To-do list

Contents

Haziq Jamil
Department of Statistics
London School of Economics and Political Science
August 16, 2017

Chapter 6

Estimation of I-probit models using variational inference

In this chapter we provide the details of the variational algorithm to estimate categorical I-prior models.

6.1 Relevant distributions

$$p(\mathbf{y}|\mathbf{y}^*) = \prod_{i=1}^n \prod_{j=1}^m p_{ij} = \prod_{i=1}^n \prod_{j=1}^m \mathbb{1}[y^*_{ij} = \max_k y^*_{ik}]^{\mathbb{1}[y_i=j]}$$

$$\begin{aligned}
p(\mathbf{y}^*|\mathbf{f}) &= \prod_{i=1}^n \prod_{j=1}^m \mathrm{N}(f_{ij}, 1) \\
&= \exp\left[-\frac{nm}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^m (y^*_{ij} - f_{ij})^2 \right] \\
&= \exp\left[-\frac{nm}{2}\log 2\pi - \frac{1}{2}\|\mathbf{y}^* - \mathbf{f}\|^2 \right]
\end{aligned}$$

$$f_{ij} = \alpha_j + \sum_{k=1}^n h_{\lambda_j}(x_i, x_k) w_{kj}$$

$$\begin{aligned}
p(\mathbf{w}) &= \prod_{i=1}^n \prod_{j=1}^m p(w_{ij}) \\
&= \prod_{i=1}^n \prod_{j=1}^m \mathrm{N}(0, 1) \\
&= \exp\left[-\frac{nm}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^m w_{ij}^2 \right] \\
&= \exp\left[-\frac{nm}{2}\log 2\pi - \frac{1}{2}\mathbf{w}^\top \mathbf{w} \right]
\end{aligned}$$

$$p(\lambda, \alpha) \propto \text{const.}$$

6.2 Mean field distributions

$$\begin{aligned} p(\mathbf{y}^*, \mathbf{w}, \alpha, \lambda | \mathbf{y}) &\equiv q(\mathbf{y}^*) q(\mathbf{w}) q(\lambda) q(\alpha) \\ &\equiv \prod_{i,j} q(y^*_{ij}) q(\mathbf{w}) q(\lambda) q(\alpha) \end{aligned}$$

The first line is by assumption, while the second line follows from an induced factorisation, as we will see later. Denote by $\tilde{q}$ the distributions which minimise the KL divergence (maximises the lower bound). Then, for each of $\xi \in \{\mathbf{y}^*, \mathbf{w}, \alpha, \lambda\}$, $\tilde{q}$ satisfies

$$\log \tilde{q}(\xi) = \mathrm{E}_{-\xi}[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \alpha, \lambda)] + \text{const.}$$

$\tilde{q}(\mathbf{y}^*)$

In this subsection, we use the notation $y^*_i = (y^*_{i1}, \ldots, y^*_{im})$ to denote the vector of length m containing the latent variables for response i. The joint distribution for $\mathbf{y}^* = (y^*_1, \ldots, y^*_n)^\top$ is a product of the distribution for each of the components y^*_i - this is a consequence of the independence structure across observations. Therefore, we can consider the variational density for each y^*_i separately.

Consider the case where y_i takes one particular value $j \in \{1, \ldots, m\}$. The mean-field density $q(y^*_i)$ for each $i = 1, \ldots, n$ is found to be

$$\begin{aligned} \log \tilde{q}(y^*_i) &= \mathbb{1}[y^*_{ij} = \max_k y^*_{ik}] \cdot \mathrm{E}_{\mathbf{w},\alpha,\lambda} \left[-\frac{1}{2} \sum_{k=1}^m (y^*_{ik} - f_{ik})^2 \right] + \text{const.} \\ &= \mathbb{1}[y^*_{ij} = \max_k y^*_{ik}] \cdot \left[-\frac{1}{2} \sum_{k=1}^m (y^*_{ik} - \tilde{f}_{ik})^2 \right] + \text{const.} \\ &\equiv \begin{cases} \prod_{k=1}^m \mathrm{N}(\tilde{f}_{ik}, 1) & \text{if } y^*_{ij} > y^*_{ik}, \forall k \neq j \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

where $\tilde{f}_{ik} = \mathrm{E}[\alpha_k] + \sum_{l=1}^m h_{\mathrm{E}[\lambda_k]}(x_i, x_l) \, \mathrm{E}[w_{il}]$, and expectations are taken under the optimal mean-field distribution $\tilde{q}$. The distribution for $q(y^*_i)$ is a truncated m-variate normal distribution such that the jth component is always largest. It is worth investigating the properties of this distribution, and we now present some relevant definitions and results.

Definition 6.1 (Conically-truncated multivariate normal distribution)**.** Let $\mathbf{X} = (X_1, \ldots, X_d)$ be a d-dimensional random variable with pdf defined as

$$p(\mathbf{x}) = \begin{cases} \prod_{i=1}^d \mathrm{N}(\mu_i, \sigma_i) & \text{if } X_j > X_i, \forall i \neq j \\ 0 & \text{otherwise} \end{cases}$$

for some $j \in \{1, \ldots, d\}$. We denote the distribution of $\mathbf{X}$ by $\mathrm{N}^{(j)}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ and $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$. The pdf of $\mathbf{X}$ has support on the set $\{\mathbb{R}^d \,|\, x_j > x_i, \forall i \neq j\}$ and the following functional form:

$$p(\mathbf{x}) = \frac{C^{-1}}{\sigma_1 \cdots \sigma_d (2\pi)^{d/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right]$$

where ϕ is the pdf of a standard normal distribution and

$$C = \mathrm{E}_Z\left[\prod_{\substack{i=1\\ i\neq j}}^{d} \Phi\left(\frac{\sigma_j}{\sigma_i}Z + \frac{\mu_j - \mu_i}{\sigma_i}\right)\right]$$

where $Z \sim \mathrm{N}(0,1)$. In the case where all variances are unity, the pdf of $\mathbf{X} \sim \mathrm{N}^{(j)}(\boldsymbol{\mu}, \mathbf{I}_d)$ is

$$p(\mathbf{x}) = \left\{(2\pi)^{d/2}\,\mathrm{E}_Z\left[\prod_{\substack{i=1\\ i\neq j}}^{d} \Phi\left(Z + \mu_j - \mu_i\right)\right]\right\}^{-1} \exp\left[-\frac{1}{2}\sum_{i=1}^{d}(x_i - \mu_i)^2\right].$$

