Variational Bayesian Model Selection for Mixture Distributions

Yassin El Hajj Chehade, Amal Chaoui
Centrale Lille, Université de Lille - Lille, France

e-mail: yassin.elhajjchehade.etu@univ-lille.fr; amal.chaoui@centrale.centralelille.fr

Abstract

Mixture modeling is a powerful technique for integrating multiple data generating processes into a single model. The main problem in the application of such models is inferring the right number of components that generated the data. Cross-validation, statistical tests or Bayesian-based approaches returning a posterior distribution on the number of components can be used to address the problem. However, these methods are usually computationally expensive. Corduneanu and Bishop (2001) proposed a variational treatment to the problem of approximating the posterior distribution in which the mixing coefficients are optimized using marginal likelihood. The method is able to recover the suitable number of components by imposing on the mixing coefficients of unwanted components to vanish.

1 Introduction

In a Gaussian mixture model, it is assumed that a given variable x can be drawn from one of M Gaussian data generating processes, each with their own set of parameters (μ_i, T_i). The likelihood of the M-mixture model then writes

$$P(D = (x_1, ..., x_N)|\pi, \mu, T) = \prod_{n=1}^{N} \left[\sum_{i=1}^{M} \pi_i \mathcal{N}(x_n|\mu_i, T_i) \right] \tag{1}$$

where $(\pi_i)_1^M$ are the mixing coefficients, verifying $\sum_{i=1}^{M} \pi_i = 1$.

A traditional Bayesian-based approach introduces prior distributions over the mixing coefficients, the number of components as well as their parameters. The optimal value of the components M is then set to maximize the posterior distribution of the mixture. It is therefore necessary to consider different values up to an upper bound, which makes the computations intractable and less efficient.

The authors proposed a method that sets an initial fixed number of components (i.e. the upper bound considered above) and the mixing coefficients associated to unneeded components converge to zero as their values are optimized using the marginal likelihood (i.e. *type 2 maximum-likelihood*). This is evaluated by means of variational Bayesian methods that maximize a tractable lower bound.

2 Bayesian Mixture Model

In the variational setting proposed in the paper, Corduneanu and Bishop (2001) introduce binary latent variables $(s_{in})_{1\leq i\leq M, 1\leq n\leq N}$, where $s_{in} = 1$ if the data point x_n is generated from component i. Conjugate priors are also defined over the means $(\mu_i)_{1\leq i\leq M}$ and covariances $(T_i)_{1\leq i\leq M}$. The joint distribution over the data D can then be expressed

$$P(D, \mu, T, s|\pi) = P(D|\mu, T, s)P(\mu)P(T)P(s|\pi) \tag{2}$$

$$P(D|\mu, T, s) = \prod_{n=1}^{N}\prod_{i=1}^{M}\mathcal{N}(x_n|\mu_i, T_i)^{s_{in}} \tag{3}$$

$$P(\mu) = \prod_{i=1}^{M}\mathcal{N}(\mu_i|0, \beta I) \tag{4}$$

$$P(T) = \prod_{i=1}^{M}\mathcal{W}(T_i|\nu_i, V) \tag{5}$$

$$P(s|\pi) = \prod_{i=1}^{M}\prod_{n=1}^{N}\pi_i^{s_{in}} \tag{6}$$

where β is a fixed parameter of small value and ν and V are respectively the degrees of freedom and the scale matrix's inverse of the Wishart distribution. The graphical representation of the probabilistic model is shown in figure 1.

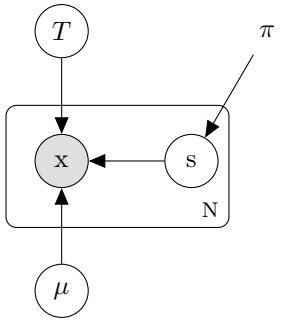

Figure 1: Graphical representation of the Gaussian mixture model

The method described in the paper consists in evaluating the marginal likelihood $P(D|\pi)$ by marginalizing $P(D, \mu, T, s|\pi)$ over $\theta = (\mu, T, s)$. However, this involves computing an integral which is generally intractable. Variational Bayesian methods consists in introducing a distribution $Q(\theta)$ that approximates the posterior distribution. It can be specified in a factorized form $Q(\theta) = Q_\mu(\mu)Q_T(T)Q_s(s)$. The true marginal log-likelihood $\log P(D|\pi)$ is approximated by considering the maximum of a lower bound $\mathcal{L}(Q)$ as defined in inequality 7.

