A Variational Bayesian Formulation for GTM: Theoretical Foundations

Iván Olier and Alfredo Vellido
Department of Computing Languages and Systems (LSI)
Technical University of Catalonia (UPC)
C/. Jordi Girona 1-3, Edifici Omega, Despatx S106
08034 - Barcelona, Spain
{iaolier,avellido}@lsi.upc.edu

Abstract

Generative Topographic Mapping (GTM) is a non-linear latent variable model of the manifold learning family that provides simultaneous visualization and clustering of high-dimensional data. It was originally formulated as a constrained mixture of Gaussian distributions, for which the adaptive parameters were determined by Maximum Likelihood (ML), using the Expectation-Maximization (EM) algorithm. In this paper, we define an alternative variational formulation of GTM that provides a full Bayesian treatment to a Gaussian Process (GP) - based variation of the model.

1 Introduction

Manifold learning models attempt to describe multivariate data in terms of low-dimensional representations, often with the goal of allowing the intuitive visualization of high-dimensional data. *Generative Topographic Mapping* (GTM) [1] is one such model that can be ascribed to the field of Statistical Machine Learning. Its probabilistic setting and functional similarity make it a principled alternative to *Self-Organizing Maps* (SOM) [2]. It can also be described as a density modelling method defined as constrained mixture of distributions. As such, it has been extended to perform missing data imputation [3, 4]; to handle data outliers robustly [4, 5]; to perform unsupervised feature selection [6, 7, 8]; and to analyse multivariate time series [8, 9], amongst other capabilities.

The GTM model has also been modified to provide active regularization in the presence of noise [10, 11]. In its basic formulation, the GTM is trained within the ML framework using EM, permitting the occurrence of data overfitting unless regularization is included, a major drawback when modelling noisy data. The regularization methods in [10, 11] were based on *Bayesian evidence* approaches. Alternatively, we could reformulate GTM within a fully Bayesian approach and endow the model with regularization capabilities based on evidence with Laplacian approximations [12], Markov Chain Monte Carlo (MCMC) methods [13], and variational techniques [14, 15]. In this paper, we chose the latter alternative to provide the theoretical foundations of a Variational GTM model based on the GTM with GP prior outlined in [10], to which a Bayesian estimation of its parameters is added. This Variational GTM should limit the negative effect of data overfitting, improving on the performance of the standard GTM with GP prior, while retaining the data visualization capabilities of the model. Variational techniques have been successfully associated to well-known methods such as Gaussian Mixture Models [16], Probabilistic PCA [17], Independent Component Analysis [18], and Hidden Markov Models [19], amongst others.

The remaining of this report is organized as follows: First, in section 2, introductions to the original GTM, the GTM with GP prior and a Bayesian approach of the GTM, are provided. This is followed, in section 3, by the description of the proposed Variational GTM.

2 Generative Topographic Mapping

2.1 The Original GTM

The neural network-inspired GTM is a nonlinear latent variable model of the manifold learning family, with sound foundations in probability theory. It performs simultaneous clustering and visualization of the observed data through a nonlinear and topology-preserving mapping from a visualization latent space in $\Re^L$(with *L* being usually 1 or 2 for visualization purposes) onto a manifold embedded in the $\Re^D$ space, where the observed data reside. The mapping that generates the manifold is carried out through a *regression function* given by:

$$\mathbf{y} = \mathbf{W}\mathbf{\Phi}\left(\mathbf{u}\right) \tag{1}$$

where $\mathbf{y} \in \Re^D$, $\mathbf{u} \in \Re^L$, $\mathbf{W}$ is the matrix that generates the mapping, and $\mathbf{\Phi}$ is a matrix with the images of *S* basis functions ϕ_s (defined as radially symmetric Gaussians in the original formulation of the model). To achieve computational tractability, the prior distribution of $\mathbf{u}$ in latent space is constrained to form a uniform discrete grid of *K* centres, analogous to the layout of the SOM units, in the form:

$$p\left(\mathbf{u}\right) = \frac{1}{K}\sum_{k=1}^{K}\delta\left(\mathbf{u} - \mathbf{u}_k\right) \tag{2}$$