Proof. A derivation of the functional form for the pdf of $X \sim \mathrm{N}^{(j)}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given. Using the fact that $\int p(x)\,\mathrm{d}x = 1$, and that

$$
\begin{aligned}
&\int \mathbb{1}[x_i < x_j, \forall i \neq j] \prod_{i=1}^{d} \mathrm{N}(\mu_i, \sigma_i^2)\,\mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&\quad = \int \mathbb{1}[x_i < x_j, \forall i \neq j] \prod_{i=1}^{d} \left[\frac{1}{\sigma_i} \phi\left(\frac{x_i - \mu_i}{\sigma} \right) \right] \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&\quad = \int \mathbb{1}[x_i < x_j, \forall i \neq j] \frac{1}{\sigma_j} \phi\left(\frac{x_j - \mu_j}{\sigma_j} \right) \prod_{\substack{i=1 \\ i \neq j}}^{d} \left[\frac{1}{\sigma_i} \phi\left(\frac{x_i - \mu_i}{\sigma_i} \right) \right] \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&\quad = \int \prod_{\substack{i=1 \\ i \neq j}}^{d} \Phi\left(\frac{x_j - \mu_i}{\sigma_i} \right) \frac{1}{\sigma_j} \phi\left(\frac{x_j - \mu_j}{\sigma_j} \right) \mathrm{d}x_j \\
&\quad = \int \prod_{\substack{i=1 \\ i \neq j}}^{d} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_i}{\sigma_i} \right) \phi(z_j)\,\mathrm{d}z_j \\
&\qquad \text{(by using the standardisation } z_j = (x_j - \mu_j)/\sigma_j\text{)} \\
&\quad = \mathrm{E}\left[\prod_{\substack{i=1 \\ i \neq j}}^{d} \Phi\left(\frac{\sigma_j}{\sigma_i} Z_j + \frac{\mu_j - \mu_i}{\sigma_i} \right) \right]
\end{aligned}
$$

the proof follows directly. □

Lemma 6.1. *Let $X \sim \mathrm{N}^{(j)}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with pdf $p(\mathbf{x})$ as defined in Definition 6.1. Then*

(i) The expectation $\mathrm{E}[\mathbf{X}] = \big(\mathrm{E}[X_1], \dots, \mathrm{E}[X_d]\big)$ is given by

$$
\mathrm{E}[X_i] = \begin{cases} \mu_i - \sigma_i C^{-1} \, \mathrm{E}_Z \left[\phi_i \prod_{k \neq i,j} \Phi_k \right] & \text{if } i \neq j \\ \mu_j - \sigma_j \sum_{i \neq j} \big(\mathrm{E}[X_i] - \mu_i \big) & \text{if } i = j \end{cases}
$$

(ii) The differential entropy $\mathcal{H}(p)$ is given by

$$
\mathcal{H}(p) = \log C + \frac{d}{2} \log 2\pi + \frac{1}{2} \sum_{i=1}^{d} \log \sigma_i^2 + \frac{1}{2} \sum_{i=1}^{d} \frac{1}{\sigma_i^2} \, \mathrm{E}[x_i - \mu_i]^2
$$

where $C = \mathrm{E}\left[\prod_{i \neq j} \Phi_i\right]$*, and we had defined*

$$\phi_i = \phi_i(Z) = \phi\left(\frac{\sigma_j Z + \mu_j - \mu_i}{\sigma_i}\right)$$
$$\Phi_i = \Phi_i(Z) = \Phi\left(\frac{\sigma_j Z + \mu_j - \mu_i}{\sigma_i}\right)$$

with $Z \sim \mathrm{N}(0, 1)$*, and* $\phi(\cdot)$ *and* $\Phi(\cdot)$ *the pdf and cdf of* Z *respectively.*

As we know, y_i takes on any one value from the set $\{1, \ldots, m\}$. Thus, we have that the distribution of $(y^*_{i1}, \ldots, y^*_{im})$ is $\mathrm{N}^{(y_i)}(\boldsymbol{\mu}_i, \mathbf{I}_m)$, where $\boldsymbol{\mu}_i = (\tilde{f}_{i1}, \ldots, \tilde{f}_{im})$. The expectation is given by

$$\mathrm{E}[y^*_{ik}] = \begin{cases} \tilde{f}_{ik} - C_i^{-1} \mathrm{E}_Z\left[\phi_{ik}(Z) \prod\limits_{l \neq k, y_i} \Phi_{il}(Z)\right] & \text{if } k \neq y_i \\ \tilde{f}_{iy_i} - \sum_{k \neq y_i}\left(\mathrm{E}[y^*_{ik}] - \tilde{f}_{ik}\right) & \text{if } k = y_i \end{cases}$$

where

$$\phi_{ik}(Z) = \phi(Z + \tilde{f}_{iy_i} - \tilde{f}_{ik})$$
$$\Phi_{ik}(Z) = \Phi(Z + \tilde{f}_{iy_i} - \tilde{f}_{ik})$$
$$C_i = \mathrm{E}_Z\left[\prod_{\substack{i=1 \\ i \neq j}}^{d} \Phi\left(Z + \tilde{f}_{iy_i} - \tilde{f}_{ik}\right)\right]$$

and $Z \sim \mathrm{N}(0, 1)$ with PDF and CDF $\phi(\cdot)$ and $\Phi(\cdot)$ respectively. In order to calculate these expectations, we need to compute the following integrals:

$$\mathrm{E}_Z\left[\phi_{ik}(Z) \prod_{l \neq k, j} \Phi_{il}(Z)\right] = \int \phi_{ik}(z) \prod_{l \neq k, j} \Phi_{il}(z)\phi(z)\,\mathrm{d}z, \quad \forall k \neq y_i$$
$$C_i = \mathrm{E}_Z\left[\prod_{l \neq j} \Phi_{il}(Z)\right] = \int \prod_{l \neq j} \Phi_{il}(z)\phi(z)\,\mathrm{d}z$$

Since these are functions of a Gaussian pdf, these can be computed rather efficiently using quadrature methods.

$\tilde{q}(\mathbf{w})$

For each $j = 1, \dots, m$, denote $\mathbf{y}_j^* = (y_{1j}^*, \dots, y_{nj}^*)^\top$ as the vector of length n containing all latent observations for each class. Then,

$$\begin{aligned}
\log \tilde{q}(\mathbf{w}) &= \mathrm{E}_{\mathbf{y}^*,\alpha,\lambda}\left[-\frac{1}{2}\sum_{j=1}^m \|\mathbf{y}_j^* - \alpha_j \mathbf{1}_n - \mathbf{H}_{\lambda_j}\mathbf{w}_j\|^2 - \frac{1}{2}\sum_{j=1}^m \|\mathbf{w}_j\|^2\right] + \text{const.} \\
&= -\frac{1}{2}\sum_{j=1}^m \mathrm{E}_{\mathbf{y}^*,\alpha,\lambda}\left[\mathbf{w}_j^\top \mathbf{H}_{\lambda_j}^2 \mathbf{w}_j + \mathbf{w}_j^\top \mathbf{w}_j - 2(\mathbf{y}_j^* - \alpha_j \mathbf{1}_n)^\top \mathbf{H}_{\lambda_j}\mathbf{w}_j\right] + \text{const.} \\
&= -\frac{1}{2}\sum_{j=1}^m \left(\mathbf{w}_j^\top (\mathrm{E}[\mathbf{H}_{\lambda_j}^2] + \mathbf{I}_n)\mathbf{w}_j - 2(\mathrm{E}[\mathbf{y}_j^*] - \mathrm{E}[\alpha_j]\mathbf{1}_n)^\top \mathrm{E}[\mathbf{H}_{\lambda_j}]\mathbf{w}_j\right) + \text{const.}
\end{aligned}$$