$$\begin{aligned}\log P(D|\pi) &= \log \int Q(\theta)\frac{P(D, \theta|\pi)}{Q(\theta)}d\theta \\ &\geq \int Q(\theta)\log\frac{P(D, \theta|\pi)}{Q(\theta)}d\theta = \mathcal{L}(Q)\end{aligned} \tag{7}$$

In fact, the difference between $\log P(D|\pi)$ and $\mathcal{L}(Q)$ gives rise to the Kullback-Leibler (KL) divergence between the approximation $Q(\theta)$ and the posterior $P(\theta|D, \pi)$ (see equation 8).

$$\log P(D|\pi) = \mathcal{L}(Q) + KL(Q||P) = \mathcal{L}(Q) - \int Q(\theta)\log\frac{P(D, \theta|\pi)}{Q(\theta)} \tag{8}$$

Minimizing the KL-divergence $KL(Q||P)$ (hence maximizing $\mathcal{L}(Q)$) leads to expressing $Q_i(\theta_i)$ as follows

$$\log Q_i(\theta_i) = \mathbb{E}_{j\neq i}\left[\log P(D, \theta)\right] + cst \tag{9}$$

To derive the expression of $Q_s(s)$ by applying equation 9, we use the approximation

$$P(\theta|\mathcal{D}) \approx Q(\theta) = Q(\mu)Q(T)Q(s).$$

The log of the joint posterior distribution of $\theta = (\mu, T, s)$ is given by :

$$\log P(\theta|\mathcal{D}) = \sum_{i=1}^{M} -\frac{1}{2}\mu_i^T(\beta I)\mu_i - \log|T_i|^{(\nu-D-1)/2} - \frac{1}{2}Tr(VT_i)$$
$$+\sum_{i=1}^{M}\sum_{n=1}^{N} s_{in}\left[\log\pi_i - \frac{D}{2}\log 2\pi + \frac{1}{2}\log|T_i| - \frac{1}{2}(x_n-\mu_i)^T T_i(x_n-\mu_i)\right] + cst \tag{10}$$

By keeping only the terms involving the parameters s_{in} and taking the expectation over the other variables, we yield

$$Q_s(s) \propto \prod_{n=1}^{N}\prod_{i=1}^{M} Q_{s_{in}}(s_{in}) \implies Q_s(s) = \prod_{n=1}^{N}\prod_{i=1}^{M} p_{in}^{s_{in}} \tag{11}$$

where p_{in} is defined

$$p_{in} = \frac{\tilde{p}_{in}}{\sum_{j=1}^{M}\tilde{p}_{in}} \tag{12}$$

$$\tilde{p}_{in} = \exp\{\log\pi_i - \frac{D}{2}\log 2\pi + \frac{1}{2}\langle\log|T_i|\rangle$$
$$-\frac{1}{2}Tr[\langle T_i\rangle(x_n x_n^T - \langle\mu_i\rangle x_n^T - x_n\langle\mu_i^T\rangle + \langle\mu_i\mu_i^T\rangle)]\} \tag{13}$$

Similarly, we obtain the expressions of the variational distributions (the details are provided in the appendix)

$$Q_\mu(\mu) = \prod_{i=1}^{M} Q_{\mu_i}(\mu_i) = \prod_{i=1}^{M}\mathcal{N}(\mu_i|m_\mu^{(i)}, T_\mu^{(i)}) \tag{14}$$
$$Q_T(T) = \prod_{i=1}^{M} Q_{T_i}(T_i) = \prod_{i=1}^{M}\mathcal{W}(T_i|\nu_T^{(i)}, V_T^{(i)}) \tag{15}$$

Using the fact $s_{in} \sim \mathcal{B}(p_{in})$, $\mu_i \sim \mathcal{N}(m_\mu^{(i)}, T_\mu^{(i)})$, and $T_\mu^{(i)} \sim \mathcal{W}(\nu_T^{(i)}, V_T^{(i)})$, the expectation of variables writes as follows

$$\langle s_{in}\rangle = p_{in} \tag{16}$$
$$\langle\mu_i\rangle = m_\mu^{(i)} \tag{17}$$
$$\langle\mu_i\mu_i^T\rangle = Cov(\mu_i) + \langle\mu_i\rangle\langle\mu_i^T\rangle = T_\mu^{(i)} + m_\mu^{(i)}m_\mu^{(i)^T} \tag{18}$$
$$\langle T_i\rangle = \nu_T^{(i)}V_T^{(i^{-1})} \tag{19}$$

