Generative, Fully Bayesian, Gaussian Pattern Classifier

Niko Brümmer
AGNITIO Research, South Africa

BOSARIS Workshop, November 2012

1 Introduction

Observable patterns live in $\mathbb{R}^N$ and belong to K different classes. We are given a *supervised training database*, $D = (\mathbf{X}, L)$. Here $\mathbf{X}$ represents patterns and L the corresponding true class labelsof patterns. Now we want to recognize the unknown classes to which new unlabelled *test* patterns belong. For this purpose we pretend there is a generative model, with unknown parameter $\mathbf{\Phi}$, that generates all train and test data. Although $\mathbf{\Phi}$ is unknown, it is assumed to be the *same* for train and test.

For the purposes of this exercise, we choose perhaps the simplest possible such generative model, which has multivariate normal, class-conditional distributions of the form:

$$P(\mathbf{x}|k, \mathbf{\Phi}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{\Lambda}^{-1}) \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^N$ is a pattern, $k \in \{1, 2, \ldots, K\}$ is a class index, $\boldsymbol{\mu}_k \in \mathbb{R}^N$ is the class conditional mean and $\mathbf{\Lambda}^{-1}$ is the N-by-N, *common within-class covariance* matrix. We shall refer to $\mathbf{\Lambda}$ as the *within-class precision*.

The model parameters are collectively referred to as $\mathbf{\Phi}$, where $\mathbf{\Phi} = (\mathbf{M}, \mathbf{\Lambda})$, and $\mathbf{M} = \begin{bmatrix}\boldsymbol{\mu}_1 & \boldsymbol{\mu}_2 & \cdots & \boldsymbol{\mu}_K\end{bmatrix}$.

Letting Π denote some prior for $\mathbf{\Phi}$, the fully Bayesian recipe[1] requires calculation of the *parameter posterior* $P(\mathbf{\Phi}|D, \Pi)$, which in turn gives the

[1] For a more complete derivation, see section 2.1 in Niko Brummer and Edward de Villiers, 'Integrating out model parameters in generative and discriminative classifiers', 2011, available online at: `http://sites.google.com/site/nikobrummer/bayesian_model_integration.pdf`.

predictive distribution:

$$P(\mathbf{x}|k, D, \Pi) = \int P(\mathbf{x}|k, \mathbf{\Phi}) P(\mathbf{\Phi}|D, \Pi)\, d\mathbf{\Phi} \tag{2}$$

where $\mathbf{x}$ is a test pattern of hypothesized class k. This predictive distribution is the end goal of the exercise, since it can be used in a straight-forward calculation to find the classification posterior:

$$P(k|\mathbf{x}, D, \Pi, \pi) = \frac{P_k P(\mathbf{x}|k, D, \Pi)}{\sum_{i=1}^{K} P_i P(\mathbf{x}|i, D, \Pi)} \tag{3}$$

where $\pi = (P_1, P_2, \ldots, P_K)$ is a given prior distribution over classes. Finally, the classification posterior can then be used to make minimum-expected-cost classification decisions.

In the rest of this document, we introduce notation for several probability distributions, motivate the form of the parameter prior and then derive the parameter posterior and predictive distribution.

2 Dramatis personae

Here we introduce notation and properties of the probability distributions which will play the following roles in this problem:

likelihood: product of multivariate Gaussians

prior for $\mathbf{\Lambda}$: Wishart

prior for $\mathbf{M}$: matrix normal, conditioned on $\mathbf{\Lambda}$.

joint prior/posterior for $\mathbf{M}, \mathbf{\Lambda}$: matrix normal Wishart

predictive distribution: multivariate T.

2.1 Multivariate Gaussian distribution

The density of the multivariate Gaussian or normal distribution, for dimensionality N, defined in terms of mean $\boldsymbol{\mu}_k$ and precision $\mathbf{\Lambda}$ is:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{\Lambda}^{-1}) = \frac{|\mathbf{\Lambda}|^{\frac{1}{2}}}{(2\pi)^{\frac{N}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)'\mathbf{\Lambda}(\mathbf{x} - \boldsymbol{\mu}_k)\right) \tag{4}$$

Our likelihood will be expressed as a product of Gaussians. Let $\mathbf{X}_k = \begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_{T_k}\end{bmatrix}$ represent T_k iid samples from this density, then

$$P(\mathbf{X}_k|\boldsymbol{\mu}_k, \mathbf{\Lambda}) = (2\pi)^{-\frac{T_k N}{2}} |\mathbf{\Lambda}|^{\frac{T_k}{2}} \exp(t_0 + t_1 + t_2) \tag{5}$$

where

$$t_0 = -\frac{1}{2}\operatorname{tr}(T_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k' \boldsymbol{\Lambda}), \qquad t_1 = \operatorname{tr}(\mathbf{f}_k \boldsymbol{\mu}_k' \boldsymbol{\Lambda}), \qquad t_2 = -\frac{1}{2}\operatorname{tr}(\mathbf{S}_k \boldsymbol{\Lambda})$$

which we have expressed in terms of the first and second order stats:

$$\mathbf{f}_k = \sum_{i=1}^{T_k} \mathbf{x}_i, \qquad\qquad \mathbf{S}_k = \mathbf{X}_k \mathbf{X}_k'$$