This way defined, the GTM can also be understood as a constrained mixture of Gaussians. A density model in data space is therefore generated for each component *k* of the mixture, which, assuming that the observed data set $\mathbf{X}$ is constituted by $\mathbf{N}$ independent, identically distributed (i.i.d.) data points $\mathbf{x}_n$, leads to the definition of a complete likelihood in the form:

$$P\left(\mathbf{X}|\mathbf{W},\beta\right) = \left(\frac{\beta}{2\pi}\right)^{ND/2}\prod_{n=1}^{N}\left\{\frac{1}{K}\sum_{k=1}^{K}\exp\left(-\frac{\beta}{2}\left\|\mathbf{x}_n - \mathbf{y}_k\right\|^2\right)\right\} \tag{3}$$

where $\mathbf{y}_k = \mathbf{W}\mathbf{\Phi}\left(\mathbf{u}_k\right)$. From Eq. 3, the adaptive parameters of the model, which are $\mathbf{W}$ and the common inverse variance of the Gaussian components, β, can be optimized by ML using the EM algorithm. Details can be found in [1].

2.2 Gaussian Process Formulation of GTM

The original formulation of GTM described in the previous section has a hard constraint imposed on the mapping from the latent space to the data space due to the finite number of basis functions used. An alternative approach is introduced in [10], where the regression function using basis functions is replaced by a smooth mapping carried out by a GP prior. This way, the likelihood takes the form:

$$P\left(\mathbf{X}|\mathbf{Z},\mathbf{Y},\beta\right) = \left(\frac{\beta}{2\pi}\right)^{ND/2}\prod_{n=1}^{N}\prod_{k=1}^{K}\left\{\exp\left(-\frac{\beta}{2}\left\|\mathbf{x}_n - \mathbf{y}_k\right\|^2\right)\right\}^{z_{kn}} \tag{4}$$

where: $\mathbf{Z} = \{z_{kn}\}$ are binary membership variables complying with the restriction $\sum_{k=1}^{K} z_{kn} = 1$ and $\mathbf{y}_k = \left(y_{k1},\ldots,y_{kD}\right)^T$ are the column vectors of a matrix $\mathbf{Y}$ and the centroids of spherical Gaussian generators. Note that the spirit of $\mathbf{y}_k$ in this approach is similar to the regression version of GTM (Eq. 1) but with a different formulation: A GP formulation is assumed introducing a prior multivariate Gaussian distribution over $\mathbf{Y}$ defined as:

$$P\left(\mathbf{Y}\right) = \left(2\pi\right)^{-KD/2}\left|\mathbf{C}\right|^{-D/2}\prod_{d=1}^{D}\exp\left(-\frac{1}{2}\mathbf{y}_{(d)}^T\mathbf{C}^{-1}\mathbf{y}_{(d)}\right) \tag{5}$$

where $\mathbf{y}_{(d)}$ is each one of the row vectors of the matrix $\mathbf{Y}$ and $\mathbf{C}$ is a matrix where each element is a covariance function that can be defined as

$$\mathbf{C}(i,j) = \mathbf{C}(\mathbf{u}_i, \mathbf{u}_j) = \nu \exp\left(-\frac{\|\mathbf{u}_i - \mathbf{u}_j\|^2}{2\alpha^2}\right), \quad i,j = 1 \dots K \tag{6}$$

and where parameter ν is usually set to 1. The α parameter controls the flexibility of the mapping from the latent space to the data space. An extended review of covariance functions can be found in [20]. An alternative GP formulation was introduced in [21], but this approach had the disadvantage of not preserving the topographic ordering in latent space, being therefore inappropiate for data visualization purposes.

Note that Eqs. 3 and 4 are equivalent if a prior multinomial distribution over $\mathbf{Z}$ in the form $P(\mathbf{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left(\frac{1}{K}\right)^{z_{kn}} = \frac{1}{K^N}$ is assumed.

Eq. 4 leads to the definition of a log-likelihood and parameters $\mathbf{Y}$ and β of this model can be optimized using the EM algorithm, in a similar way to the parameters $\mathbf{W}$ and β in the regression formulation. Some basic details are provided in [10].

