# Supplementary Material

**Joint Discriminative Bayesian Dictionary and Classifier Learning**

## 1 Joint probability distribution

According to the proposed model, the joint probability distribution over the data of the $c^{\text{th}}$ class can be expressed as:

$$P(\{\mathbf{y}_i^c\}, \{\mathbf{h}_i^c\}, \mathbf{\Phi}, \mathbf{\Psi}, \{\mathbf{z}_i^c\}, \{\mathbf{s}_i^c\}, \{\mathbf{t}_i^c\}, \{\pi_k^c\}, \lambda_s^c, \lambda_t^c, \lambda_y, \lambda_h) =$$
$$\prod_{i=1}^{|\mathcal{I}_c|} \mathcal{N}\left(\mathbf{y}_i^c | \mathbf{\Phi}(\mathbf{z}_i^c \odot \mathbf{s}_i^c), 1/\lambda_{y_o}\mathbf{I}_L\right) \text{Gam}\left(\lambda_y | e_o, f_o\right) \mathcal{N}\left(\mathbf{h}_i^c | \mathbf{\Psi}(\mathbf{z}_i^c \odot \mathbf{t}_i^c), 1/\lambda_{h_o}\mathbf{I}_C\right) \text{Gam}\left(\lambda_h | e_o, f_o\right)$$
$$\prod_{k=1}^{|\mathcal{K}|} \mathcal{N}\left(\boldsymbol{\varphi}_k | \mathbf{0}, 1/\lambda_{\varphi_o}\mathbf{I}_L\right) \mathcal{N}\left(\boldsymbol{\psi}_k | \mathbf{0}, 1/\lambda_{\psi_o}\mathbf{I}_C\right)$$
$$\prod_{i=1}^{|\mathcal{I}_c|} \prod_{k=1}^{|\mathcal{K}|} \text{Bernoulli}\left(z_{ik}^c | \pi_{k_o}^c\right) \text{Beta}\left(\pi_k^c | \frac{a_o}{K}, \frac{b_o(K-1)}{K}\right)$$
$$\prod_{k=1}^{|\mathcal{K}|} \mathcal{N}\left(\mathbf{s}_i^c | \mathbf{0}, 1/\lambda_{s_o}^c \mathbf{I}_{|\mathcal{K}|}\right) \text{Gam}\left(\lambda_s^c | c_o, d_o\right) \mathcal{N}\left(\mathbf{t}_i^c | \mathbf{0}, 1/\lambda_{t_o}^c \mathbf{I}_{|\mathcal{K}|}\right) \text{Gam}\left(\lambda_t^c | c_o, d_o\right).$$

## 2 Gibbs sampling equations

We have made use of the following theorem [1] while driving the Gibbs Sampling equations for our model:

**Theorem 1 [1]:** If prior probability over $\mathbf{y_1}$ is given as $p(\mathbf{y}_1) = \mathcal{N}(\mathbf{y}_1|\boldsymbol{\mu}_o, \mathbf{\Lambda}_o^{-1})$ and the likelihood function is defined as $p(\mathbf{y}_2|\mathbf{y}_1) = \mathcal{N}(\mathbf{y}_2|\mathbf{A}\mathbf{y}_1 + \mathbf{b}, \mathbf{L^{-1}})$, then the posterior probability distribution over $\mathbf{y_1}$ can be written as $p(\mathbf{y}_1|\mathbf{y}_2) = \mathcal{N}(\mathbf{y}_1|\boldsymbol{\mu}, \mathbf{\Lambda}^{-1})$, where:

$$\mathbf{\Lambda} = \mathbf{\Lambda}_o + \mathbf{A}^T\mathbf{L}\mathbf{A}$$
$$\boldsymbol{\mu} = \mathbf{\Lambda}^{-1}(\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \mathbf{\Lambda}_o\boldsymbol{\mu}_o).$$

Below, we derive the sampling equations. The sampling is performed in our approach in an iterative manner. The sampling sequence is the same as the sequence of the equations given below.