2.2 Wishart distribution

We use the notation $\boldsymbol{\Lambda} > 0$ to indicate that N-by-N matrix $\boldsymbol{\Lambda}$ is positive definite. The probability density of the *Wishart distribution* can be defined,[2] for $\boldsymbol{\Lambda} > 0$, as:

$$\mathcal{W}(\boldsymbol{\Lambda}|a, \mathbf{B}) = \frac{\left|\frac{1}{2}\mathbf{B}\right|^{\frac{a}{2}}}{\Gamma_N\left(\frac{a}{2}\right)} |\boldsymbol{\Lambda}|^{\frac{a-N-1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}(\mathbf{B}\boldsymbol{\Lambda})\right) \tag{6}$$

where $a > N - 1$; $\mathbf{B}$ is N-by-N positive definite; and Γ_N is the *multivariate gamma function*, defined as:

$$\Gamma_N(x) = \pi^{\frac{N(N-1)}{4}} \prod_{i=1}^{N} \Gamma\left(x + \frac{1-i}{2}\right) \tag{7}$$

with $\Gamma(x) = \Gamma_1(x)$ the usual gamma function. The expected value of the Wishart density is $\langle \boldsymbol{\Lambda} \rangle = a\mathbf{B}^{-1}$.

Since the Wishart density integrates to one, we find the useful result:

$$\int_{\boldsymbol{\Lambda}>0} |\boldsymbol{\Lambda}|^{\frac{a-N-1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}(\mathbf{B}\boldsymbol{\Lambda})\right) d\boldsymbol{\Lambda} = \Gamma_N\left(\frac{a}{2}\right) \left|\frac{1}{2}\mathbf{B}\right|^{-\frac{a}{2}} \tag{8}$$

The Wishart density will form the prior for the within-class precision, $\boldsymbol{\Lambda}$.

2.3 Matrix Normal

The prior for the class means, $\mathbf{M}$, will be formed by a matrix normal density, where the means are dependent on the precision *as well as each other*.

[2] Our notation for the Wishart is parametrized for convenience in terms of $\mathbf{B}$. In Bishop's book (appendix B), for example, the Wishart is defined in terms of $\mathbf{B}^{-1}$.

The *matrix normal* density, for an N-by-K matrix $\mathbf{M}$, can be expressed as:

$$\mathcal{M}(M|\mathbf{\Theta},\mathbf{R},\mathbf{\Lambda}) = \frac{|\mathbf{R}|^{\frac{N}{2}}\,|\mathbf{\Lambda}|^{\frac{K}{2}}}{(2\pi)^{\frac{NK}{2}}} \exp\left(-\frac{1}{2}\operatorname{tr}\big(\mathbf{R}(\mathbf{M}-\mathbf{\Theta})'\mathbf{\Lambda}(\mathbf{M}-\mathbf{\Theta})\big)\right) \tag{9}$$

where the location parameter, $\mathbf{\Theta}$ is N-by-K and where there are two positive definite precision parameters: $\mathbf{R}$, K-by-K and $\mathbf{\Lambda}$, N-by-N. The matrix normal is related to the multivariate Gaussian as follows:[3]

$$\mathcal{M}(M|\mathbf{\Theta},\mathbf{R},\mathbf{\Lambda}) = \mathcal{N}(\operatorname{vec}(\mathbf{M})|\operatorname{vec}(\mathbf{\Theta}),\mathbf{R}^{-1}\otimes\mathbf{\Lambda}^{-1}) \tag{10}$$

This prior can independently describe the prior belief about the location of the data and the accuracy of the resulting pattern recognizer. If for example $\mathbf{\Theta} = \mathbf{0}$ and $\mathbf{R} = \alpha\mathbf{I}$ is isotropic, then larger α would imply smaller accuracy, by forcing the means closer (relative to the distance induced by $\mathbf{\Lambda}$), but it would also force their common location to be near the (arbitrarily chosen) origin. By allowing a more general form for $\mathbf{R}$, we can achieve non-informativeness about the location, but still achieve a conservative (regularizing) prior belief about accuracy. See appendix A.

2.3.1 Marginal

Let c_k denote the k-th element on the diagonal of $\mathbf{R}^{-1}$. Then we can express the marginal for column k of $\mathbf{M}$ as:[4]

$$P(\boldsymbol{\mu}_k|\mathbf{\Theta},\mathbf{R},\mathbf{\Lambda}) = \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{\theta}_k, c_k\mathbf{\Lambda}^{-1}) \tag{11}$$

where $\boldsymbol{\theta}_k$ is column k of $\mathbf{\Theta}$.