Let $\mathbf{A}_j = \mathrm{E}[\mathbf{H}_{\lambda_j}^2] + \mathbf{I}_n$ and $\mathbf{a}_j = \mathrm{E}[\mathbf{H}_{\lambda_j}](\mathrm{E}[\mathbf{y}_j^*] - \mathrm{E}[\alpha_j]\mathbf{1}_n)$. Then, using the fact that

$$\mathbf{w}_j^\top \mathbf{A}_j \mathbf{w}_j - 2\mathbf{a}_j^\top \mathbf{w}_j = (\mathbf{w}_j - \mathbf{A}_j^{-1}\mathbf{a}_j)^\top \mathbf{A}_j (\mathbf{w}_j - \mathbf{A}_j^{-1}\mathbf{a}_j),$$

we see the $\log \tilde{q}(\mathbf{w})$ is a sum of quadratic terms in $\mathbf{w}_j$, and we recognise this as the kernel of the product of indepdendent multivariate normal densities. Therefore, for each $j = 1, \dots, m$,

$$\tilde{q}(\mathbf{w}_j) \equiv \mathrm{N}(\mathbf{A}_j^{-1}\mathbf{a}_j, \mathbf{A}_j^{-1}),$$

and $\tilde{q}(\mathbf{w}) = \prod_{j=1}^m \tilde{q}(\mathbf{w}_j)$. Because of this induced factorisation, we can obtain mean-field densities for each $\mathbf{w}_j$ separately. For convenience later in deriving the lower bound, we note that the second moment of $\tilde{q}(\mathbf{w}_j)$ is equal to $\mathrm{E}[\mathbf{w}_j \mathbf{w}_j^\top] = \mathbf{A}_j^{-1}(\mathbf{I}_n + \mathbf{a}_j \mathbf{a}_j^\top \mathbf{A}_j^{-1}) =: \widetilde{\mathbf{W}}_j$.

$\tilde{q}(\lambda)$

For $j = 1, \dots, m$,

$$\begin{aligned}
\log \tilde{q}(\lambda_j) &= \mathrm{E}_{\mathbf{y}^*,\mathbf{w},\alpha}\left[-\frac{1}{2}\sum_{j=1}^m \|\mathbf{y}_j^* - \alpha_j \mathbf{1}_n - \lambda_j \mathbf{H}\mathbf{w}_j\|^2\right] + \text{const.} \\
&= -\frac{1}{2}\sum_{j=1}^m \mathrm{E}_{\mathbf{y}^*,\mathbf{w},\alpha}\left[\lambda_j^2 \mathbf{w}_j^\top \mathbf{H}^2 \mathbf{w}_j - 2\lambda_j (\mathbf{y}_j^* - \alpha_j \mathbf{1}_n)^\top \mathbf{H}\mathbf{w}_j\right] + \text{const.} \\
&= -\frac{1}{2}\sum_{j=1}^m \left(\lambda_j^2 \operatorname{tr}\left(\mathbf{H}^2 \mathrm{E}[\mathbf{w}_j \mathbf{w}_j^\top]\right) - 2\lambda_j (\mathrm{E}[\mathbf{y}_j^*] - \mathrm{E}[\alpha_j]\mathbf{1}_n)^\top \mathbf{H}\,\mathrm{E}[\mathbf{w}_j]\right) + \text{const.}
\end{aligned}$$

By completing the squares, we recognise this is as the kernel of the product of independent univariate normal densities. Thus, each $\lambda_j \sim \mathrm{N}(d_j/c_j, 1/c_j)$, where

$$c_j = \operatorname{tr}\left(\mathbf{H}^2\,\mathrm{E}[\mathbf{w}_j\mathbf{w}_j^\top]\right) \quad \text{and} \quad d_j = (\mathrm{E}[\mathbf{y}_j^*] - \mathrm{E}[\alpha_j]\mathbf{1}_n)^\top \mathbf{H}\,\mathrm{E}[\mathbf{w}_j].$$

Supposing we use the same covariance kernel (and therefore scale parameter) for each regression class, the distribution for λ is easily seen as

$$\lambda \sim \mathrm{N}\left(\frac{\sum_{j=1}^m d_j}{\sum_{j=1}^m c_j}, \frac{1}{\sum_{j=1}^m c_j}\right).$$

$\tilde{q}(\alpha)$

For $j = 1, \dots, m$, denote $\mathbf{H}_i$ as the row vector of the kernel matrix $\mathbf{H}$. Then,

$$\begin{aligned}
\log \tilde{q}(\alpha) &= \mathrm{E}_{\mathbf{y}^*,\mathbf{w},\lambda}\left[-\frac{1}{2}\sum_{j=1}^m\sum_{i=1}^n \left(y_{ij}^* - \alpha_j - \lambda_j \textstyle\sum_{k=1}^n h(x_i,x_k)w_{kj}\right)^2\right] + \text{const.} \\
&= -\frac{1}{2}\sum_{j=1}^m \mathrm{E}_{\mathbf{y}^*,\mathbf{w},\lambda}\left[n\alpha_j^2 - 2\alpha_j\sum_{i=1}^n (y_{ij}^* - \lambda_j\mathbf{H}_i\mathbf{w}_j)\right] + \text{const.} \\
&= -\frac{n}{2}\sum_{j=1}^m \left[\left(\alpha_j - \frac{1}{n}\sum_{i=1}^n (\mathrm{E}[y_{ij}^*] - \mathrm{E}[\lambda_j]\mathbf{H}_i\mathbf{w}_j)\right)^2\right] + \text{const.}
\end{aligned}$$

which is of course the kernel of the product of m univariate normal densities, each with mean and variance

$$\tilde{\alpha}_j = \frac{1}{n}\sum_{i=1}^n \left(\mathrm{E}[y_{ij}^*] - \mathrm{E}[\lambda_j]\mathbf{H}_i\,\mathrm{E}[\mathbf{w}_j]\right) \quad \text{and} \quad v_{\alpha_j} = \frac{1}{n}.$$

Suppose that we use a single intercept parameter α. In this case, α is is also normally distributed with mean and variance

$$\tilde{\alpha} = \frac{1}{nm}\sum_{j=1}^m\sum_{i=1}^n \left(\mathrm{E}[y_{ij}^*] - \mathrm{E}[\lambda_j]\mathbf{H}_i\,\mathrm{E}[\mathbf{w}_j]\right) \quad \text{and} \quad v_\alpha = \frac{1}{nm}.$$