2.3 Bayesian GTM

The specification of a full Bayesian model of GTM can be completed by defining priors over the parameters $\mathbf{Z}$ and β. Since z_{kn} are defined as binary values, a multinomial distribution can be chosen for $\mathbf{Z}$:

$$P(\mathbf{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} p_{kn}^{z_{kn}} \tag{7}$$

where p_{kn} is the parameter of the distribution.

As in [17], a Gamma distribution[1] is chosen to be the prior over β:

$$P(\beta) = \Gamma(\beta | d_\beta, s_\beta) \tag{8}$$

where d_β and s_β are the parameters of the distribution. Therefore, the joint probability $P(\mathbf{X}, \mathbf{Z}, \mathbf{Y}, \beta)$ is given by:

$$P(\mathbf{X}, \mathbf{Z}, \mathbf{Y}, \beta) = P(\mathbf{X}|\mathbf{Z}, \mathbf{Y}, \beta) P(\mathbf{Z}) P(\mathbf{Y}) P(\beta) \tag{9}$$

This expression can be maximized through evidence methods using the Laplace approximation [12] or, alternatively, using Markov Chain Monte Carlo [13] or variational [14, 15] methods.

3 Variational GTM

3.1 Motivation of the Use of Variational Inference

A basic problem in Statistical Machine Learning is the computation of the marginal likelihood $P(\mathbf{X}) = \int P(\mathbf{X}, \Theta)\, d\Theta$, where $\Theta = \{\theta_i\}$ is the set of parameters defining the model. Depending of the complexity of the model, the analytical computation of this integral could be intractable. Variational inference allows approximating the marginal likelihood through Jensen's inequality as follows:

$$\begin{aligned} \ln P(\mathbf{X}) &= \ln \int P(\mathbf{X}, \Theta)\, d\Theta = \ln \int Q(\Theta) \frac{P(\mathbf{X}, \Theta)}{Q(\Theta)} d\Theta \\ &\geq \int Q(\Theta) \ln \frac{P(\mathbf{X}, \Theta)}{Q(\Theta)} d\Theta = F(Q) \end{aligned} \tag{10}$$

[1] The Gamma distribution is defined as follows: $\Gamma(\nu | d_\nu, s_\nu) = \frac{s_\nu^{d_\nu} \nu^{d_\nu - 1} \exp^{-s_\nu \nu}}{\Gamma(d_\nu)}$

The function $F(Q)$ is a lower bound function such that its convergence guarantees the convergence of the marginal likelihood. The goal in variational methods is choosing a suitable form for the density $Q(\Theta)$ in such a way that $F(Q)$ can be readily evaluated and yet which is sufficiently flexible that the bound is reasonably tight. A reasonable approximation for $Q(\Theta)$ is based on the assumption that it factorizes over each one of the parameters as $Q(\Theta) = \prod_i Q_i(\theta_i)$. That assumed, $F(Q)$ can be maximized leading the optimal distributions:

$$Q_i(\theta_i) = \frac{\exp \langle \ln P(\mathbf{X}, \Theta) \rangle_{k \neq i}}{\int \exp \langle \ln P(\mathbf{X}, \Theta) \rangle_{k \neq i} \, d\theta_i} \tag{11}$$

where $\langle \, . \, \rangle_{k \neq i}$ denotes an expectation with respect to the distributions $Q_k(\theta_k)$ for all $k \neq i$.

3.2 Variational Distributions

In order to apply the variational principles to the Bayesian GTM within the framework described in the previous section, a Q distribution of the form:

$$Q(\mathbf{Z}, \mathbf{Y}, \beta) = Q(\mathbf{Z}) \, Q(\mathbf{Y}) \, Q(\beta) \tag{12}$$

is assumed, where natural choices of $Q(\mathbf{Z})$, $Q(\mathbf{Y})$ and $Q(\beta)$ are similar distributions to the priors $P(\mathbf{Z})$, $P(\mathbf{Y})$ and $P(\beta)$, respectively. Thus,

$$Q(\mathbf{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \tilde{p}_{kn}^{z_{kn}} \tag{13}$$

$$Q(\mathbf{Y}) = \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{y}_{(d)} | \tilde{\mathbf{m}}^{(d)}, \tilde{\mathbf{\Sigma}}\right) \tag{14}$$

$$Q(\beta) = \Gamma\left(\beta | \tilde{d}_\beta, \tilde{s}_\beta\right) \tag{15}$$