2.4 Matrix Normal Wishart

For $\mathbf{M}$, N-by-K and $\mathbf{\Lambda}$, N-by-N positive definite, the joint *matrix normal Wishart* density can be expressed as:

$$\begin{aligned}
&\mathcal{MW}(\mathbf{M},\mathbf{\Lambda}|\mathbf{\Theta},\mathbf{R},a,\mathbf{B}) \\
&= \frac{|\mathbf{R}|^{\frac{N}{2}}\,|\mathbf{\Lambda}|^{\frac{K}{2}}}{(2\pi)^{\frac{NK}{2}}} \exp\left(-\frac{1}{2}\operatorname{tr}\big(\mathbf{R}(\mathbf{M}-\mathbf{\Theta})'\mathbf{\Lambda}(\mathbf{M}-\mathbf{\Theta})\big)\right) \mathcal{W}(\mathbf{\Lambda}|a,\mathbf{B}) \\
&= \frac{\left|\frac{1}{2}\mathbf{B}\right|^{\frac{a}{2}}}{\Gamma_N\left(\frac{a}{2}\right)} \frac{|\mathbf{R}|^{\frac{N}{2}}}{(2\pi)^{\frac{NK}{2}}} |\mathbf{\Lambda}|^{\frac{a+K-N-1}{2}} \exp(e_1+e_2+e_3)
\end{aligned} \tag{12}$$

[3]Here vec stacks the columns of a matrix into a single vector and $\otimes$ denotes the Kronecker matrix product. Keep in mind that $(\mathbf{R}\otimes\mathbf{\Lambda})^{-1} = \mathbf{R}^{-1}\otimes\mathbf{\Lambda}^{-1}$.

[4]See e.g. Bishop's book, equation (B.51).

where

$$e_1 = -\frac{1}{2}\operatorname{tr}(\mathbf{R}\mathbf{M}'\mathbf{\Lambda}\mathbf{M}), \quad e_2 = -\frac{1}{2}\operatorname{tr}\big((\mathbf{B}+\mathbf{\Theta}\mathbf{R}\mathbf{\Theta}')\mathbf{\Lambda}\big), \quad e_3 = \operatorname{tr}(\mathbf{\Theta}\mathbf{R}\mathbf{M}'\mathbf{\Lambda})$$

and where the parameters $\mathbf{\Theta}, \mathbf{R}, a, \mathbf{B}$ are as introduced above.

2.5 Multivariate T

Below, when expressing the predictive distribution, we shall need the solution to an integral of the form:

$$\begin{aligned}
&\int_{\mathbf{\Lambda}>0} \mathcal{N}(\mathbf{x}|\boldsymbol{\theta}, \beta^{-1}\mathbf{\Lambda}^{-1})\mathcal{W}(\mathbf{\Lambda}|a,\mathbf{B})\, d\mathbf{\Lambda} \\
&= \frac{\left|\frac{1}{2}\mathbf{B}\right|^{\frac{a}{2}} \beta^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}\Gamma_N\left(\frac{a}{2}\right)} \int |\mathbf{\Lambda}|^{\frac{a+1-N-1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\big((\beta\mathbf{d}\mathbf{d}'+\mathbf{B})\mathbf{\Lambda}\big)\right) d\mathbf{\Lambda}
\end{aligned} \tag{13}$$

where we have defined $\mathbf{d} = \mathbf{x} - \boldsymbol{\theta}$ for convenience. We solve the integral using (8), and then simplify using (7) and the matrix determinant lemma:[5]

$$\begin{aligned}
&\int_{\mathbf{\Lambda}>0} \mathcal{N}(\mathbf{x}|\boldsymbol{\theta}, \beta^{-1}\mathbf{\Lambda}^{-1})\mathcal{W}(\mathbf{\Lambda}|a,\mathbf{B})\, d\mathbf{\Lambda} \\
&= \frac{\left|\frac{1}{2}\mathbf{B}\right|^{\frac{a}{2}} \beta^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}\Gamma_N\left(\frac{a}{2}\right)} \Gamma_N\left(\frac{a+1}{2}\right) \left|\frac{1}{2}(\beta\mathbf{d}\mathbf{d}'+\mathbf{B})\right|^{-\frac{a+1}{2}} \\
&= \left(\frac{\beta}{\pi}\right)^{\frac{N}{2}} \frac{\Gamma_N\left(\frac{a+1}{2}\right)}{\Gamma_N\left(\frac{a}{2}\right)} \frac{|\mathbf{B}|^{\frac{a}{2}}}{|\beta\mathbf{d}\mathbf{d}'+\mathbf{B}|^{\frac{a+1}{2}}} \\
&= \left(\frac{\beta}{\pi}\right)^{\frac{N}{2}} \frac{\Gamma\left(\frac{a+1}{2}\right)}{\Gamma\left(\frac{a+1-N}{2}\right)} \frac{|\mathbf{B}|^{\frac{a}{2}}}{|\beta\mathbf{d}\mathbf{d}'+\mathbf{B}|^{\frac{a+1}{2}}} \\
&= \frac{\Gamma\left(\frac{a+1}{2}\right)}{\Gamma\left(\frac{a+1-N}{2}\right)} \left|\frac{\pi}{\beta}\mathbf{B}\right|^{-\frac{1}{2}} \left(1+\beta\mathbf{d}'\mathbf{B}^{-1}\mathbf{d}\right)^{-\frac{a+1}{2}} \\
&= \mathcal{T}_N(\mathbf{x}|\boldsymbol{\theta}, \beta^{-1}\mathbf{B}, a)
\end{aligned} \tag{14}$$

This is a *multivariate T distribution*, for which we have introduced the notation[6] $\mathcal{T}_N$.