**Sample $\boldsymbol{\varphi}_k$:** According to the proposed model, we can write the posterior distribution over the $k^{\text{th}}$ dictionary atom $p(\boldsymbol{\varphi}_k|-)$ as follows:

$$p(\boldsymbol{\varphi}_k|-) \propto \prod_{i=1}^{N} \mathcal{N}(\mathbf{y}_i|\mathbf{\Phi}(\mathbf{z}_i \odot \mathbf{s}_i), \lambda_{y_o}^{-1}\mathbf{I}_L)\mathcal{N}(\boldsymbol{\varphi}_k|\mathbf{0}, \lambda_{\varphi_o}^{-1}\mathbf{I}_L).$$

We can write the mean of the likelihood function in terms of $\boldsymbol{\varphi}_k$ as:

$$\mathbf{y}_{i_{\varphi_k}} = \mathbf{y}_i - \mathbf{\Phi}(\mathbf{z}_i \odot \mathbf{s}_i) + \boldsymbol{\varphi}_k(z_{ik} \odot s_{ik}).$$

where $\mathbf{y}_{i_{\varphi_k}}$ denotes the contribution of the $k^{\text{th}}$ dictionary atom in approximating $\mathbf{y}_i$. Hence, the posterior distribution over $\boldsymbol{\varphi}_k$ can be re-written as:

$$p(\boldsymbol{\varphi}_k|-) \propto \prod_{i=1}^{N} \mathcal{N}(\mathbf{y}_{i_{\varphi_k}}|\boldsymbol{\varphi}_k(z_{ik}.s_{ik}), \lambda_{y_o}^{-1}\mathbf{I}_L)\mathcal{N}(\boldsymbol{\varphi}_k|\mathbf{0}, \lambda_{\varphi_o}^{-1}\mathbf{I}_L).$$

Exploiting the results of Theorem 1, the posterior over the dictionary atoms can be expressed as:

$$p(\boldsymbol{\varphi}_k|-) = \mathcal{N}(\boldsymbol{\varphi}_k|\boldsymbol{\mu}_k, \lambda_{\varphi}^{-1}\mathbf{I}_L), \text{where,}$$

$$\lambda_{\varphi} = \lambda_{\varphi_o} + \lambda_{y_o}\sum_{i=1}^{N}(z_{ik}.s_{ik})^2, \quad \boldsymbol{\mu}_k = \lambda_{y_o}\lambda_{\varphi}^{-1}\sum_{i=1}^{N}(z_{ik}.s_{ik})\mathbf{y}_{i_{\varphi_k}}.$$

We have arrived at the above expressions by placing $\mathbf{A} = \sum_{i=1}^{N}(z_{ik}.s_{ik})$ and $\mathbf{b} = \mathbf{0}$ in the results of Theorem 1. Note that, we have intentionally dropped the super-script '$c$' from the above expressions. This is because, the dictionary atoms are updated using the training data of all the classes simultaneously. The same is true for updating the columns $\boldsymbol{\psi}_k$ of the classifier $\boldsymbol{\Psi}$.

**Sample $\boldsymbol{\psi}_k$:** The posterior distribution $p(\boldsymbol{\psi}_k|-)$ over the $k^{\text{th}}$ column of $\boldsymbol{\Psi}$ can be written as:

$$p(\boldsymbol{\psi}_k|-) \propto \prod_{i=1}^{N} \mathcal{N}(\mathbf{h}_i|\boldsymbol{\Psi}(\mathbf{z}_i \odot \mathbf{t}_i), \lambda_{h_o}^{-1}\mathbf{I}_C)\mathcal{N}(\boldsymbol{\psi}_k|\mathbf{0}, \lambda_{\psi_o}^{-1}\mathbf{I}_C).$$

With the same reasoning as for sampling $\boldsymbol{\varphi}_k$, we can sample $\boldsymbol{\psi}_k$ from $p(\boldsymbol{\psi}_k|-) = \mathcal{N}(\boldsymbol{\psi}_k|\boldsymbol{\mu}_k, \lambda_{\psi}^{-1}\mathbf{I}_C)$, where

$$\lambda_{\psi} = \lambda_{\psi_o} + \lambda_{h_o}\sum_{i=1}^{N}(z_{ik}.t_{ik})^2, \quad \boldsymbol{\mu}_k = \lambda_{h_o}\lambda_{k}^{-1}\sum_{i=1}^{N}(z_{ik}.t_{ik})\mathbf{h}_{i_{\psi_k}}.$$

**Sample $z_{ik}^c$:** Once the dictionary and the classifier have been sampled, we must sample $z_{ik}^c$ based on the updated dictionary and the classifier. The posterior probability distribution over $z_{ik}^c$ can be expressed as, $\forall i \in \mathcal{I}_c$, $\forall k \in \mathcal{K}$:

$$p(z_{ik}^c|-) \propto \mathcal{N}(\mathbf{y}_{i_{\varphi_k}}^c|\boldsymbol{\varphi}_k(z_{ik}^c.s_{ik}^c), \lambda_{y_o}^{-1}\mathbf{I}_L)\ \mathcal{N}(\mathbf{h}_{i_{\varphi_k}}^c|\boldsymbol{\psi}_k(z_{ik}^c.t_{ik}^c), \lambda_{h_o}^{-1}\mathbf{I}_C)\ \text{Bernoulli}(z_{ik}^c|\pi_{k_o}^c).$$

