W^TW)^{-1}W^T(y-\mu)$.

$\nabla_x^2:-u\tau W^TW-u\Longrightarrow\mathrm{Cov}(x|y,u)=u^{-1}[\tau W^TW+I]^{-1}$.

$$\boxed{p(x|y,u)=\mathcal{N}(x;\tau(I+\tau W^TW)^{-1}W^T(y-\mu),u^{-1}[I+\tau W^TW]^{-1})}.$$

Let's make M-step for the objective:

$$\langle\mathcal{L}\rangle=\langle\log p(\{y_n,x_n,u_n,z_n\}_{n=1}^N)\rangle\propto^{+}\sum_{n,k=1}^{N,K}\langle z_{nk}\left[\log\pi_k-\tfrac{u_n\tau_k}{2}\|y_n-\mu_k-W_kx_n\|_2^2-\tfrac{u_n}{2}\|x_n\|_2^2-\right.$$

$$\left.-\log Z(\nu_k)+(\nu_k-1)\log u_n-\nu_ku_n+\tfrac{d}{2}\log u_n+\tfrac{D}{2}[\log u_n+\log\tau_k]\right\rangle_{\text{posterior}},$$

$$\text{posterior}:=p(x_n,u_n,z_{nk}=1|y_n)=p(x_n|u_n,y_n,z_{nk}=1)p(u_n|y_n,z_{nk}=1)p(z_{nk}=1|y_n).$$

M step: π_k

$\mathcal{L}_{nk}(\pi)=z\log\pi,\langle\nabla\mathcal{L}_{nk}\rangle=\langle z\rangle\frac{1}{\pi},\langle\nabla_{\pi_k}\mathcal{L}\rangle=\frac{1}{\pi_k}\sum_{n=1}^N r_{nk}$. Adding constraint $\sum_k\pi_k=1$:

$$\langle\nabla_{\pi_k}\mathcal{L}\rangle+\nabla_{\pi_k}\lambda(1-\sum_k\pi_k)=0\Longrightarrow\boxed{\pi_k=\frac{1}{N}\sum_{n=1}^N r_{nk}}.$$

M step: μ_k

$$\mathcal{L}_{nk}(\mu)=-\tfrac{u\tau}{2}\|y-\mu-W\|_2^2,\nabla\mathcal{L}_{nk}=\tau u(y-\mu-Wx).$$

$$\langle\nabla_{\mu_k}\mathcal{L}\rangle=\sum_{n=1}^N r_{nk}\langle u\rangle_{nk}\langle u\rangle_{nk}(y_n-\mu_k-W_k\langle x\rangle_{nk})=0\Longrightarrow$$

$$\boxed{\mu_k=\left[\sum_{n=1}^N r_{nk}\langle u\rangle_{nk}\right]^{-1}\sum_{n=1}^N r_{nk}\langle u\rangle_{nk}(y_n-W_k\langle x\rangle_{nk})}.$$

M step: W_k

$\mathcal{L}_{nk} = -\frac{u\tau}{2}\|y-\mu-Wx\|_2^2 = -\frac{u\tau}{2}\left(\text{tr}[W^TWxx^T] - 2\text{tr}[Wx(y-\mu)^T]\right)$,
To take derivative we could sketch: $\text{tr}[(W+H)^T(W+H)] \approx \text{tr}[W^TW + 2W^TH]$.

$$\nabla_W \mathcal{L}_{nk} : -u\tau xx^TW + u\tau x(y-\mu)^T, \langle \nabla_{W_k}\mathcal{L}\rangle = \tau\left[\sum_{n=1}^N r_{nk}\langle u\rangle_{nk}\langle x\rangle_{nk}(y_n-\mu_k)^T - \langle uxx^T\rangle_{nk}\right]W_k = 0 \Longrightarrow$$

$$\boxed{W_k = \left[\sum_{n=1}^N r_{nk}\langle uxx^T\rangle_{nk}\right]^{-1}\left(\sum_{n=1}^N r_{nk}\langle u\rangle_{nk}\langle x\rangle_{nk}(y_n-\mu_k)^T\right)}.$$

M step: ν_k

$$\left(\sum_{n=1}^N r_{nk}\right)\frac{d}{d\nu_k}\log Z(\nu_k,\nu_k) = \sum_{n=1}^N r_{nk}[\langle \log u\rangle_{nk} - \langle u\rangle_{nk}].$$

To find θ, we need to make several fixed-point iteration of such updates. To finalize, we need to write explicitly expectations from M-step updates, using posterior distributions that we found before. Note, that all parameters here is fixed, i.e. "old".

E-step

$$\boxed{\langle x\rangle_{nk} = \tau_k(I+\tau_k W_k^TW_k)^{-1}W_k^T(y_n-\mu_k)},$$

$$\boxed{\langle uxx^T\rangle_{nk} = [I+\tau_k W^TW]^{-1} + \langle u\rangle_{nk}\langle x\rangle_{nk}\langle x\rangle_{nk}^T},$$

$$\boxed{\langle u\rangle_{nk} = [\tfrac{D}{2}+\nu_k]\left[\nu_k + \tfrac{1}{2}(y_n-\mu_k)^T(W_kW_k^T+\tau_k^{-1}I)^{-1}(y_n-\mu_k)\right]^{-1}},$$

$$\boxed{\langle \log u\rangle_{nk} = \psi(\tfrac{D}{2}+\nu_k) - \log\left[\nu_k + \tfrac{1}{2}(y_n-\mu_k)^T(W_kW_k^T+\tau^{-1}I)^{-1}(y_n-\mu_k)\right]},$$

where ψ is a digamma function,

$$\boxed{\langle z_{nk}\rangle = \frac{\pi_k \text{St}(y;\nu_k,\mu_k,\tau_k,W_k)}{\sum_l \pi_l\ \text{St}(y_n;\nu_l,\mu_l,\tau_l,W_l)}}.$$

At the homework all $\tau_k = 1$, so I don't derive M step for it here, but it could be done in the same several lines. The result could be found here. Note, that we use parametrization of the normal distribution with covariance and in the link with inverse of the covariance.