[5] $|\mathbf{B}+\beta\mathbf{d}\mathbf{d}'| = (1+\beta\mathbf{d}'\mathbf{B}^{-1}\mathbf{d})\,|\mathbf{B}|$

[6] Again, our notation is for convenience here. It differs (just cosmetically) from the way Bishop (his appendix B), defines his 'multivariate Student's t' and also (again cosmetically) from the 'Box and Tiao T distribution' in Minka's 'Inferring a Gaussian'.

3 Parameter inference

Given the supervised database D and a conjugate prior, we do a Bayesian inference of the parameters $\mathbf{\Phi} = (\mathbf{M}, \mathbf{\Lambda})$. The likelihood is the product of Gaussians of all the data in D. We use the matrix normal Wishart as conjugate prior and obtain a posterior of the same form.

3.1 Likelihood

For the purpose of parameter inference, the supervised database, D is represented by the triple statistic of the form $T_k, \mathbf{f}_k, \mathbf{S}_k$, for each class $k \in \{1, 2, \ldots, K\}$. For convenience we define: $\mathbf{T}$ to be the diagonal matrix, with diagonal elements $T_1, T_2, \ldots, T_K$; $T = \mathrm{tr}(\mathbf{T})$; $\mathbf{S} = \sum_k \mathbf{S}_k$ and $\mathbf{F} = \begin{bmatrix}\mathbf{f}_1 \cdots \mathbf{f}_K\end{bmatrix}$. Recall that the mean parameters are represented as $\mathbf{M} = \begin{bmatrix}\boldsymbol{\mu}_1 \cdots \boldsymbol{\mu}_K\end{bmatrix}$.

Recalling section 2.1, the parameter likelihood is:

$$P(\mathbf{X}|\mathbf{M}, \mathbf{\Lambda}, L) \propto |\mathbf{\Lambda}|^{\frac{T}{2}} \exp(E_1 + E_2 + E_3) \tag{15}$$

where

$$E_1 = -\frac{1}{2}\mathrm{tr}(\mathbf{T}\mathbf{M}'\mathbf{\Lambda}\mathbf{M}), \qquad E_2 = -\frac{1}{2}\mathrm{tr}(\mathbf{S}\mathbf{\Lambda}), \qquad E_3 = \mathrm{tr}(\mathbf{F}\mathbf{M}'\mathbf{\Lambda})$$

3.2 Parameter prior

We assign a matrix normal Wishart prior, with a zero location parameter, $\mathbf{\Theta} = \mathbf{0}$. Letting $\Pi = (\mathbf{R}, a, \mathbf{B})$, our conjugate prior is of the form (recall section 2.4):

$$\begin{aligned} P(\mathbf{M}, \mathbf{\Lambda}|\Pi) &= \mathcal{MW}(\mathbf{M}, \mathbf{\Lambda}|\mathbf{0}, \mathbf{R}, a, \mathbf{B}) \\ &\propto |\mathbf{\Lambda}|^{\frac{a+K-N-1}{2}} \exp(e_1 + e_2) \end{aligned} \tag{16}$$

where

$$e_1 = -\frac{1}{2}\mathrm{tr}(\mathbf{R}\mathbf{M}'\mathbf{\Lambda}\mathbf{M}), \qquad e_2 = -\frac{1}{2}\mathrm{tr}(\mathbf{B}\mathbf{\Lambda})$$

We choose $\mathbf{R}$ as derived in appendix A:

$$\mathbf{R} = r\mathbf{I} - \frac{r^2}{Kr + \epsilon}\mathbf{1}\mathbf{1}' \tag{17}$$

where ϵ is very small, to be non-informative about the location of the data and where increasing r regularizes the solution by making means that are close together more probable.

3.3 Parameter posterior

The parameter posterior can now be derived as:

$$\begin{aligned} P(\mathbf{M}, \mathbf{\Lambda}|D, \Pi) &\propto P(\mathbf{M}, \mathbf{\Lambda}|\Pi) P(\mathbf{X}|\mathbf{M}, \mathbf{\Lambda}, L) \\ &\propto |\mathbf{\Lambda}|^{\frac{a+T+K-N-1}{2}} \exp(s_1 + s_2 + s_3) \end{aligned} \tag{18}$$

where

$$s_1 = -\frac{1}{2}\operatorname{tr}\big((\mathbf{R}+\mathbf{T})\mathbf{M}'\mathbf{\Lambda}\mathbf{M}\big), \quad s_2 = -\frac{1}{2}\operatorname{tr}\big((\mathbf{S}+\mathbf{B})\mathbf{\Lambda}\big), \quad s_3 = \operatorname{tr}(\mathbf{F}\mathbf{M}'\Lambda)$$