Now, using Eqs. 13 to 15 in Eq. 11, the following expressions for the variational parameters $\tilde{\mathbf{\Sigma}}, \tilde{\mathbf{m}}^{(d)}, \tilde{p}_{kn}, \tilde{d}_\beta$ and $\tilde{s}_\beta$ can be obtained:

$$\tilde{\mathbf{\Sigma}} = \left(\langle \beta \rangle \sum_{n=1}^{N} \mathbf{G}_n + \mathbf{C}^{-1} \right)^{-1} \tag{16}$$

$$\tilde{\mathbf{m}}_{(d)} = \langle \beta \rangle \, \tilde{\mathbf{\Sigma}} \sum_{n=1}^{N} x_{nd} \langle \mathbf{z}_n \rangle \tag{17}$$

$$\tilde{p}_{kn} = p_{kn} \exp\left(-\frac{\langle \beta \rangle}{2} \left\langle \|\mathbf{x}_n - \mathbf{y}_k\|^2 \right\rangle_{\mathbf{Y}} \right) \tag{18}$$

$$\tilde{d}_\beta = d_\beta + \frac{ND}{2} \tag{19}$$

$$\tilde{s}_\beta = s_\beta + \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \langle z_{kn} \rangle \left\langle \|\mathbf{x}_n - \mathbf{y}_k\|^2 \right\rangle \tag{20}$$

where $\mathbf{z}_n$ corresponds to each row vector of $\mathbf{Z}$ and $\mathbf{G}_n$ is a diagonal matrix of size $K \times K$ with elements $\langle \mathbf{z}_n \rangle$. The moments in the previous equations are defined as:

$$\langle z_{kn} \rangle = \tilde{p}_{kn} \tag{21}$$

$$\langle \beta \rangle = \frac{\tilde{d}_\beta}{\tilde{s}_\beta} \tag{22}$$

$$\begin{aligned}
\langle \|\mathbf{x}_n - \mathbf{y}_k\| \rangle_{\mathbf{Y}} &= \left\langle (\mathbf{x}_n - \mathbf{y}_k)^T (\mathbf{x}_n - \mathbf{y}_k) \right\rangle_{\mathbf{Y}} \\
&= \sum_{d=1}^{D} x_{nd}^2 - 2x_{nd} \langle y_{kd} \rangle_{\mathbf{Y}} + \langle y_{kd}^2 \rangle_{\mathbf{Y}} \\
&= \sum_{d=1}^{D} x_{nd}^2 - 2x_{nd} \tilde{m}_k^{(d)} + \tilde{\Sigma}_{kk} + \left(\tilde{m}_k^{(d)} \right)^2 \\
&= D\tilde{\Sigma}_{kk} + \|\mathbf{x}_n - \tilde{\mathbf{m}}_k\|^2
\end{aligned} \tag{23}$$

Details on these calculations can be found in the appendix.

3.3 Lower Bound

Finally, and according to Eqs. 9 and 10, the lower bound function $F(Q)$ is derived from:

$$F(Q) = \int Q(\mathbf{Z}) Q(\mathbf{Y}) Q(\beta) \ln \frac{P(\mathbf{X}|\mathbf{Z}, \mathbf{Y}, \beta) P(\mathbf{Z}) P(\mathbf{Y}) P(\beta)}{Q(\mathbf{Z}) Q(\mathbf{Y}) Q(\beta)} d\mathbf{Z} d\mathbf{Y} d\beta \tag{24}$$