6.3 Monitoring the lower bound

A convergence criterion would be when there is no more significant increase in the lower bound $\mathcal{L}$, as defined by

$$\begin{aligned}
\mathcal{L} &= \int q(\mathbf{y}^*, \mathbf{w}, \lambda, \alpha) \log \left[\frac{p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \lambda, \alpha)}{q(\mathbf{y}^*, \mathbf{w}, \lambda, \alpha)} \right] \mathrm{d}\mathbf{y}^* \, \mathrm{d}\mathbf{w} \, \mathrm{d}\lambda \, \mathrm{d}\alpha \\
&= \mathrm{E}[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \lambda, \alpha)] - \mathrm{E}[\log q(\mathbf{y}^*, \mathbf{w}, \lambda, \alpha)] \\
&= \cancel{\mathrm{E} \left[\log \prod_{i=1}^n \prod_{j=1}^m p(y_i | y_{ij}^*) \right]} + \mathrm{E}\left[\log p(\mathbf{y}^* | \mathbf{f})\right] + \mathrm{E}\left[\log p(\mathbf{w})\right] + \cancel{\mathrm{E}\left[\log p(\lambda)\right]} + \cancel{\mathrm{E}\left[\log p(\alpha)\right]} \\
&\quad - \mathrm{E}\left[\log q(\mathbf{y}^*)\right] - \mathrm{E}\left[\log q(\mathbf{w})\right] - \mathrm{E}\left[\log q(\lambda)\right] - \mathrm{E}\left[\log q(\alpha)\right]
\end{aligned}$$

Note that the categorical pmf $p(y_i|y_{ij}^*)$ becomes degenerate once the latent variables are known, so this term is cancelled out. With the exception of $q(\mathbf{y}^*)$, all of the distributions are Gaussian. The following results will be helpful.

Definition 6.2 (Differential entropy)**.** The differential entropy $\mathcal{H}$ of a pdf $p(x)$ is given by

$$\mathcal{H}(p) = -\int p(x) \log p(x) \, \mathrm{d}x = -\mathrm{E}_p[\log p(x)].$$

Lemma 6.2. *Let $p(x)$ be the pdf of a random variable x. Then if*

(i) p is a univariate normal distribution with mean μ and variance σ^2,

$$\mathcal{H}(p) = \frac{1}{2}(1 + \log 2\pi) + \frac{1}{2} \log \sigma^2$$

(ii) p is a d-dimensional normal distribution with mean μ and variance Σ,

$$\mathcal{H}(p) = \frac{d}{2}(1 + \log 2\pi) + \frac{1}{2} \log |\Sigma|$$

Terms involving distributions of $\mathbf{y}^*$

$$
\begin{aligned}
\mathrm{E}\left[\log p(\mathbf{y}^*|\mathbf{f})\right] - \mathrm{E}\left[\log q(\mathbf{y}^*)\right] &= \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{E}\left[\log p(y^*_{ij}|f_{ij})\right] + \sum_{i=1}^{n} \mathcal{H}\big(q(y^*_i)\big) \\
&= \cancel{\sum_{i=1}^{n}\sum_{j=1}^{m} \left(-\frac{1}{2}\log 2\pi - \frac{1}{2}\,\mathrm{E}[y^*_{ij} - f_{ij}]^2\right)} \\
&\quad + \cancel{\sum_{i=1}^{n}\sum_{j=1}^{m} \left(\frac{1}{2}\log 2\pi + \frac{1}{2}\,\mathrm{E}[y^*_{ij} - f_{ij}]^2\right)} + \sum_{i=1}^{n} \log C_i
\end{aligned}
$$

Terms involving distributions of $\mathbf{w}$

$$
\begin{aligned}
\mathrm{E}\left[\log p(\mathbf{w})\right] - \mathrm{E}\left[\log q(\mathbf{w})\right] &= \sum_{j=1}^{m} \Big(\mathrm{E}\left[\log p(\mathbf{w}_j)\right] - \mathrm{E}\left[\log q(\mathbf{w}_j)\right] \Big) \\
&= \sum_{j=1}^{m} \left(-\frac{n}{2}\log 2\pi - \frac{1}{2}\,\mathrm{E}[\mathbf{w}_j^\top \mathbf{w}_j] + \mathcal{H}\big(q(\mathbf{w}_j)\big) \right) \\
&= \sum_{j=1}^{m} \left(\cancel{-\frac{n}{2}\log 2\pi} - \frac{1}{2}\operatorname{tr}\Big(\mathrm{E}[\mathbf{w}_j \mathbf{w}_j^\top]\Big) + \frac{n}{2}(1 + \cancel{\log 2\pi}) - \frac{1}{2}\log|\mathbf{A}_j| \right) \\
&= \frac{nm}{2} - \frac{1}{2}\sum_{j=1}^{m} \Big(\operatorname{tr}\widetilde{\mathbf{W}}_j + \log|\mathbf{A}_j| \Big)
\end{aligned}
$$

Terms involving distribution of $q(\lambda)$

$$
\begin{aligned}
-\,\mathrm{E}\left[\log q(\lambda)\right] &= \sum_{j=1}^{m} \mathcal{H}\big(q(\lambda_j)\big) \\
&= \sum_{j=1}^{m} \left(\frac{1}{2}(1 + \log 2\pi) - \frac{1}{2}\log c_j \right) \\
&= \frac{m}{2}(1 + \log 2\pi) - \frac{1}{2}\sum_{j=1}^{m} \log c_j
\end{aligned}
$$

or if using single λ

$$-\operatorname{E}\left[\log q(\lambda)\right] = \frac{1}{2}(1+\log 2\pi) - \frac{1}{2}\log\sum_{j=1}^{m} c_j.$$

Terms involving distribution of $q(\alpha)$

$$\begin{aligned}
-\operatorname{E}\left[\log q(\alpha)\right] &= \sum_{j=1}^{m} \mathcal{H}\big(q(\alpha_j)\big) \\
&= \frac{m}{2}(1+\log 2\pi - \log n)
\end{aligned}$$

or if using single α

$$-\operatorname{E}\left[\log q(\alpha)\right] = \frac{1}{2}(1+\log 2\pi - \log nm).$$

The lower bound

$$\begin{aligned}
\mathcal{L} &= \sum_{i=1}^{n} \log C_i + \frac{nm}{2} - \frac{1}{2}\sum_{j=1}^{m}\Big(\operatorname{tr}\widetilde{\mathbf{W}}_j + \log|\mathbf{A}_j|\Big) \\
&\quad + \frac{m}{2}(1+\log 2\pi) - \frac{1}{2}\sum_{j=1}^{m}\log c_j + \frac{m}{2}(1+\log 2\pi - \log n) \\
&= \frac{m}{2}\big(n + 2(1+\log 2\pi) - \log n\big) - \frac{1}{2}\sum_{j=1}^{m}\Big(\operatorname{tr}\widetilde{\mathbf{W}}_j + \log|\mathbf{A}_j| + \log c_j\Big) + \sum_{i=1}^{n}\log C_i
\end{aligned}$$

Of course, if using either single α or single λ, then the formula needs to be adjusted accordingly.

6.4 The variational algorithm

Since there is a cyclic dependence of the parameters on each other, we employ a sequential update algorithm. In what follows, a tilde on the parameters indicate that these are the expectations of the parameters given the optimal factorised distributions $\tilde{q}$ derived earlier.