which also has a matrix normal Wishart form, so that:

$$P(\mathbf{M}, \mathbf{\Lambda}|D, \Pi) = \mathcal{MW}(\mathbf{M}, \mathbf{\Lambda}|\mathbf{M}^*, \mathbf{R}^*, a^*, \mathbf{B}^*) \tag{19}$$

where we can identify the parameters by comparing (18) with (12):

$$\begin{aligned} \mathbf{M}^* &= \mathbf{F}(\mathbf{R}^*)^{-1}, & \mathbf{R}^* &= \mathbf{R} + \mathbf{T} \\ a^* &= a + T, & \mathbf{B}^* &= \mathbf{B} + \mathbf{S} - \mathbf{F}(\mathbf{R}^*)^{-1}\mathbf{F}' \end{aligned}$$

3.3.1 Analysis

Here we analyze some properties of the parameter posterior. First, we examine the determinant of $\mathbf{R}^*$, which needs to be non-zero for the posterior to be proper:

$$|\mathbf{R}^*| = |r\mathbf{I} + \mathbf{T}| \, \frac{\epsilon + r\sum_{i=1}^{K} \frac{T_i}{r+T_i}}{\epsilon + rK} \tag{20}$$

We can stress this determinant by zeroing T_i, ϵ, r, but not all at once. We can keep the determinant non-zero in the following cases:

- If $r > 0$, then we can let $\epsilon = 0$ and we can even let one of the $T_i = 0$. We will be interested in this case if we consider open-set recognition, where there is no training data for a class.
- If all $T_i > 0$, then we can first let $r = 0$ and then let $\epsilon \to 0$ (or let $\epsilon = 0$ and then $r \to 0$), then we get $\mathbf{R}^* = \mathbf{T}$.

If we simultaneously let $\epsilon, r \to 0$ and one of the $T_i = 0$, then the posterior becomes improper. This corresponds to common sense: $T_i = 0$ says we have not observed data for class i, $r = 0$ says we cannot infer the location of the unseen class from the others (which we have seen), and $\epsilon = 0$ says we have no prior idea where the data is located.

We also need $a^* = a + T \geq N$. If we take $a \to 0$ in the non-informative limit, we need at least N data points in total, otherwise the posterior density becomes undefined.

3.3.2 With non-informative prior

Here we examine the form of the parameter posterior, when we take the prior parameters to the non-informative limits: $\epsilon \to 0$, $a \to 0$, $\mathbf{B} \to \mathbf{0}$.

If we let $\epsilon \to 0$, then we need $T \geq 1$ to make $\mathbf{R}^*$ and the matrix normal factor in the posterior non-singular. This can be shown by computing the determinant at $\epsilon = 0$:[7]

$$|\mathbf{R}^*|_{\epsilon=0} = |r\mathbf{I} + \mathbf{T}| \left(1 - \frac{1}{K} \sum_{i=1}^{K} \frac{r}{r + T_i}\right) \tag{21}$$

which becomes non-zero as soon as at least one of the $T_i > 0$. In what follows, we shall assume $\epsilon = 0$, unless otherwise stated.

If we also let $a \to 0$, then we need $a^* = T \geq N$ to ensure that the Wishart normalization factor $\Gamma_N\left(\frac{a^*}{2}\right)$ remains finite.

If we let $\mathbf{B} \to \mathbf{0}$, we also need $\mathbf{B}^* = \mathbf{S} - \mathbf{F}(\mathbf{R}^*)^{-1}\mathbf{F}' > 0$. To understand this better, we compute $(\mathbf{R}^*)^{-1}$:

$$(\mathbf{R}^*)^{-1} = (r\mathbf{I} + \mathbf{T})^{-1} + \frac{r\boldsymbol{\rho}\boldsymbol{\rho}'}{\epsilon + \sum_{i=1}^{K} \frac{T_i}{r+T_i}}, \text{ where } \quad \boldsymbol{\rho} = \begin{bmatrix} (r+T_1)^{-1} \\ \vdots \\ (r+T_K)^{-1} \end{bmatrix} \tag{22}$$

Provided $T > 0$, this remains well-behaved at $\epsilon = 0$, for all $0 \leq r \leq \infty$. Notice that:

$$(\mathbf{R}^*)^{-1}\big|_{\epsilon=0,r=0} = \mathbf{T}^{-1}, \qquad (\mathbf{R}^*)^{-1}\big|_{\epsilon=0,r\to\infty} = T^{-1}\mathbf{1}\mathbf{1}' \tag{23}$$

Recalling $\mathbf{M}^* = [\boldsymbol{\mu}_1^* \cdots \boldsymbol{\mu}_K^*] = [\mathbf{f}_1 \cdots \mathbf{f}_K](\mathbf{R}^*)^{-1}$, we see:

$$\boldsymbol{\mu}_i^*|_{\epsilon=0,r=0} = T_i^{-1}\mathbf{f}_i, \qquad \boldsymbol{\mu}_i^*|_{\epsilon=0,r\to\infty} = \left(\sum_{i=1}^{K} T_i\right)^{-1} \sum_{i=1}^{K} \mathbf{f}_i \tag{24}$$

With no regularization, at $r = 0$, the posterior expectations for the class means are just the averages of the data for each class. With infinite regularization, all the class means are forced to be identical and equal to the global average.