Finally, we can express the lower bound on the marginal log-likelihood as follows (again, the details of computations are provided in the appendix)

$$\mathcal{L}(Q) = \langle\log P(\mathcal{D}|\mu,T,s)\rangle + \langle\log P(s|\pi)\rangle + \langle\log P(\mu)\rangle + \langle\log P(T)\rangle$$
$$-\langle\log Q(\mu)\rangle - \langle\log Q(T)\rangle - \langle\log Q(s)\rangle \tag{20}$$

3 Implementation and Results

The maximization of the lower bound $\mathcal{L}(Q)$ with respect to π enables to approximate the marginal log-likelihood. However, due to the coupled dependence of the parameters, an EM procedure is adopted to adjust the parameters through a sequential update in which $\mathcal{L}(Q)$ is maximized w.r.t. π (M-step)

and then the variational distributions $Q_\mu(\mu)$, $Q_s(s)$, and $Q_T(T)$ account for the current update of the parameters (E-step).

The algorithm is initialized with a mixture of 10 components with equal mixing coefficients. The coefficients of undesired components converge to 0 and are removed when their values go below 10^{-5}.

We verified the relevance of the results through 3 experiments (using the same data generated in the original paper):

- **600 data points from a mixture of 5 mixtures** (means $[0, 0], [3, -3], [3, 3], [-3, 3], [-3, -3]$ and covariances $[1, 0; 0, 1], [1, 0.5; 0 : 5, 1], [1, -0.5; -0.5, 1]$ (see figure 2);

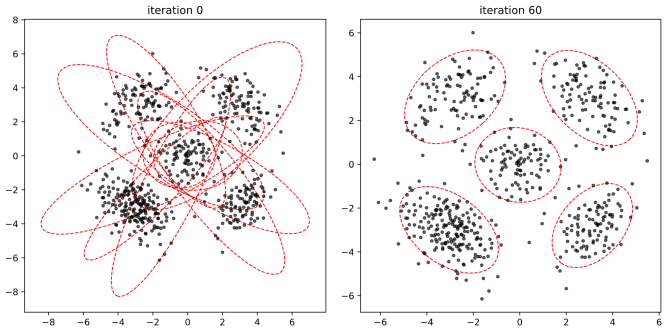

Figure 2: Configurations obtained using 600 data points from a mixture of 5 Gaussians

- **900 data points from three mixtures** (means $[0, -2], [0, 0], [0, 2]$ and same covariance $[2, 0; 0, 0.2]$)(see figure 3);

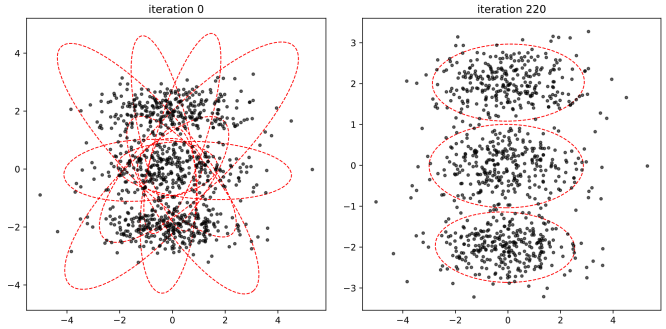

Figure 3: Configurations obtained using 900 data points from a mixture of 3 Gaussians with the same covariance

- **400 data points from three mixtures** (mean $[0, 0]$ and covariances $[1, 0; 0, 0.2], [0, 02; -0.08; -0.08, 1.5]$, $[0.5, 0.4; 0.05, 0.05]$)(see figure 3).

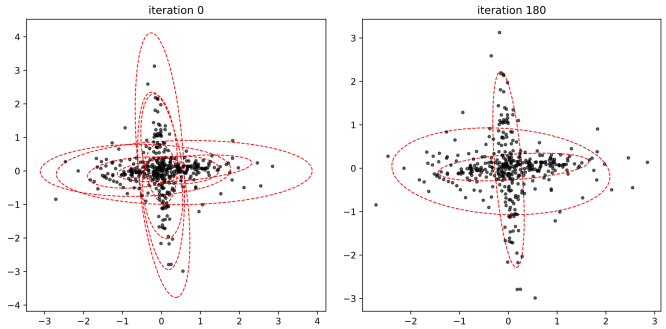

Figure 4: Configurations obtained using 400 data points from a mixture of 3 Gaussians with the same mean

Figure 5 illustrates the convergence of the likelihood lower bound during the first experiment.