1. Is this variational EM... or CAVI?

STEP 1: Update $\tilde{\mathbf{y}}^{*(t+1)}$ given $\tilde{\mathbf{w}}^{(t)}$, $\tilde{\lambda}^{(t)}$, and $\tilde{\alpha}^{(t)}$

STEP 2: Update $\tilde{\mathbf{w}}^{(t+1)}$ given $\tilde{\mathbf{y}}^{*(t+1)}$, $\tilde{\lambda}^{(t)}$, and $\tilde{\alpha}^{(t)}$

STEP 3: Update $\tilde{\lambda}^{(t+1)}$ given $\tilde{\mathbf{y}}^{*(t+1)}$, $\tilde{\mathbf{w}}^{(t+1)}$, and $\tilde{\alpha}^{(t)}$

STEP 4: Update $\tilde{\alpha}^{(t+1)}$ given $\tilde{\mathbf{y}}^{*(t+1)}$, $\tilde{\mathbf{w}}^{(t+1)}$, and $\tilde{\lambda}^{(t+1)}$

Algorithm 1 VB-EM algorithm for the probit I-prior model

procedure INITIALISE
  for $j = 1, \dots, m$ **do**
    $\tilde{\mathbf{w}}_j^{(0)} \leftarrow \mathbf{0}_n$
    $\tilde{\alpha}_j^{(0)} \leftarrow \mathrm{N}(0,1)$
    $\tilde{\lambda}_j^{(0)} \leftarrow \mathrm{N}(0,1)$
    $\tilde{\lambda}_j^{sq(0)} \leftarrow (\tilde{\lambda}_j^{(0)})^2$ ▷ this is $\mathrm{E}[\lambda_j^2]$
    $\mathbf{H}_{\lambda_j}^{(0)} \leftarrow \tilde{\lambda}_j^{(0)}\mathbf{H}$
    $\mathbf{H}_{\lambda_j}^{sq(0)} \leftarrow \tilde{\lambda}_j^{sq(0)}\mathbf{H}^2$
  end for
end procedure

procedure UPDATE FOR $\tilde{\mathbf{f}}$ (time t)
  for $j = 1, \dots, m$ **do**
    $\tilde{\mathbf{f}}_j^{(t+1)} \leftarrow \tilde{\alpha}_j^{(t)}\mathbf{1}_n + \mathbf{H}_{\lambda_j}\tilde{\mathbf{w}}_j^{(t)}$
  end for
  $\tilde{\mathbf{f}}^{(t+1)} \leftarrow \big(\tilde{\mathbf{f}}_1^{(t+1)}, \dots, \tilde{\mathbf{f}}_m^{(t+1)}\big)^\top$
end procedure

procedure UPDATE FOR y^*_{ij} (time t)
  for $i = 1, \dots, n$ **do**
    $j \leftarrow y_i$
    $C_i^{(t+1)} \leftarrow \prod_{k \neq j} \Phi\left((\tilde{f}_{ij}^{(t+1)} - \tilde{f}_{ik}^{(t+1)})/\sqrt{2}\right)$
    for $k = 1, \dots, j-1, j+1, \dots, m$ **do**
      $D_{ik} \leftarrow \mathrm{E}_Z\left[\phi_k(Z + \tilde{f}_{ij}^{(t+1)} - \tilde{f}_{ik}^{(t+1)}) \prod_{l \neq k,j} \Phi_l(Z + \tilde{f}_{ij}^{(t+1)} - \tilde{f}_{ik}^{(t+1)})\right]$
      $\tilde{y}_{ik}^{*(t+1)} \leftarrow \tilde{f}_{ik}^{(t+1)} - D_{ik}/C_i^{(t+1)}$
    end for
    $\tilde{y}_{ij}^{*(t+1)} \leftarrow \tilde{f}_{ij}^{(t+1)} - \sum_{k \neq j}\big(\tilde{y}_{ik}^{*(t+1)} - \tilde{f}_{ik}^{(t+1)}\big)$
  end for
end procedure

procedure UPDATE FOR $\mathbf{w}_j$ (time t)
    for $j = 1, \ldots, m$ **do**
        $\tilde{\mathbf{y}}_j^{*(t+1)} \leftarrow (\tilde{y}_{1j}^{(t+1)}, \ldots, \tilde{y}_{nj}^{(t+1)})^\top$
        $\mathbf{A}_j \leftarrow \mathbf{H}_{\lambda_j}^{sq(t)} + \mathbf{I}_n$
        $\mathbf{a}_j \leftarrow \mathbf{H}_\lambda(\tilde{\mathbf{y}}_j^{*(t+1)} - \tilde{\alpha}_j^{(t)}\mathbf{1}_n)$
        $\tilde{\mathbf{w}}_j^{(t+1)} \leftarrow \mathbf{A}_j^{-1}\mathbf{a}_j$
        $\widetilde{\mathbf{W}}_j^{(t+1)} \leftarrow \mathbf{A}_j^{-1}(\mathbf{I}_n + \mathbf{a}_j\mathbf{a}_j^\top\mathbf{A}_j^{-1})$
        $\text{logdetA}_j^{(t+1)} \leftarrow \log|\mathbf{A}_j|$
    end for
end procedure

procedure UPDATE FOR λ (time t)
    for $j = 1, \ldots, m$ **do**
        $c_j^{(t+1)} \leftarrow \operatorname{tr}\left(\mathbf{H}^2\widetilde{\mathbf{W}}_j\right)$
        $d_j \leftarrow (\tilde{\mathbf{y}}_j^{*(t+1)} - \tilde{\alpha}_j^{(t)}\mathbf{1}_n)^\top\mathbf{H}\tilde{\mathbf{w}}_j^{(t+1)}$
        $\tilde{\lambda}_j^{(t+1)} \leftarrow d_j / c_j^{(t+1)}$
        $\tilde{\lambda}_j^{sq(t+1)} \leftarrow 1/c_j^{(t)} + (d_i/c_i^{(t+1)})^2$
    end for
    if single λ **then** $\forall j$
        $\tilde{\lambda}_j^{(t+1)} \leftarrow \sum_j d_j \Big/ \sum_j c_j^{(t+1)}$
        $\tilde{\lambda}_j^{sq(t+1)} \leftarrow 1 \Big/ \sum_j c_j^{(t+1)} + \left(\sum_j d_j \Big/ \sum_j c_j^{(t+1)}\right)^2$
    end if
    call UPDATE KERNEL MATRICES
end procedure

procedure UPDATE KERNEL MATRICES (time t)
    for $j = 1, \ldots, m$ **do**
        $\mathbf{H}_{\lambda_j}^{(t+1)} \leftarrow \tilde{\lambda}_j^{(t+1)}\mathbf{H}$
        $\mathbf{H}_{\lambda_j}^{sq(t+1)} \leftarrow \tilde{\lambda}_j^{sq(t+1)}\mathbf{H}^2$
    end for
end procedure