[7] We can use the matrix inversion lemma, because $\mathbf{R}^*$ is a rank one update of the invertible, diagonal matrix $r\mathbf{I} + \mathbf{T}$.

4 Predictive distribution

We can now finally derive the predictive distribution, which is the end goal of the exercise. Thanks to the conjugacy, the integral over the parameter posterior can be found in closed form:

$$\begin{aligned} & P(\mathbf{x}|k, D, \Pi) \\ &= \int P(\mathbf{x}|k, \mathbf{\Phi}) P(\mathbf{\Phi}|D, \Pi)\, d\mathbf{\Phi} \\ &= \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{\Lambda}^{-1}) \mathcal{MW}(\mathbf{M}, \mathbf{\Lambda}|\mathbf{M}^*, \mathbf{R}^*, a^*, \mathbf{B}^*)\, d\boldsymbol{\mu}_1 \cdots d\boldsymbol{\mu}_K\, d\mathbf{\Lambda} \\ &= \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{\Lambda}^{-1}) \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{\mu}_k^*, c_k^* \mathbf{\Lambda}^{-1})\, d\boldsymbol{\mu}_k \mathcal{W}(\mathbf{\Lambda}|a^*, \mathbf{B}^*)\, d\mathbf{\Lambda} \end{aligned} \tag{25}$$

where we used the result of section 2.3.1, with c_k^* the k-th diagonal element of $(\mathbf{R}^*)^{-1}$ and $\boldsymbol{\mu}_k^*$ the k-th column of $\mathbf{M}^*$. Next, we integrate out $\boldsymbol{\mu}_k$ by simply adding variances and finally use (14) to integrate out $\mathbf{\Lambda}$:

$$\begin{aligned} P(\mathbf{x}|k, D, \Pi) &= \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k^*, (c_k^* + 1)\mathbf{\Lambda}^{-1}) \mathcal{W}(\mathbf{\Lambda}|a^*, \mathbf{B}^*)\, d\mathbf{\Lambda} \\ &= \mathcal{T}_N(\mathbf{x}|\boldsymbol{\mu}_k^*, (c_k^* + 1)\mathbf{B}^*, a^*) \end{aligned} \tag{26}$$

4.1 Non-informative prior

We let $\epsilon \to 0$, so that $\mathbf{R} = r\mathbf{C}_K$ and $\mathbf{R}^* = \mathbf{T} + r\mathbf{C}_K$. We let $a \to 0$, so that $a^* = T$ and $\mathbf{B} \to \mathbf{0}$, so that $\mathbf{B}^* = \mathbf{S} - \mathbf{F}(\mathbf{R}^*)^{-1}\mathbf{F}'$. For recognizing the class of $\mathbf{x}$, via the posterior (3), we can ignore all factors of (26) which are not conditioned on the class. This gives:

$$P(\mathbf{x}|k, D, \Pi) \propto (c_k^* + 1)^{-\frac{N}{2}} \left(1 + \frac{(\mathbf{x} - \boldsymbol{\mu}_k^*)'(\mathbf{B}^*)^{-1}(\mathbf{x} - \boldsymbol{\mu}_k^*)}{c_k^* + 1}\right)^{-\frac{T+1}{2}} \tag{27}$$

where $\boldsymbol{\mu}_k^*$ is column k of $\mathbf{M}^* = \mathbf{F}(\mathbf{R}^*)^{-1}$.

This solution has the desirable property that it is invariant to invertible affine transformations of the data: if we transform the train and test data with the same transform, then the classification posterior will not change.

5 Model evidence

Here we assume $a, \mathbf{B}, \epsilon$ are given (we will eventually take them to the non-informative limits at zero), but we are interested in how the model changes as

a function of r, so that we can decide which value to use for it, possibly using an ML or MAP to estimate. Recall that the supervised database is denoted $\mathbf{X}, L$, where $\mathbf{X}$ represents the data for all the classes and L represents the class labels. We need to compute the model *evidence*:

$$\begin{aligned} P(\mathbf{X}|r, L, a, \mathbf{B}, \epsilon) &= \int P(\mathbf{X}, \mathbf{M}, \mathbf{\Lambda}|r, L, a, \mathbf{B}, \epsilon)\, d\mathbf{M}\, d\mathbf{\Lambda} \\ &= \int P(\mathbf{X}|\mathbf{M}, \mathbf{\Lambda}, L) P(\mathbf{M}, \mathbf{\Lambda}|r, a, \mathbf{B}, \epsilon)\, d\mathbf{M}\, d\mathbf{\Lambda} \end{aligned} \tag{28}$$