The model optimization automatically recovers the number of generating mixtures in all proposed synthetic data sets. Besides, the variational mixture model is even capable to estimate the means and covariances of the mixture components.

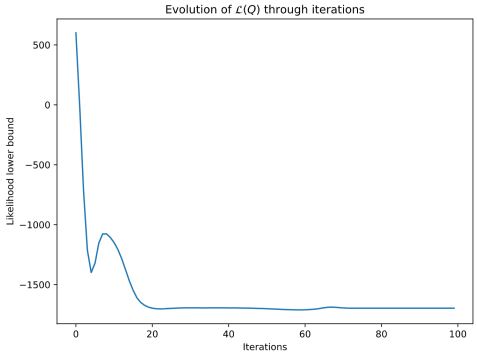

Figure 5: Convergence of the likelihood lower bound

4 Conclusion

In this report, we briefly reviewed the variational Bayesian method for mixture Gaussians proposed by Corduneanu and Bishop (2001). The approach is based on setting an initial fixed number of components and allowing to eliminate unwanted components by setting the associated mixing coefficients to 0. The advantage of the framework provided, apart from inferring the true number of components used to generate data, lies in its ability to estimate the means and covariances of the Gaussian mixtures.

References

Corduneanu, A., & Bishop, C. (2001, 01). Variational bayesian model selection for mixture distribution. *Artificial Intelligence and Statistics*, *18*, 27-34.

Appendix: Derivation of the parameters expressions

The standard Variational Inference procedure factorizes the posterior distributions for the latent variables and the component parameters.

$$P(\mu, T, s|\mathcal{D}) \approx Q(\theta) = Q(\mu)Q(T)Q(s)$$

We recall the Bayesian model

$$P(\mathcal{D}|\mu, T, s) = \prod_{n=1}^{N}\prod_{i=1}^{M}\mathcal{N}(x_n|\mu_i, T_i)^{s_{in}} \tag{21}$$

$$P(s|\pi) = \prod_{n=1}^{N}\prod_{i=1}^{M}\pi_i^{s_{in}} \tag{22}$$

$$P(\mu) = \prod_{i=1}^{M}\mathcal{N}(\mu_i|0, \beta I) \tag{23}$$

$$P(T) = \prod_{i=1}^{M}\mathcal{W}(T_i|\nu, V) \tag{24}$$

The log of the joint posterior distribution of $\theta = (\mu, T, s)$ is given by :

$$\log P(\theta|\mathcal{D}) = \sum_{i=1}^{M} \left[-\frac{1}{2}\mu_i^T(\beta I)\mu_i - \log|T_i|^{(\nu-D-1)/2} - \frac{1}{2}Tr(VT_i) \right]$$
$$+ \sum_{i=1}^{M}\sum_{n=1}^{N} s_{in} \left[\log \pi_i - \frac{D}{2}\log 2\pi + \frac{1}{2}\log|T_i| - \frac{1}{2}(x_n - \mu_i)^T T_i (x_n - \mu_i) \right] + cst \tag{25}$$

Expression of $Q_s(s)$***:***

Now using the above equation, the best variation distribution for each parameter is computed by keeping only the terms involving the parameter and taking the expectation over the other variables. In case of $Q(s)$, we obtain

$$\begin{aligned}
\log Q(s_{in}) &= \langle s_{in}[\log \pi_i - \frac{D}{2}\log 2\pi + \frac{1}{2}\log|T_i| - \frac{1}{2}(x_n - \mu_i)^T T_i (x_n - \mu_i)]\rangle \\
&= s_{in}[\log \pi_i - \frac{D}{2}\log 2\pi + \frac{1}{2}\langle \log|T_i|\rangle \\
&- \frac{1}{2}Tr\{\langle T_i\rangle \langle (x_n - \mu_i)(x_n - \mu_i)^T\rangle\}] + cst
\end{aligned} \tag{26}$$