Integrating out, we obtain:

$$\begin{aligned}
F(Q) &= \langle \ln P(\mathbf{X}|\mathbf{Z}, \mathbf{Y}, \beta) \rangle + \langle \ln P(\mathbf{Z}) \rangle + \langle \ln P(\mathbf{Y}) \rangle + \langle \ln P(\beta) \rangle \\
&\quad - \langle \ln Q(\mathbf{Z}) \rangle - \langle \ln Q(\mathbf{Y}) \rangle - \langle \ln Q(\beta) \rangle
\end{aligned} \tag{25}$$

where the moments are expressed as:

$$\begin{aligned}
\langle \ln P(\mathbf{X}|\mathbf{Z}, \mathbf{Y}, \beta) \rangle &= \frac{ND}{2} \langle \ln \beta \rangle - \frac{ND}{2} \ln 2\pi \\
&\quad - \frac{\langle \beta \rangle}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \langle z_{kn} \rangle \left\langle \|\mathbf{x}_n - \mathbf{y}_k\|^2 \right\rangle
\end{aligned} \tag{26}$$

$$\langle \ln P(\mathbf{Z}) \rangle = \sum_{n=1}^{N} \sum_{k=1}^{K} \langle z_{kn} \rangle \ln p_{kn} \tag{27}$$

$$\langle \ln P(\mathbf{Y}) \rangle = -\frac{KD}{2} \ln 2\pi - \frac{D}{2} \ln |\mathbf{C}| - \frac{1}{2} \sum_{d=1}^{D} \left\langle \mathbf{y}_{(d)}^T \mathbf{C}^{-1} \mathbf{y}_{(d)} \right\rangle \tag{28}$$

$$\langle \ln P(\beta) \rangle = d_\beta \ln s_\beta - \ln \Gamma(d_\beta) + (d_\beta - 1) \langle \ln \beta \rangle - s_\beta \langle \beta \rangle \tag{29}$$

$$\langle \ln Q(\mathbf{Z}) \rangle = \sum_{n=1}^{N} \sum_{k=1}^{K} \langle z_{kn} \rangle \ln \tilde{p}_{kn} \tag{30}$$

$$\langle \ln Q(\mathbf{Y}) \rangle = -\frac{KD}{2} \ln 2\pi - \frac{D}{2} \ln \left| \tilde{\mathbf{\Sigma}} \right| - \frac{KD}{2} \tag{31}$$

$$\langle \ln Q(\beta) \rangle = \tilde{d}_\beta \ln \tilde{s}_\beta - \ln \Gamma\left(\tilde{d}_\beta \right) + \left(\tilde{d}_\beta - 1 \right) \langle \ln \beta \rangle - \tilde{s}_\beta \langle \beta \rangle \tag{32}$$

and

$$\langle \ln \beta \rangle = \psi\left(\tilde{d}_\beta \right) - \ln \tilde{s}_\beta \tag{33}$$

$$\left\langle \mathbf{y}_{(d)}^T \mathbf{C}^{-1} \mathbf{y}_{(d)} \right\rangle = \text{tr}\left[\mathbf{C}^{-1} \left(\tilde{\mathbf{\Sigma}} + \tilde{\mathbf{m}}^{(d)} \left(\tilde{\mathbf{m}}^{(d)} \right)^T \right) \right] \tag{34}$$

In the previous expressions, $\Gamma(\cdot)$ are Gamma functions, and $\psi(\cdot)$ is the Digamma function.

4 Conclusions and Future Work

In this brief report, the theoretical foundations of a Variational Bayesian GTM model have been laid. This model should prevent, even if partially, the phenomenon of data overfitting, and, therefore, show good generalization capabilities. Immediate future work should be directed towards assessing such capabilities as well as the limitations of the proposed model, through the design of experiments using noisy artificial data.

References

[1] Bishop, C.M., Svensen, M., Williams, C.R.I.: GTM: The Generative Topographic Mapping. Neural Comput. **10**(1) (1998) 215–234

[2] Kohonen, T.: Self-Organizing Maps (3rd ed). Springer-Verlag, Berlin (2001)

[3] Olier, I., Vellido, A.: Comparative assessment of the robustness of missing data imputation through Generative Topographic Mapping. In: Lect. Notes Comput. Sc. Volume 3512. (2005) 787–794

[4] Vellido, A.: Missing data imputation through GTM as a mixture of t-distributions. Neural Networks **19**(10) (2006) 1624–1635

[5] Vellido, A., Lisboa, P.J.G.: Handling outliers in brain tumour MRS data analysis through robust topographic mapping. Comput. Biol. Med. **3**(10) (2006) 1049–1063

[6] Vellido, A., Lisboa, P.J.G., Vicente, D.: Robust analysis of MRS brain tumour data using t-GTM. Neurocomputing **69**(7-9) (2006) 754–768