It is straight forward to show that based on the above mentioned posterior

$$p(z_{ik}^c = 1|-) \propto \pi_{k_o}^c.\exp\Big(-\frac{(\mathbf{y}_{i_{\varphi_k}}^c - \boldsymbol{\varphi}_k s_{ik}^c)^\intercal\lambda_{y_o}\mathbf{I}_L(\mathbf{y}_{i_{\varphi_k}}^c - \boldsymbol{\varphi}_k s_{ik}^c)}{2}\Big).\exp\Big(-\frac{(\mathbf{h}_{i_{\psi_k}} - \boldsymbol{\psi}_k t_{ik}^c)^\intercal\lambda_{h_o}\mathbf{I}_C(\mathbf{h}_{i_{\psi_k}} - \boldsymbol{\psi}_k t_{ik}^c)}{2}\Big)$$

$$\propto \pi_{k_o}^c \underbrace{\exp\Big(-\frac{\lambda_{y_o}}{2}\mathbf{y}_{i_{\varphi_k}}^{c\intercal}\mathbf{y}_{i_{\varphi_k}}^c\Big)}_{\xi_1}.\overbrace{\exp\Big(-\frac{\lambda_{y_o}}{2}(\boldsymbol{\varphi}_k^\intercal\boldsymbol{\varphi}_k s_{ik}^{c\,2} - 2s_{ik}^c\mathbf{y}_{i_{\varphi_k}}^{c\intercal}\boldsymbol{\varphi}_k)\Big)}^{\xi_2}...$$

$$.\underbrace{\exp\Big(-\frac{\lambda_{h_o}}{2}\mathbf{h}_{i_{\psi_k}}^{c\intercal}\mathbf{h}_{i_{\psi_k}}\Big)}_{\xi_3}.\overbrace{\exp\Big(-\frac{\lambda_{h_o}}{2}(\boldsymbol{\psi}_k^\intercal\boldsymbol{\psi}_k t_{ik}^{c\,2} - 2t_{ik}\mathbf{h}_{i_{\psi_k}}^{c\intercal}\boldsymbol{\psi}_k)\Big)}^{\xi_4}.$$

Let $p_1 = \pi^c_{k_o}\xi_1\xi_2\xi_3\xi_4$. We can derive an expression for $p(z^c_{ik} = 0|-)$ in a similar fashion, that comes out to be:

$$p(z^c_{ik} = 0|-) \propto (1 - \pi^c_{k_o}) \exp\Big( - \frac{\lambda_{y_o}}{2} \mathbf{y}^{c\intercal}_{i_{\varphi_k}} \mathbf{y}^c_{i_{\varphi_k}}\Big). \exp\Big( - \frac{\lambda_{h_o}}{2} \mathbf{h}^{c\intercal}_{i_{\psi_k}} \mathbf{h}^c_{i_{\psi_k}}\Big).$$

Let $p_o = (1 - \pi^c_{k_o})\xi_1\xi_3$. Using $p_1$ and $p_o$, $z^c_{ik}$ can be sampled from the following normalized Bernoulli distribution:

$$z^c_{ik} \sim \text{Bernoulli}\Big(\frac{p_1}{p_1 + p_0}\Big).$$

Simplifying further:

$$z^c_{ik} \sim \text{Bernoulli}\Big(\frac{\pi^c_{k_o}\xi}{1 - \pi^c_{k_o} + \xi\pi^c_{k_o}}\Big),$$

where, $\xi = \xi_2\xi_4$.