Here $P(\mathbf{M}, \mathbf{\Lambda}|r, a, \mathbf{B}, \epsilon)$ is the matrix normal Wishart parameter prior as defined in sections 3.2 and 2.4. Omitting the constant $(2\pi)^{-\frac{NK}{2}}$, we have:

$$\begin{aligned} P(\mathbf{M}, \mathbf{\Lambda}|r, a, \mathbf{B}) \propto & \frac{\left|\frac{1}{2}\mathbf{B}\right|^{\frac{a}{2}}}{\Gamma_N\left(\frac{a}{2}\right)} \\ & |\mathbf{R}|^{\frac{N}{2}}\, |\mathbf{\Lambda}|^{\frac{a+K-N-1}{2}} \exp\Big[-\frac{1}{2}\operatorname{tr}(\mathbf{R}\mathbf{M}'\mathbf{\Lambda}\mathbf{M}) - \frac{1}{2}\operatorname{tr}(\mathbf{B}\mathbf{\Lambda})\Big] \end{aligned} \tag{29}$$

Recall that $\mathbf{R}$ is dependent on r and ϵ. Since we are interested only in inferring r, we could omit $\frac{\left|\frac{1}{2}\mathbf{B}\right|^{\frac{a}{2}}}{\Gamma_N\left(\frac{a}{2}\right)}$, but we show them here to stress the fact that if $a = 0$ and $\mathbf{B} = \mathbf{0}$, then the prior and hence also the evidence become improper. The same holds at $\epsilon = 0$.

The other factor in the integrand is the likelihood, given by (15):

$$P(\mathbf{X}|\mathbf{M}, \mathbf{\Lambda}, L) \propto |\mathbf{\Lambda}|^{\frac{T}{2}} \exp\Big[-\frac{1}{2}\operatorname{tr}(\mathbf{T}\mathbf{M}'\mathbf{\Lambda}\mathbf{M}) - \frac{1}{2}\operatorname{tr}(\mathbf{S}\mathbf{\Lambda}) + \operatorname{tr}(\mathbf{F}\mathbf{M}'\mathbf{\Lambda})\Big] \tag{30}$$

Multiplying prior and likelihood, we get the integrand:

$$\begin{aligned} \mathcal{I}(\mathbf{M}, \mathbf{\Lambda}) = & |\mathbf{R}|^{\frac{N}{2}}\, |\mathbf{\Lambda}|^{\frac{T+a+K-N-1}{2}} \\ & \exp\Big[-\frac{1}{2}\operatorname{tr}((\mathbf{T}+\mathbf{R})\mathbf{M}'\mathbf{\Lambda}\mathbf{M}) - \frac{1}{2}\operatorname{tr}((\mathbf{S}+\mathbf{B})\mathbf{\Lambda}) + \operatorname{tr}(\mathbf{F}\mathbf{M}'\mathbf{\Lambda})\Big] \end{aligned} \tag{31}$$

This is proportional to the matrix normal Wishart parameter posterior that we found above, with parameters $a^*, \mathbf{B}^*, \mathbf{M}^*, \mathbf{R}^*$. If we nomalize this, the integral is one, so that[8]:

$$\int \mathcal{I}(\mathbf{M}, \mathbf{\Lambda})\, d\mathbf{M}\, d\mathbf{\Lambda} \propto |\mathbf{R}|^{\frac{N}{2}} \left|\frac{1}{2}\mathbf{B}^*\right|^{-\frac{a^*}{2}} |\mathbf{R}^*|^{-\frac{N}{2}} \tag{32}$$

We can simplify:

$$\frac{|\mathbf{R}|}{|\mathbf{R}^*|} = \frac{r^K}{\prod_{k=1}^K (r + T_i)} \times \frac{\epsilon}{\epsilon + r\sum_{k=1}^K \frac{T_i}{r+T_i}} \tag{33}$$

[8] omitting factors independent of r

A Appendix: Inducing dependence between the means

Let there be a common location parameter $\boldsymbol{\theta}$ and let the class means be distributed about this location as $\mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{\theta}, (r\boldsymbol{\Lambda})^{-1})$. Increasing the precision parameter r, *decreases* prior belief in accuracy, because then it becomes more unlikely that the means will be far apart. The joint density for $\mathbf{M} = \begin{bmatrix}\boldsymbol{\mu}_1 \cdots \boldsymbol{\mu}_K\end{bmatrix}$ can be expressed conveniently as:

$$\mathcal{M}(\mathbf{M}|\boldsymbol{\theta}\mathbf{1}', r\mathbf{I}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{\theta}, (r\boldsymbol{\Lambda})^{-1})$$

where $\mathbf{1}'$ is a row of K ones. Now let the location parameter be distributed as $\mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, (\epsilon\boldsymbol{\Lambda})^{-1})$, which we can make non-informative by making ϵ very small, and we integrate out $\boldsymbol{\theta}$ to find the marginal:[9]