Finally, we get the following expression for $Q(s)$:

$$Q(s) \propto \prod_{n=1}^{N}\prod_{i=1}^{M} Q(s_{in}) \tag{27}$$

$$Q(s) = \prod_{n=1}^{N}\prod_{i=1}^{M} p_{in}^{s_{in}} \tag{28}$$

where p_{in} is defined

$$p_{in} = \frac{\tilde{p}_{in}}{\sum_{j=1}^{M}\tilde{p}_{in}} \tag{29}$$

$$\tilde{p}_{in} = \exp\{\log \pi_i - \frac{D}{2}\log 2\pi + \frac{1}{2}\langle \log|T_i|\rangle$$
$$-\frac{1}{2}Tr[\langle T_i\rangle (x_n x_n^T - \langle \mu_i\rangle x_n^T - x_n\langle \mu_i^T\rangle + \langle \mu_i \mu_i^T\rangle)]\} \tag{30}$$

Expression of $Q_\mu(\mu)$***:***

In the same way, we derive the expression of $Q(\mu)$

$$Q(\mu) = \prod_{i=1}^{M} Q(\mu_i) \tag{31}$$

where

$$\begin{aligned}
\log Q(\mu_i) &= \langle -\frac{1}{2}\mu_i^T(\beta I)\mu_i - \frac{1}{2}\sum_{n=1}^{N} s_{in}(x_n - \mu_i)^T T_i (x_n - \mu_i)\rangle + cst \\
&= -\frac{1}{2}\mu_i^T(\beta I + \sum_{n=1}^{N}\langle s_{in}\rangle\langle T_i\rangle)\mu_i + (\sum_{n=1}^{N} x_n^T\langle s_{in}\rangle\langle T_i\rangle)\mu_i + cst
\end{aligned} \tag{32}$$

We recognize the expression of a Gaussian

$$Q(\mu_i) = \mathcal{N}(\mu_i|m_\mu^{(i)}, T_\mu^{(i)}) \tag{33}$$

where we have defined

$$T_\mu^{(i)} = \beta I + \langle T_i \rangle \sum_{n=1}^{N} \langle s_{in} \rangle \tag{34}$$

$$m_\mu^{(i)} = {T_\mu^{(i)}}^{-1} \langle T_i \rangle \sum_{n=1}^{N} x_n \langle s_{in} \rangle \tag{35}$$

Expression of $Q_T(T)$***:***
Similarly, we derive the expression of $Q(T)$

$$Q(T) = \prod_{i=1}^{M} Q(T_i) \tag{36}$$

where

$$\begin{aligned}
\log Q(T_i) &= \langle (\nu - D - 1)/2 \log |T_i| - \frac{1}{2} Tr(V T_i) - \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu_i)^T T_i (x_n - \mu_i) \\
&= [(\nu + \sum_{n=1}^{N} \langle s_{in} \rangle) - D - 1]/2 \log |T_i| \\
&- \frac{1}{2} Tr\{(V + \sum_{n=1}^{N} \langle s_{in} \rangle \langle (x_n - \mu_i)(x_n - \mu_i)^T \rangle) T_i\} \rangle
\end{aligned} \tag{37}$$

We recognize the expression of the Wishart distribution

$$Q(T_i) = \mathcal{W}(T_i | \nu_T^{(i)}, V_T^{(i)}) \tag{38}$$

where

$$\nu_T^{(i)} = \nu + \sum_{n=1}^{N} \langle s_{in} \rangle \tag{39}$$

$$V_T^{(i)} = V + \sum_{n=1}^{N} \left[x_n x_n^T \langle s_{in} \rangle - x_n \langle s_{in} \rangle \langle \mu_i^T \rangle - \langle \mu_i \rangle x_n^T \langle s_{in} \rangle + \langle \mu_i \mu_i^T \rangle \langle s_{in} \rangle \right] \tag{40}$$

Conclusion:

- $s_{in} \sim \mathcal{B}(p_{in})$;
- $\mu_i \sim \mathcal{N}(m_\mu^{(i)}, T_\mu^{(i)})$
- $T_\mu^{(i)} \sim \mathcal{W}(\nu_T^{(i)}, V_T^{(i)})$