[7] Vellido, A.: Assessment of an unsupervised feature selection method for Generative Topographic Mapping. In: Lect. Notes Comput. Sc. Volume 4132. (2006) 361–370

[8] Olier, I., Vellido, A.: Time Series Relevance Determination through a topology-constrained Hidden Markov model. In: Lect. Notes Comput. Sc. Volume 4224. (2006) 40–47

[9] Olier, I., Vellido, A.: Capturing the dynamics of multivariate time series through visualization using Generative Topographic Mapping Through Time. In: Proceedings of IEEE ICEIS 2006, Islamabad, Pakistan. (2006)

[10] Bishop, C.M., Svensen, M., Williams, C.R.I.: Developments of the Generative Topographic Mapping. Neurocomputing **21**(1–3) (1998) 203–224

[11] Vellido, A., El-Deredy, W., Lisboa, P.J.G.: Selective smoothing of the Generative Topographic Mapping. IEEE T. Neural Networ. **14**(4) (2003) 847–852

[12] MacKay, D.J.C.: A practical Bayesian framework for back-propagation networks. Neural Comput. **4**(3) (1992) 448–472

[13] Andrieu, C., de Freitas, N., Doucet, A., Jordan, M.I.: An introduction to MCMC for machine learning. Mach. Learn. **50** (2003) 5–43

[14] Beal, M.: Variational algorithms for approximate Bayesian inference. PhD thesis, The Gatsby Computational Neuroscience Unit, Univ. College London (2003)

[15] Jakkola, T., Jordan, M.I.: Bayesian parameter estimation via variational methods. Stat. Comput. **10** (2000) 25–33

[16] Attias, H.: A variational Bayesian framework for graphical models. In: Lect. Notes Comput. Sc. (1999) 209–215

[17] Bishop, C.M.: Variational principal components. In: Proceedings Ninth Intern. Conf. on Artificial Neural Networks. Volume 1. (1999) 509–514

[18] Valpola, H.: Ensemble learning for Independent Component Analysis. In: Proceedings of the First International Workshop on Independent Component Analysis and Blind Signal Separation, ICA'99, Aussois, France. (1999) 7–12

[19] MacKay, D.J.C.: Ensemble learning for Hidden Markov Models. Unpublished Manuscript (1997)

[20] Abrahamsen, P.: A review of Gaussian random fields and correlation functions. Technical Report 917, Norwegian Computing Center, Oslo, Norway (1997)

[21] Utsugi, A.: Bayesian sampling and ensemble learning in Generative Topographic Mapping. Neural Process. Lett. **12** (2000) 277–290

A Computation of the Variational Parameters

Details on the derivation of some of the variational parameters of the model, starting from the Eq. 11, are provided next.

A.1 Derivation of $\tilde{p}_{kn}$

The variational distribution $Q(\mathbf{Z})$ takes the following form:

$$\ln Q(\mathbf{Z}) = \int Q(\mathbf{Y}) \int Q(\beta) \left[\ln P(\mathbf{X}|\mathbf{Z}, \mathbf{Y}, \beta) + \ln P(\mathbf{Z})\right] d\beta d\mathbf{Y} + c \tag{35}$$

where c groups the constant expressions. Then, the Eqs. 4, 7, and 13 are included in Eq. 35 to obtain:

$$\sum_{n=1}^{N}\sum_{k=1}^{K} z_{kn} \ln \tilde{p}_{kn} = \left\langle -\frac{\beta}{2} \sum_{n=1}^{N}\sum_{k=1}^{K} z_{kn} \|\mathbf{x}_n - \mathbf{y}_k\|^2 \right\rangle_{\mathbf{Y},\beta} + \sum_{n=1}^{N}\sum_{k=1}^{K} z_{kn} \ln p_{kn} + c \tag{36}$$

Thus, the variational parameter $\tilde{p}_{kn}$ is obtained from the Eq. 36 as follows:

$$\ln \tilde{p}_{kn} = \left\langle -\frac{\beta}{2} \|\mathbf{x}_n - \mathbf{y}_k\|^2 \right\rangle_{\mathbf{Y},\beta} + \ln p_{kn} \tag{37}$$