procedure UPDATE FOR α (time t)
  if single α **then**
    $\tilde{\alpha}^{(t+1)} \leftarrow \frac{1}{nm} \sum_{j=1}^{m} \sum_{i=1}^{n} \left(\tilde{y}_{ij}^{*(t+1)} - \tilde{\lambda}_j^{(t+1)} \mathbf{H}_i \tilde{\mathbf{w}}_j^{(t+1)} \right)$
  else
    for $j = 1, \dots, m$ **do**
      $\tilde{\alpha}_j^{(t+1)} \leftarrow \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{y}_{ij}^{*(t+1)} - \tilde{\lambda}_j^{(t+1)} \mathbf{H}_i \tilde{\mathbf{w}}_j^{(t+1)} \right)$
    end for
  end if
end procedure

procedure CALCULATE LOWER BOUND (time t)
  $\mathcal{L}^{(t)} \leftarrow \frac{1}{2}\big(nm - \log nm + 3(1 + \log 2\pi)\big) - \frac{1}{2} \left(\operatorname{logdet} \mathrm{A}^{(t)} + \operatorname{tr} \widetilde{\mathbf{W}}^{(t)} + \sum_{i=1}^{2} \log c_i^{(t)} \right) + \sum_{i=1}^{n} \log C_i^{(t)}$
end procedure

procedure THE VB-EM ALGORITHM
  $t \leftarrow 0$
  while $\mathcal{L}^{(t+1)} - \mathcal{L}^{(t)} > \delta$ **or** $t < t_{max}$ **do**
    call UPDATE FOR $\mathbf{y}^*$
    call UPDATE FOR $\mathbf{w}$
    call UPDATE FOR λ
    call UPDATE FOR α
    call CALCULATE LOWER BOUND
    $t \leftarrow t + 1$
  end while
end procedure

return $(\hat{\mathbf{y}}^*, \hat{\mathbf{w}}, \hat{\lambda}, \hat{\alpha}) \leftarrow (\tilde{\mathbf{y}}^{*(t)}, \tilde{\mathbf{w}}^{(t)}, \tilde{\lambda}^{(t)}, \tilde{\alpha}^{(t)})$ ▷ converged parameter estimates
return $(\hat{y}_1, \dots, \hat{y}_n) \leftarrow \left(\arg\max_{k=1}^{m} \hat{y}_{1k}^*, \dots, \arg\max_{k=1}^{m} \hat{y}_{nk}^* \right)$ ▷ predicted classes
for $i = 1, \dots, n$ **do**
  for $j = 1, \dots, m$ **do**
    return $\hat{p}_{ij} \leftarrow \prod_{\substack{k=1 \\ k \neq j}}^{m} \Phi \left(\frac{\hat{y}_{ij}^* - \hat{y}_{ik}^*}{\sqrt{2}} \right)$ ▷ predicted probabilities
  end for
end for

Appendix A

Appendix for I-probit

A.1 Proof of Lemma 6.1

Proof. (i) Due to the independence structure in the pdf of $\mathbf{X}$, it is easy to consider the expectations of each of the components separately and marginalising out the rest of the components. For $i \neq j$, we have

$$
\begin{aligned}
\mathrm{E}[x_i] &= C^{-1} \int \cdots \int \mathbb{1}[x_k < x_j, \forall k \neq j] \cdot x_i \prod_{k=1}^{d} \frac{1}{\sigma_k} \phi\left(\frac{x_k - \mu_k}{\sigma_k}\right) \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&= C^{-1} \iint \mathbb{1}[x_i < x_j] \frac{x_i}{\sigma_i} \phi\left(\frac{x_i - \mu_i}{\sigma_i}\right) \prod_{k \neq i,j} \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) \frac{1}{\sigma_j} \phi\left(\frac{x_j - \mu_j}{\sigma_j}\right) \mathrm{d}x_i \, \mathrm{d}x_j \\
&= C^{-1} \iint \mathbb{1}[\sigma_i z_i + \mu_i < \sigma_j z_j + \mu_j](\sigma_i z_i + \mu_i)\phi(z_i) \prod_{k \neq i,j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \phi(z_j) \, \mathrm{d}z_i \, \mathrm{d}z_j \\
&= \mu_i C^{-1} \iint \mathbb{1}[z_i < (\sigma_j z_j + \mu_j - \mu_i)/\sigma_i]\phi(z_i) \prod_{k \neq i,j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \phi(z_j) \, \mathrm{d}z_i \, \mathrm{d}z_j \\
&\quad + \sigma_i C^{-1} \iint \mathbb{1}[z_i < (\sigma_j z_j + \mu_j - \mu_i)/\sigma_i] z_i \phi(z_i) \prod_{k \neq i,j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \phi(z_j) \, \mathrm{d}z_i \, \mathrm{d}z_j \\
&= \mu_i C^{-1} \overbrace{\int \prod_{k \neq j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \phi(z_j) \, \mathrm{d}z_j}^{C} \\
&\quad + \sigma_i C^{-1} \int \mathbb{1}[z_i < (\sigma_j z_j + \mu_j - \mu_i)/\sigma_i] z_i \phi(z_i) \prod_{k \neq i,j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \phi(z_j) \, \mathrm{d}z_i \, \mathrm{d}z_j
\end{aligned}
$$

The integral involving z_i in the second part of the sum is recognised as the (unnormalised) expectation of the lower-tail of a univariate standard normal distribution truncated at $\tau_{ij} = (\sigma_j z_j + \mu_j - \mu_i)/\sigma_i$. That is,

$$\mathrm{E}[Z_i | Z_i < \tau_{ij}] = \left[\Phi(\tau_{ij})\right]^{-1} \int \mathbb{1}[z_i < \tau_{ij}] z_i \phi(z_i) \,\mathrm{d}z_i = -\frac{\phi(\tau_{ij})}{\Phi(\tau_{ij})}$$

Plugging this expression back into the derivation of this expectation, we get

$$\begin{aligned}
\mathrm{E}[X_i] &= \mu_i - \sigma_i C^{-1} \int \phi\left(\frac{\sigma_j z_j + \mu_j - \mu_i}{\sigma_i}\right) \prod_{k \neq i,j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \phi(z_j) \,\mathrm{d}z_j \\
&= \mu_i - \sigma_i C^{-1}\, \mathrm{E}\left[\phi\left(\frac{\sigma_j Z_j + \mu_j - \mu_i}{\sigma_i}\right) \prod_{k \neq i,j} \Phi\left(\frac{\sigma_j Z_j + \mu_j - \mu_k}{\sigma_k}\right)\right].
\end{aligned}$$