**Sample $s^c_{ik}$:** We can write the following regarding the posterior probability distribution over $s^c_{ik}$:

$$p(s^c_{ik}|-) \propto \mathcal{N}(\mathbf{y}^c_{i_{\varphi_k}}|\boldsymbol{\varphi}_k(z^c_{ik}.s^c_{ik}), \lambda^{-1}_{y_o}\mathbf{I}_L)\mathcal{N}(s^c_{ik}|0, \lambda^{-1}_{s_o}).$$

Exploiting the results of Theorem 1, $s^c_{ik}$ can be sampled from $\mathcal{N}(s^c_{ik}|\mu_s, \lambda^{-1}_s)$, where:

$$\begin{aligned}
\lambda_s &= \lambda_{s_o} + (\boldsymbol{\varphi}_k z^c_{ik})^\intercal \lambda_{y_o}\mathbf{I}_L(\boldsymbol{\varphi}_k z^c_{ik}) \\
&= \lambda_{s_o} + \lambda_{y_o} z^{c\,2}_{ik} \boldsymbol{\varphi}^\intercal_k \boldsymbol{\varphi}_k, \\
\mu_s &= \lambda^{-1}_s\Big((\boldsymbol{\varphi}_k z^c_{ik})^\intercal \lambda_{y_o}\mathbf{I}_L \;\; \mathbf{y}^c_{i_{\varphi_k}}\Big) \\
&= \lambda^{-1}_s \lambda_{y_o} z^c_{ik} \boldsymbol{\varphi}^\intercal_k \mathbf{y}^c_{i_{\varphi_k}}.
\end{aligned}$$

**Sample $t^c_{ik}$:** Using the same reasoning as for $s^c_{ik}$, we can sample $t^c_{ik}$ from $\mathcal{N}(t^c_{ik}|\mu_t, \lambda^{-1}_t)$, where:

$$\lambda_t = \lambda_{t_o} + \lambda_{h_o} z^{c\,2}_{ik} \boldsymbol{\psi}^\intercal_k \boldsymbol{\psi}_k, \quad \mu_t = \lambda^{-1}_t \lambda_{h_o} z^c_{ik} \boldsymbol{\psi}^\intercal_k \mathbf{h}^c_{i_{\psi_k}}.$$

**Sample $\pi_k$:** We can write the posterior distribution over $\pi^c_k$ as follows:

$$\begin{aligned}
p(\pi^c_k|-) &\propto \prod_{i \in \mathcal{I}_c} \text{Bernoulli}(z^c_{ik}|\pi^c_{k_o})\text{Beta}(\pi^c_{k_o}|a_o/K, b_o(K-1)/K) \\
&= {}^c\pi_{k_o}^{\sum_{i=1}^{|\mathcal{I}_c|} z^c_{ik}} (1 - \pi^c_{k_o})^{|\mathcal{I}_c| - \sum_{i=1}^{|\mathcal{I}_c|} z^c_{ik}} \times {}^c\pi_{k_o}^{\frac{a_o}{K} - 1} (1 - \pi^c_{k_o})^{\frac{b_o(K-1)}{K} - 1} \\
&= {}^c\pi_{k_o}^{\frac{a_o}{K} + \sum_{i=1}^{|\mathcal{I}_c|} z^c_{ik} - 1} (1 - \pi^c_{k_o})^{\frac{b_o(K-1)}{K} + |\mathcal{I}_c| - \sum_{i=1}^{|\mathcal{I}_c|} z^c_{ik} - 1} \\
&= \text{Beta}\left(\frac{a_o}{K} + \sum_{i=1}^{|\mathcal{I}_c|} z^c_{ik}, \frac{b_o(K-1)}{K} + |\mathcal{I}_c| - \sum_{i=1}^{|\mathcal{I}_c|} z^c_{ik}\right).
\end{aligned}$$

Thus, we sample $\pi^c_k$ from the above mentioned Beta probability distribution. Note that, in the above derivation we wrote $\pi^c_k$ as ${}^c\pi_k$ for readability only.

**Sample $\lambda_s^c$ :** To compute $\lambda_s^c$, we treat $s_{ik}^c$ for all the dictionary atoms simultaneously (we do the same for $\lambda_t^c$ below). We consider $\mathbf{s}_i^c \in \mathbb{R}^K$ to be a sample of a Gaussian distribution with isotropic precision. This simplification allows us to efficiently infer the posterior distribution over $\lambda_s^c$ without significantly compromising the performance of our approach. The posterior distribution over $\lambda_s^c$ can be expressed as:

$$\begin{aligned}
p(\lambda_s^c|-) &\propto \prod_{i\in\mathcal{I}_c} \mathcal{N}(\mathbf{s}_i^c|\mathbf{0}, 1/\lambda_{s_o}^c \mathbf{I}_{|\mathcal{K}|})\mathrm{Gam}(\lambda_s^c|c_o, d_o) \\
&= \frac{1}{(2\pi)^{\frac{|\mathcal{I}_c|.|\mathcal{K}|}{2}} \det(1/\lambda_{s_o}^c \mathbf{I}_{|\mathcal{K}|})^{\frac{|\mathcal{I}_c|}{2}}} \exp\Big(-\frac{\lambda_{s_o}^c}{2}\sum_{i=1}^{|\mathcal{I}_c|}\mathbf{s}_i^{c\intercal}\mathbf{s}_i^c\Big)\ \frac{1}{\Gamma(c_o)} h_o^{g_o} \lambda_{s_o}^{c\ d_o-1}\exp(-d_o\lambda_{s_o}^c)
\end{aligned}$$

where $\Gamma(.)$ is the well-known gamma function and $\det(.)$ denotes the determinant of a matrix. Neglecting the constants in the right hand side of the above equation, and making use of the property $\det(\lambda\mathbf{I}_{|\mathcal{K}|}) = \lambda^{|\mathcal{K}|}$:

$$\begin{aligned}
p(\lambda_s^c|-) &\propto \lambda_{s_o}^{c\ \frac{|\mathcal{I}_c|.|\mathcal{K}|}{2}} \exp\Big(-\frac{\lambda_{s_o}^c}{2}\sum_{i=1}^{|\mathcal{I}_c|}\mathbf{s}_i^{c\intercal}\mathbf{s}_i^c\Big)\ \lambda_{s_o}^{c_o-1}\exp(-d_o\lambda_{s_o}^c) \\
&= \lambda_{s_o}^{c\ \frac{|\mathcal{I}_c|.|\mathcal{K}|}{2}+c_o-1} \exp\Big(-\lambda_{s_o}^c\big(\frac{1}{2}\sum_{i=1}^{|\mathcal{I}_c|}\mathbf{s}_i^{c\intercal}\mathbf{s}_i^c + d_o\big)\Big) \\
&\propto \mathrm{Gam}\left(\frac{|\mathcal{I}_c||\mathcal{K}|}{2} + c_o, \frac{1}{2}\sum_{i=1}^{|\mathcal{I}_c|}\mathbf{s}_i^{c\intercal}\mathbf{s}_i^c + d_o\right).
\end{aligned}$$

Therefore, we sample $\lambda_s^c$ as:

$$\lambda_s^c \sim \mathrm{Gam}\left(\frac{|\mathcal{I}_c||\mathcal{K}|}{2} + c_o, \frac{1}{2}\sum_{i=1}^{|\mathcal{I}_c|}||\mathbf{s}_i^c||_2^2 + d_o\right),$$

where, $||.||_2$ denotes the $\ell_2$-norm of a vector.

**Sample $\lambda_t^c$ :** Similarly, we can sample $\lambda_t^c$ from the following Gamma probability distribution:

$$\lambda_t^c \sim \mathrm{Gam}\left(\frac{|\mathcal{I}_c||\mathcal{K}|}{2} + c_o, \frac{1}{2}\sum_{i=1}^{|\mathcal{I}_c|}||\mathbf{t}_i^c||_2^2 + d_o\right).$$

**Sample $\lambda_y$ :** The posterior over $\lambda_y$ can be written as:

$$p(\lambda_y|-) \propto \prod_{i=1}^{N}\mathcal{N}(\mathbf{y}_i|\mathbf{\Phi}(\mathbf{z}_i\odot\mathbf{s}_i), \lambda_{y_o}^{-1}\mathbf{I}_L)\mathrm{Gam}(\lambda_y|e_o, f_o).$$

Again, we have intentionally dropped the superscript '$c$' because the computation is performed over the training data of all classes simultaneously. Following similar steps as in the derivations for $\lambda_s^c$ and $\lambda_t^c$ we can show that $\lambda_y$ must be sampled as follows:

$$\lambda_y \sim \mathrm{Gam}\left(\frac{LN}{2} + e_o, \frac{1}{2}\sum_{i=1}^{N}||\mathbf{y}_i - \mathbf{\Phi}(\mathbf{z}_i\odot\mathbf{s}_i)||_2^2 + f_o\right).$$

**Sample $\lambda_h$ :** Correspondingly, $\lambda_h$ can be sampled as the following:

$$\lambda_h \sim \mathrm{Gam}\left(\frac{CN}{2} + e_o, \frac{1}{2}\sum_{i=1}^{N}||\mathbf{h}_i - \mathbf{\Psi}(\mathbf{z}_i\odot\mathbf{t}_i)||_2^2 + f_o\right).$$

# References

[1] Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006)