$$\begin{aligned}
&P(\mathbf{M}|r, \epsilon, \boldsymbol{\Lambda}) \\
&= \int_{\mathbb{R}^N} \mathcal{M}(\mathbf{M}|\boldsymbol{\theta}\mathbf{1}', r\mathbf{I}, \boldsymbol{\Lambda}) \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, (\epsilon\boldsymbol{\Lambda})^{-1})\, d\boldsymbol{\theta} \\
&\propto \int_{\mathbb{R}^N} \exp\Big(-\frac{r}{2}\operatorname{tr}\big(r(\mathbf{M} - \boldsymbol{\theta}\mathbf{1}')'\boldsymbol{\Lambda}(\mathbf{M} - \boldsymbol{\theta}\mathbf{1}')\big)\Big) \exp\Big(-\frac{\epsilon}{2}\boldsymbol{\theta}'\boldsymbol{\Lambda}\boldsymbol{\theta}\Big)\, d\boldsymbol{\theta} \\
&= \exp\Big(-\frac{r}{2}\operatorname{tr}(\mathbf{M}'\boldsymbol{\Lambda}\mathbf{M})\Big) \int_{\mathbb{R}^N} \exp\Big(-\frac{rK + \epsilon}{2}\boldsymbol{\theta}'\boldsymbol{\Lambda}\boldsymbol{\theta} + r\boldsymbol{\theta}'\boldsymbol{\Lambda}\mathbf{M}\mathbf{1}\Big)\, d\boldsymbol{\theta} \\
&\propto \exp\left(-\frac{r}{2}\operatorname{tr}(\mathbf{M}'\boldsymbol{\Lambda}\mathbf{M}) + \frac{r}{2}\frac{1}{K + \frac{\epsilon}{r}}\operatorname{tr}(\mathbf{1}\mathbf{1}'\mathbf{M}'\boldsymbol{\Lambda}\mathbf{M})\right) \\
&\propto \mathcal{M}(\mathbf{M}|\mathbf{0}, \mathbf{R}, \boldsymbol{\Lambda})
\end{aligned} \tag{34}$$

where

$$\mathbf{R} = r\left(\mathbf{I} - \frac{1}{K + \frac{\epsilon}{r}}\mathbf{1}\mathbf{1}'\right) \tag{35}$$

Notice that if $\epsilon = 0$, then $\mathbf{R} = r\mathbf{C}_K$, where $\mathbf{C}_K$ is the *centering matrix* of size K, which is singular. In this case, $\mathcal{M}(\mathbf{M}|\mathbf{0}, \mathbf{R}, \boldsymbol{\Lambda})$ is improper. For $\epsilon > 0$, $\mathbf{R}$ is diagonally dominant and therefore positive definite.

We can use the matrix inversion and matrix determinant lemmas to find:

$$\mathbf{R}^{-1} = r^{-1}\mathbf{I} + \epsilon^{-1}\mathbf{1}\mathbf{1}', \qquad |\mathbf{R}| = \frac{\epsilon r^K}{\epsilon + Kr} \tag{36}$$

[9] We follow these steps: (i) Ignore all factors not dependent on $\mathbf{M}$ or $\boldsymbol{\theta}$. (ii) Complete the square involving $\boldsymbol{\theta}$ terms to form a Gaussian, which integrates to 1. (iii) Identify the parameters of the resulting matrix normal distribution.

Notice that the elements of the covariance matrix $\mathbf{R}^{-1}$ are just sums of the compounded variances r^{-1} and ϵ^{-1}.

A.1 Caution

Caution is advisable when working with improper priors and it is good practice to *not* immediately set $\epsilon = 0$ and then continue with the improper result. A non-zero ϵ should be retained until the end of the whole calculation and if the result is a function of ϵ, then as a last step one may pass to the limit as $\epsilon \to 0$. In this problem, we shall find that this limit is unproblematic: even if there is just a single data point, the *posterior* matrix normal distribution remains proper at $\epsilon \to 0$.

A.2 Centering and WCCN

In order to get a better intuitive understanding of the role that this prior plays, we discuss its form at the limit $\epsilon \to 0$:

$$\mathcal{M}(\mathbf{M}|\mathbf{0}, r\mathbf{C}_K, \mathbf{\Lambda}) \propto \exp\Big(-\frac{r}{2}\operatorname{tr}(\mathbf{C}_K\mathbf{M}'\mathbf{\Lambda}\mathbf{M})\Big) \tag{37}$$

This can be interpreted by pointing out that the centering matrix, $\mathbf{C}_K$, is symmetric and idempotent[10], so that this prior effectively forms a *regularization penalty* on $\mathbf{M}$ of the form:

$$r\operatorname{tr}(\mathbf{C}_K'\mathbf{M}'\mathbf{\Lambda}\mathbf{M}\mathbf{C}_K) \tag{38}$$

where multiplication by $\mathbf{C}_K$ effects centering, and multiplication by $\mathbf{\Lambda}$ effects within class covariance normalization (WCCN).

[10] $\mathbf{C}_K = \mathbf{C}_K'$ and $\mathbf{C}_K^2 = \mathbf{C}_K$.