The expectation for the jth component is

$$\begin{aligned}
\mathrm{E}[X_j] &= C^{-1} \int \cdots \int \mathbb{1}[x_k < x_j, \forall k \neq j] \cdot x_j \prod_{k=1}^{d} \frac{1}{\sigma_k} \phi\left(\frac{x_k - \mu_k}{\sigma_k}\right) \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&= C^{-1} \int x_j \prod_{k \neq j} \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) \cdot \frac{1}{\sigma_j} \phi\left(\frac{x_j - \mu_j}{\sigma_j}\right) \mathrm{d}x_j \\
&= C^{-1} \int (\sigma_j z_j + \mu_j) \prod_{k \neq j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \cdot \phi(z_j) \,\mathrm{d}z_j \\
&= \mu_j C^{-1} \overbrace{\int \prod_{k \neq j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \cdot \phi(z_j) \,\mathrm{d}z_j}^{C} \\
&\qquad + \sigma_j C^{-1} \int \prod_{k \neq j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \cdot z_j \phi(z_j) \,\mathrm{d}z_j \\
&= \mu_j + \sigma_j C^{-1}\, \mathrm{E}\left[Z_j \prod_{k \neq j} \Phi\left(\frac{\sigma_j Z_j + \mu_j - \mu_k}{\sigma_k}\right)\right] \\
&= \mu_j + \sigma_j \sum_{\substack{i=1 \\ i \neq j}}^{d} \sigma_i C^{-1}\, \mathrm{E}\left[\phi\left(\frac{\sigma_j Z_j + \mu_j - \mu_i}{\sigma_i}\right) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi\left(\frac{\sigma_j Z_j + \mu_j - \mu_k}{\sigma_k}\right)\right] \\
&= \mu_j - \sigma_j \sum_{i \neq j} \big(\mathrm{E}[X_i] - \mu_i\big)
\end{aligned}$$

where we have made use of Lemma A.1 in the second last step of the above.

(ii) The differential entropy is given by

$$
\begin{aligned}
\mathcal{H}(p) &= -\int p(\mathbf{x}) \log p(\mathbf{x}) \,\mathrm{d}\mathbf{x} = -\operatorname{E}[\log p(\mathbf{x})] \\
&= -\operatorname{E}\left[-\log C - \frac{d}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{d}\log\sigma_i^2 - \frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right] \\
&= \log C + \frac{d}{2}\log 2\pi + \frac{1}{2}\sum_{i=1}^{d}\log\sigma_i^2 + \frac{1}{2}\sum_{i=1}^{d}\frac{1}{\sigma_i^2}\operatorname{E}[x_i-\mu_i]^2.
\end{aligned}
$$

□

Lemma A.1. *Let $Z \sim \mathrm{N}(0,1)$. Then for all $m \in \{\mathbb{N} \,|\, m > 1\}$ and $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$,*

$$
\operatorname{E}\left[Z \prod_{\substack{k=1\\k\neq j}}^{m} \Phi(\sigma_k Z + \mu_k)\right] = \sum_{\substack{i=1\\i\neq j}}^{m} \operatorname{E}\left[\sigma_i \phi(\sigma_i Z + \mu_i) \prod_{\substack{k=1\\k\neq i,j}}^{m} \Phi(\sigma_k Z + \mu_k)\right]
$$

for some $j \in \{1, \ldots, m\}$.

Proof. Use the fact that for any differentiable function g, $\operatorname{E}[Zg(Z)] = \operatorname{E}[g'(Z)]$, and apply the result with the function $g_m : z \mapsto \prod_{k\neq j} \Phi(\sigma_k z + \mu_k)$. All that is left is to derive the derivative of g, and we use an inductive proof to do this.

We adopt the following notation for convenience:

$$
\begin{aligned}
\phi_i &= \phi(\sigma_i z + \mu_i) \\
\Phi_i &= \Phi(\sigma_i z + \mu_i)
\end{aligned}
$$

The simplest case is when $m = 2$, which can be trivially shown to be true. Without loss of generality, let $j = 1$. Then

$$
\begin{aligned}
g_2(z) &= \Phi_2 \\
\Rightarrow g_2'(z) &= \sigma_2\phi_2 = \sum_{\substack{i=1\\i\neq 1}}^{2}\left[\sigma_i\phi_i \sum_{\substack{k=1\\k\neq 1,2}}^{2} \Phi_k\right].
\end{aligned}
$$

Now assume that the inductive hypothesis holds for some $m \in \{\mathbb{N} \,|\, m > 1\}$. That is, the derivative of

$$
g_m(z) = \prod_{\substack{k=1\\k\neq j}}^{m} \Phi_k
$$

which is

$$g'_m(z) = \sum_{\substack{i=1 \\ i \neq j}}^{m} \left[\sigma_i \phi_i \prod_{\substack{k=1 \\ k \neq i,j}}^{m} \Phi_k \right],$$

is assumed to be true. Assume that without loss of generality, $j \neq m+1$. Then the derivative of

$$g_{m+1}(z) = \prod_{\substack{k=1 \\ k \neq j}}^{m+1} \Phi_k = g_m(z)\Phi_{m+1}$$

is found to be

$$\begin{aligned}
g'_{m+1}(z) &= \sigma_{m+1}\phi_{m+1} g_m(z) + g'_m(z)\Phi_{m+1} \\
&= \sigma_{m+1}\phi_{m+1} \prod_{\substack{k=1 \\ k \neq j}}^{m} \Phi_k + \sum_{\substack{i=1 \\ i \neq j}}^{m} \left[\sigma_i \phi_i \prod_{\substack{k=1 \\ k \neq i,j}}^{m} \Phi_k \right] \Phi_{m+1} \\
&= \sigma_{m+1}\phi_{m+1} \prod_{\substack{k=1 \\ k \neq j,m+1}}^{m+1} \Phi_k + \sum_{\substack{i=1 \\ i \neq j}}^{m} \left[\sigma_i \phi_i \prod_{\substack{k=1 \\ k \neq i,j}}^{m+1} \Phi_k \right] \\
&= \sum_{\substack{i=1 \\ i \neq j}}^{m+1} \left[\sigma_i \phi_i \prod_{\substack{k=1 \\ k \neq i,j}}^{m+1} \Phi_k \right] \\
&= g'_{m+1}(z).
\end{aligned}$$

Thus, by induction and linearity of expectations, the proof is complete. □

A.2 Proof for ...

Lemma A.2. *Let $p(x)$ be the pdf of a random variable x. Then if*

(i) p is a univariate normal distribution with mean μ and variance σ^2,

$$\mathcal{H}(p) = \frac{1}{2}(1 + \log 2\pi) + \frac{1}{2}\log \sigma^2$$

(ii) p is a d-dimensional normal distribution with mean μ and variance Σ,

$$\mathcal{H}(p) = \frac{d}{2}(1 + \log 2\pi) + \frac{1}{2}\log |\Sigma|$$

(iii) p is distribution of the ***upper-tail*** *of a univariate, one-sided normal distribution*

truncated at zero with mean μ and variance 1,

$$\mathcal{H}(p) = \frac{1}{2}\log 2\pi + \frac{1}{2}\left(\mathrm{E}[x^2] + \mu^2 - 2\mu\,\mathrm{E}[x]\right) + \log\Phi(\mu)$$

*(iv) p is distribution of the **lower-tail** of a univariate, one-sided normal distribution truncated at zero with mean μ and variance 1,*

$$\mathcal{H}(p) = \frac{1}{2}\log 2\pi + \frac{1}{2}\left(\mathrm{E}\,x^2 + \mu^2 - 2\mu\,\mathrm{E}\,x\right) + \log\left(1 - \Phi(\mu)\right)$$

Proof.