Expectation of the variables:
The expectation of variables used in the previous expressions are then given as follows :

$$\langle s_{in} \rangle = p_{in} \tag{41}$$

$$\langle \mu_i \rangle = m_\mu^{(i)} \tag{42}$$

$$\langle \mu_i \mu_i^T \rangle = Cov(\mu_i) + \langle \mu_i \rangle \langle \mu_i^T \rangle = T_\mu^{(i)} + m_\mu^{(i)} {m_\mu^{(i)}}^T \tag{43}$$

$$\langle T_i \rangle = \nu_T^{(i)} V_T^{(i^{-1})} \tag{44}$$

The lower bound $\mathcal{L}(Q)$**:**
Expression of the lower bound writes

$$\begin{aligned}\mathcal{L}(Q) &= \int Q(\theta)\log\frac{P(\mathcal{D},\theta|\pi)}{Q(\theta)}d\theta \\ &= \langle\log P(\mathcal{D}|\mu,T,s)\rangle + \langle\log P(s|\pi)\rangle + \langle\log P(\mu)\rangle + \langle\log P(T)\rangle \\ &- \langle\log Q(\mu)\rangle - \langle\log Q(T)\rangle - \langle\log Q(s)\rangle \end{aligned} \tag{45}$$

Using (21), (22),(23), (24), we derive the following

$$\begin{aligned}\langle\log P(\mathcal{D}|\mu,T,s)\rangle = \sum_{i=1}^{M}\sum_{n=1}^{N}\langle s_{in}\rangle\{&\frac{1}{2}\langle\log|T_i|\rangle - \frac{D}{2}\log(2\pi) \\ &- \frac{1}{2}Tr[\langle T_i\rangle\langle(x_n-\mu_i)(x_n-\mu_i)^T\rangle]\}\end{aligned} \tag{46}$$

$$\langle\log P(s|\pi)\rangle = \sum_{i=1}^{M}\sum_{n=1}^{N}\langle s_{in}\rangle\log\pi_i \tag{47}$$

$$\langle\log P(\mu)\rangle = \frac{MD}{2}\log\frac{\beta}{2\pi} - \frac{\beta}{2}\sum_{i=1}^{M}\langle\mu_i^T\mu_i\rangle \tag{48}$$

We recall the density of the multivariate gamma distribution

$$\Gamma_p(a) = \pi^{p(p-1)/4}\prod_{j=1}^{p}\Gamma[a+(1-j)/2] \tag{49}$$

$$\begin{aligned}\langle\log P(T)\rangle = &-\frac{\nu MD}{2}\log 2 + \frac{\nu M}{2}\log|V| - \frac{MD(D-1)}{4}\log\pi \\ &- M\sum_{j=1}^{D}\log\Gamma[(\nu+1-j)/2] + (\nu-D-1)/2\sum_{i=1}^{M}\langle\log|T_i|\rangle \\ &- \frac{1}{2}Tr(V\sum_{i=1}^{M}\langle T_i\rangle)\end{aligned} \tag{50}$$

Using (28) & (41)

$$\langle\log Q(s)\rangle = \sum_{i=1}^{M}\sum_{n=1}^{N}\langle s_{in}\rangle\log p_{in} = \sum_{i=1}^{M}\sum_{n=1}^{N}\langle s_{in}\rangle\log\langle s_{in}\rangle \tag{51}$$

Using (31) & (33)

$$\begin{aligned}\langle\log Q(\mu)\rangle &= \sum_{i=1}^{M}[-\frac{D}{2}\log(2\pi) + \frac{1}{2}\log|T_\mu^{(i)}| - \frac{1}{2}Tr(T_\mu^{(i)}T_\mu^{(i)^{-1}})] \\ &= \sum_{i=1}^{M}[-\frac{D}{2}(1+\log(2\pi)) + \frac{1}{2}\log|T_\mu^{(i)}|]\end{aligned} \tag{52}$$

Using (36) & (38)

$$\begin{aligned}\langle\log Q(T)\rangle = \sum_{i=1}M\{&-\frac{\nu_T^{(i)}D}{2}\log 2 - \frac{D(D-1)}{4}\log\pi - \sum_{j=1}^{D}\log\Gamma(\frac{\nu_T^{(i)}+1-j}{2}) \\ &+ \frac{\nu_T^{(i)}}{2}\log|V_T^{(i)}| + \frac{\nu_T^{(i)}-D-1}{2}\langle\log|T_i|\rangle - \frac{1}{2}Tr(V_T^{(i)}\langle T_i\rangle)\}\end{aligned} \tag{53}$$