This equation is resolved for $\tilde{p}_{kn}$ leading the following expression:

$$\tilde{p}_{kn} = p_{kn} \exp\left(-\frac{\langle\beta\rangle}{2} \left\langle \|\mathbf{x}_n - \mathbf{y}_k\|^2 \right\rangle_{\mathbf{Y}}\right) \tag{38}$$

where $\langle\beta\rangle$ is the mean of β for the Gamma distribution $Q(\beta)$ given by Eq. 22 and the calculation of $\left\langle \|\mathbf{x}_n - \mathbf{y}_k\|^2 \right\rangle_{\mathbf{Y}}$ is shown in Eq. 23.

A.2 Derivation of $\tilde{\mathbf{\Sigma}}$ and $\tilde{\mathbf{m}}$

Solving Eq. 11 for $Q(\mathbf{Y})$ leads to the following expression:

$$\ln Q(\mathbf{Y}) = \int Q(\mathbf{Z}) \int Q(\beta) \left[\ln P(\mathbf{X}|\mathbf{Z}, \mathbf{Y}, \beta) + \ln P(\mathbf{Y})\right] d\beta d\mathbf{Z} + c \tag{39}$$

Now, using Eqs. 4, 5 and 14, Eq. 39 is solved first for $\tilde{\Sigma}$ and then for $\tilde{\mathbf{m}}_{(d)}$ as follows:

$$\begin{aligned} \sum_{d=1}^{D} \left(\mathbf{y}_{(d)} - \tilde{\mathbf{m}}_{(d)}\right)^T \tilde{\mathbf{\Sigma}}^{-1} \left(\mathbf{y}_{(d)} - \tilde{\mathbf{m}}_{(d)}\right) &= \langle\beta\rangle \sum_{n=1}^{N}\sum_{k=1}^{K} \langle z_{kn}\rangle \|\mathbf{x}_n - \mathbf{y}_k\|^2 \\ &\quad + \sum_{d=1}^{D} \mathbf{y}_{(d)}^T \mathbf{C}^{-1} \mathbf{y}_{(d)} + c \end{aligned} \tag{40}$$

then:

$$\tilde{\boldsymbol{\Sigma}} = \left(\langle \beta \rangle \sum_{n=1}^{N} \mathbf{G}_n + \mathbf{C}^{-1} \right)^{-1} \tag{41}$$

and

$$\tilde{\mathbf{m}}_{(d)} = \langle \beta \rangle \, \tilde{\boldsymbol{\Sigma}} \sum_{n=1}^{N} x_{nd} \langle \mathbf{z}_n \rangle \tag{42}$$

A.3 Derivation of $\tilde{d}_\beta$ and $\tilde{s}_\beta$

Once again, the Eq. 11 is solved in a similar way to the previous subsections, but now with respect to $Q(\beta)$ as follows:

$$\ln Q(\beta) = \int Q(\mathbf{Z}) \int Q(\mathbf{Y}) \left[\ln P(\mathbf{X}|\mathbf{Z}, \mathbf{Y}, \beta) + \ln P(\beta) \right] d\mathbf{Y} d\mathbf{Z} + c \tag{43}$$

$$\begin{aligned} \tilde{d}_\beta \ln \tilde{s}_\beta + \left(\tilde{d}_\beta - 1 \right) \ln \beta - \tilde{s}_\beta \beta \quad &= \quad \frac{ND}{2} \ln \beta - \frac{\beta}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \langle z_{kn} \rangle \left\langle \|\mathbf{x}_n - \mathbf{y}_k\|^2 \right\rangle \\ &\quad + d_\beta \ln s_\beta + (d_\beta - 1) \ln \beta - s_\beta \beta + c \end{aligned} \tag{44}$$

Equating the factors with respect to the term $\ln \beta$, the variational parameter $\tilde{d}_\beta$ is obtained:

$$\tilde{d}_\beta = \frac{ND}{2} + d_\beta \tag{45}$$

Similarly, equating with respect to β the variational parameter $\tilde{s}_\beta$ is obtained:

$$\tilde{s}_\beta = \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \langle z_{kn} \rangle \left\langle \|\mathbf{x}_n - \mathbf{y}_k\|^2 \right\rangle \tag{46}$$