Case (i): $-\log p(x) = \frac{1}{2}\log 2\pi + \frac{1}{2}\log\sigma^2 + \frac{1}{2}(x-\mu)^2$. Then

$$\begin{aligned}
\mathcal{H}(p) &= \mathrm{E}_x\left[\frac{1}{2}\log 2\pi + \frac{1}{2}\log\sigma^2 + \frac{1}{2\sigma^2}(x-\mu)^2\right] \\
&= \frac{1}{2}\log 2\pi + \frac{1}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\cancelto{\sigma^2}{\mathrm{E}(x-\mu)^2} \\
&= \frac{1}{2}(1 + \log 2\pi) + \frac{1}{2}\log\sigma^2
\end{aligned}$$

Case (ii): $-\log p(x) = \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\Sigma| + \frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)$. Then

$$\begin{aligned}
\mathcal{H}(p) &= \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\Sigma| + \frac{1}{2}\mathrm{E}_x\left[(x-\mu)^\top\Sigma^{-1}(x-\mu)\right] \\
&= \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\Sigma| + \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}\cancelto{\Sigma}{\mathrm{E}_x\left[(x-\mu)(x-\mu)^\top\right]}\right) \\
&= \frac{d}{2}(1 + \log 2\pi) + \frac{1}{2}\log|\Sigma|
\end{aligned}$$

For the next two cases, we state the following properties of a truncated normal distribution without proof.

Lemma A.3. *Let $x \sim \mathrm{N}(\mu, \sigma^2)$ with x lying in the interval (a, b). Then we say that x follows a truncated normal distribution, and*

(i) the mean of x (conditional on $a < x < b$) is

$$\mathrm{E}[x] = \mu + \sigma\frac{\phi(\alpha) - \phi(\beta)}{Z},$$

(ii) the variance of x (conditional on $a < x < b$) is

$$\mathrm{Var}[x] = \sigma^2 \left[1 + \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{Z} - \left(\frac{\phi(\alpha) - \phi(\beta)}{Z}\right)^2\right], \textit{ and}$$

(iii) the entropy of the pdf of x (conditional on $a < x < b$) is

$$\mathcal{H} = \frac{1}{2}\log 2\pi + \frac{1}{2}\log\sigma^2 + \log Z + \frac{1}{2} + \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{2Z},$$

where $\alpha = (a-\mu)/\sigma$, $\beta = (b-\mu)/\sigma$, and $Z = \Phi(\beta) - \Phi(\alpha)$, and ϕ and Φ are the pdf and cdf of a standard normal distribution respectively.

In the special case when $\sigma = 1$ (the case we are interested in), then with some manipulation, one arrives at the following expression for the entropy of the pdf p of a truncated normal distribution:

$$\begin{aligned}
\mathcal{H}(p) &= \frac{1}{2}\log 2\pi + \cancel{\frac{1}{2}\log\sigma^2} + \log Z + \frac{1}{2}\left(1 + \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{Z}\right) \\
&= \frac{1}{2}\log 2\pi + \log Z + \frac{1}{2}\left(\mathrm{Var}[x] + \left(\frac{\phi(\alpha) - \phi(\beta)}{Z}\right)^2\right) \\
&= \frac{1}{2}\log 2\pi + \log Z + \frac{1}{2}\left(\mathrm{E}[x^2] - \mathrm{E}^2[x] + (\mathrm{E}[x] - \mu)^2\right) \\
&= \frac{1}{2}\log 2\pi + \log Z + \frac{1}{2}\left(\mathrm{E}[x^2] + \mu^2 - 2\mu\,\mathrm{E}[x]\right)
\end{aligned}$$

We now continue with the proof.

Case (iii): Using Lemma A.3 with $a = 0$, $b = +\infty$, and $\sigma = 1$, we get that $Z = 1 - \Phi(-\mu) = \Phi(\mu)$. Therefore, the entropy of p is given by

$$\mathcal{H}(p) = \frac{1}{2}\log 2\pi + \frac{1}{2}\left(\mathrm{E}[x^2] + \mu^2 - 2\mu\,\mathrm{E}[x]\right) + \log\Phi(\mu)$$

Case (iv): Again, using Lemma A.3 with $a = -\infty$, $b = 0$, and $\sigma = 1$, we get that $Z = \Phi(-\mu) = 1 - \Phi(\mu)$. Therefore, the entropy of p is given by

$$\mathcal{H}(p) = \frac{1}{2}\log 2\pi + \frac{1}{2}\left(\mathrm{E}[x^2] + \mu^2 - 2\mu\,\mathrm{E}[x]\right) + \log\left(1 - \Phi(\mu)\right)$$

□

A.3 Distribution of $\tilde{q}(\mathbf{y}^*)$ for binary case

Case: $y_i = 1$

$$\begin{aligned}
\log \tilde{q}(y_i^*) &= \mathbb{1}[y_i^* \geq 0] \cdot \operatorname{E}_{\mathbf{w},\alpha,\lambda}\left[-\frac{1}{2}(y_i^* - \alpha - \lambda \mathbf{H}_i \mathbf{w})^2\right] + \text{const.} \\
&= \mathbb{1}[y_i^* \geq 0] \cdot \left[-\frac{1}{2}\Big(y_i^{*2} - 2\operatorname{E}_{\mathbf{w},\alpha,\lambda}[\alpha + \lambda \mathbf{H}_i \mathbf{w}] y_i\Big)\right] + \text{const.} \\
&= \mathbb{1}[y_i^* \geq 0] \left[-\frac{1}{2}(y_i^* - \tilde{\eta}_i)^2\right] + \text{const.} \\
&\equiv \begin{cases} \mathrm{N}(\tilde{\eta}_i, 1) & \text{if } y_i^* \geq 0 \\ 0 & \text{if } y_i^* < 0 \end{cases}
\end{aligned}$$

where

$$\tilde{\eta}_i = \operatorname{E}\alpha + \operatorname{E}\lambda \mathbf{H}_i \operatorname{E}\mathbf{w}$$

by independence of $q(\mathbf{w})$, $q(\alpha)$ and $q(\lambda)$. $\tilde{q}(y_i^*)$ is recognised as being the upper-tail of a one-sided normal distribution truncated at zero. The mean is

$$\operatorname{E}[y_i^* | y_i^* \geq 0] = \tilde{\eta}_i + \frac{\phi(\tilde{\eta}_i)}{\Phi(\tilde{\eta}_i)}$$

where ϕ and Φ are, respectively, the pdf and cdf of a standard normal distribution.

Case: $y_i = 0$

Following the same argument, we can deduce that $q(y_i^*)$ in this case would be the lower-tail of a one-sided normal distribution truncated at zero. The mean is

$$\operatorname{E}[y_i^* | y_i^* < 0] = \tilde{\eta}_i + \frac{\phi(\tilde{\eta}_i)}{\Phi(\tilde{\eta}_i) - 1}.$$

List of Figures

List of Tables

List of Theorems

List of Definitions

List of Symbols

$\mathrm{N}_p(\mu, \Sigma)$	p-dimensional multivariate normal distribution with mean vector μ and covariance Σ.
$\sim$	Is distributed as.
$\otimes$	The tensor product.