Haziq Jamil
Department of Statistics
London School of Economics and Political Science
January 13, 2018

Chapter 5

I-priors for Categorical Responses

5.1 Preliminary

Observe data $\{(y_1, x_1), \dots, (y_n, x_n)\}$ where each $x_i \in \mathcal{X}$. Let $y_i \in \{1, \dots, m\}$ and write $y_i = (y_{i1}, \dots, y_{im})$ where $y_{ij} = 1$ if $y_i = j$ and 0 otherwise. For $j = 1, \dots, m$, attempt to model

$$
\begin{gathered}
y_{ij} = \alpha_j + f_j(x_i) + \epsilon_{ij} \\
(\epsilon_{i1}, \dots, \epsilon_{im})^\top \overset{\text{iid}}{\sim} \mathrm{N}_m(\mathbf{0}, \mathbf{\Psi}^{-1})
\end{gathered}
$$

Define $f_j(x) = f(x, j)$ such that each f_j belong to some RKHS $\mathcal{F}$ (and not separate RKHSs $\mathcal{F}_j$). The reproducing kernel of $\mathcal{F}$ is $h : (\mathcal{X} \times \{1, \dots, m\})^2 \to \mathbb{R}$ as defined by

$$
h\big((x, j), (x', j')\big) = a(j, j') h_\eta(x, x').
$$

Choices for $a : \{1, \dots, m\} \times \{1, \dots, m\} \to \mathbb{R}$ include

1. The Pearson kernel
$$
a(j, j') = \frac{\delta_{jj'}}{\mathrm{P}(X = j)} - 1
$$

2. The Identity kernel
$$
a(j, j') = \delta_{jj'}
$$

The kernel h_η may be any of the usual kernels, i.e. canonical, fBm, Pearson, SE, polynomial, etc. and this kernel depends on the hyperparameters η. Denote $\mathbf{H}$ as the $n \times n$ matrix with (r, s) entries equal to $h_\eta(x_r, x_s)$ for $r, s \in \{1, \dots, \}$. Similarly, denote $\mathbf{A}$ as the $m \times m$ matrix with (k, j) entries equal to $a(k, j)$. Note that for the identity kernel, $\mathbf{A} = \mathbf{I}_m$.

The regression model in vector form:

$$\overbrace{\begin{pmatrix} y_{i1} \\ \vdots \\ y_{im} \end{pmatrix}}^{\mathbf{y}_i} = \overbrace{\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix}}^{\boldsymbol{\alpha}} + \overbrace{\begin{pmatrix} f_1(x_i) \\ \vdots \\ f_m(x_i) \end{pmatrix}}^{\mathbf{f}(x_i)} + \overbrace{\begin{pmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{im} \end{pmatrix}}^{\boldsymbol{\epsilon}_i}$$

$$\boldsymbol{\epsilon}_i \overset{\text{iid}}{\sim} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Psi}^{-1}).$$

An I-prior on the regression function $f : \mathcal{X} \times \{1, \dots, m\} \to \mathbb{R}$ takes the form

$$f_j(x) = f(x, j) = \sum_{k=1}^{m} \sum_{i=1}^{n} a(j,k) h_\eta(x, x_i) w_{ij}$$

where $(w_{i1}, \dots, w_{im})^\top \overset{\text{iid}}{\sim} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Psi})$.

Rearrange the n observations per class. Let $\mathbf{f}_j = \big(f_j(x_1), \dots, f_j(x_n)\big)^\top \in \mathbb{R}^n$. We can write the I-prior as $\mathbf{f}_j = \mathbf{A}_{jj} \cdot \mathbf{H}\mathbf{w}_j$ Therefore, $\mathbf{f}_j \sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi}_{jj}\mathbf{A}_{jj}\mathbf{H}^2)$, and

$$\begin{aligned} \mathrm{Cov}(\mathbf{f}_j, \mathbf{f}_k) &= \mathrm{Cov}(\mathbf{A}_{jj} \cdot \mathbf{H}\mathbf{w}_j, \mathbf{A}_{kk} \cdot \mathbf{H}\mathbf{w}_k) \\ &= \mathbf{A}_{jj}\mathbf{A}_{kk} \cdot \mathbf{H} \, \mathrm{Cov}(\mathbf{w}_j, \mathbf{w}_k)\mathbf{H} \\ &= \mathbf{A}_{jj}\mathbf{A}_{kk}\boldsymbol{\Psi}_{jk}\mathbf{H}^2. \end{aligned}$$

Write $\mathbf{f} = (\mathbf{f}_1^\top, \dots, \mathbf{f}_m^\top)^\top \mathbb{R}^{nm}$. Then this is multivariate normal with mean $\mathbf{0}$ and covariance matrix equal to

$$\begin{aligned} \mathrm{Var}\, \mathbf{f} &= (\mathbf{A} \otimes \mathbf{H})(\boldsymbol{\Psi} \otimes \mathbf{I}_n)(\mathbf{A} \otimes \mathbf{H}) \\ &= (\mathbf{A}\boldsymbol{\Psi}\mathbf{A} \otimes \mathbf{H}^2) \\ &= \big(\boldsymbol{\Omega}_{jk}\mathbf{H}^2\big)_{j,k=1}^{n} \end{aligned}$$

where $\boldsymbol{\Omega}_{jk} = (\mathbf{A}\boldsymbol{\Psi}\mathbf{A})_{jk}$. Out of interest, this can be expressed as a matrix normal distribution. Write $\mathbf{w}$ as the $n \times m$ matrix with entries equal to w_{ij}. Then $\mathbf{w} \sim \mathrm{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$ and $\mathbf{f} = \mathbf{H}\mathbf{w}\mathbf{A} \sim \mathrm{MN}_{n,m}(\mathbf{0}, \mathbf{H}^2, \mathbf{A}\boldsymbol{\Psi}\mathbf{A})$.

5.1.1 Special case: $\boldsymbol{\Psi} = \mathbf{I}_m$ with identity kernel

In this case, the matrix normal distribution for $\mathbf{f}$ is $\mathrm{MN}_{n,m}(\mathbf{0}, \mathbf{H}^2, \mathbf{I}_m)$. That is to say, the *columns* of $\mathbf{f}$, i.e. $\mathbf{f}_j$, are iid observations $\mathbf{f}_j \overset{\text{iid}}{\sim} \mathrm{N}_n(\mathbf{0}, \mathbf{H}^2)$.