Haziq Jamil
Department of Statistics
London School of Economics and Political Science

1 Binary probit I-prior models

Consider binary response variables $y_1, \ldots, y_n$

$$y_i \sim \text{Bern}(p_i).$$

We also assume that there exists some continuous underlying latent variables $y_1^*, \ldots, y_n^*$ such that

$$y_i = \begin{cases} 1 & \text{if } y_i^* \geq 0 \\ 0 & \text{if } y_i^* < 0 \end{cases}$$

We consider modelling these latent variables according to the regression problem

$$y_i^* = \alpha + f(x_i) + \epsilon_i,$$

where α is an intercept and the ϵ_is are iid normal with mean zero and some variance $1/\psi$. We can model these underlying latent variables with an I-prior. Let $x_i = (x_{i1}, \ldots, x_{ip})$ be a set of covariates for each of the i observations. Define the kernel matrix as $\mathbf{H}$, where the (i, j) entries of this matrix are

$$\mathbf{H}(i, j) = h(x_i, x_j)$$

with some positive definite kernel function h defined on the set of covariates. An I-prior on the regression function f is then

$$f(x_i) = \lambda \sum_{j=1}^{n} h(x_i, x_j) w_j$$

where λ is the scale parameter of the RKHS with h as the reproducing kernel, and the w_j are iid normal with mean zero and variance ψ. In this case, we can show that

$$\begin{aligned} p_i &= \text{P}[y_i = 1] \\ &= \text{P}[y_i^* \geq 0] \\ &= \text{P}[\alpha + f(x_i) + \epsilon_i \geq 0] \\ &= \text{P}[\epsilon_i < \alpha + f(x_i)] \\ &= \Phi\left(\psi^{1/2}\left(\alpha + \lambda \textstyle\sum_{j=1}^{n} h(x_i, x_j) w_j\right)\right) \end{aligned}$$

where Φ is the cdf of a standard normal density.

1.1 Estimation

The parameters to be estimated are the intercept α and the RKHS scale parameters λ. Denote these collectively by $\theta = (\alpha, \lambda)$, and by p the relevant density/probability mass functions. The

likelihood from a single observation y_i is given by

$$\begin{aligned} L(\theta|\mathbf{y}) &= \int p(\mathbf{y}|\mathbf{w},\theta)p(\mathbf{w})\,\mathrm{d}\mathbf{w} \\ &= \int \frac{e^{-\sum_{i=1}^n w_i/2}}{(2\pi)^{n/2}} \prod_{i=1}^n p(y_i|\mathbf{w},\theta)\,\mathrm{d}\mathbf{w} \end{aligned}$$

which cannot be evaluated analytically. We try five strategies:

1. Naive MC integral *(**BAD**), too high dimensionality*
2. MC-EM integral *(**BAD**), no convergence*
3. Laplace approximation of the integral *(**GOOD**, but slow)*
4. Modified EM using modes *(**GOOD**, fast, but not sure why it works. Convergence issues)*
5. Fully Bayes estimation using HMC *(**GOOD**, slow, not as accurate)*

In I-prior models with continuous responses, the EM algorithm provided a stable way of obtaining MLEs. However, in the binary case, the E-step involves the conditional density $p(\mathbf{w}|\mathbf{y})$ which is difficult to deal with. Some ways to overcome this was to estimate the E-step via MCMC (method 2), or use the posterior modes of $p(\mathbf{w}|\mathbf{y})$ instead of the posterior means. Out of all the methods, Laplace approximation gave reasonable results, and can be used as a benchmark.

1.2 Location and scale of ϵ_i

For simplicity, we can assume the errors ϵ_i to have a known variance equal to one. If the variance is scaled by ψ', then

$$\begin{aligned} y_i^* &= \alpha + f(x_i) + \psi'\epsilon_i \\ \Rightarrow \frac{y_i^*}{\psi'} &= \frac{\alpha}{\psi'} + \frac{f(x_i)}{\psi'} + \epsilon_i \end{aligned}$$

then the model is unchanged since now the y_i^* and the function f similarly scaled. Note that the value of y_i (0 or 1) depends on the sign of y_i^* and not the scale.

Similarly, the threshold does not matter, because moving the threshold means just moving the location of the function f.

2 Variational inference for probit I-prior models

2.1 Some theory

Variational inference is a deterministic approximation of finding maximum likelihood estimates when the likelihood involves intractable integrals. The term 'variational' comes from variational calculus - the mathematical analysis that deals with optimising functionals.

In standard calculus, we deal with input variables (θ, say) and a function of θ (p, say). We are then interested in solving the maximisation problem

$$\arg\max_{\theta} p(\theta).$$

In the case where p is a likelihood function then the solution to this problem is the maximum likelihood estimate. Typically we derive this by solving first-order conditions ($\delta p(\theta)/\delta\theta = 0$).

Variational calculus allows us to solve maximisation problems involving functionals. In this case, the inputs are functions p, and functionals are merely mappings from a set of functions to the real numbers. An example of a functional is the entropy of a pdf

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

We can pose similar optimisation problems with functionals, such as

$$\arg\max_{p} \mathcal{H}(p),$$

for which the solution is a probability distribution which maximises the entropy function over all possible set of functions p. Variational calculus itself is not in any way a form of approximation. However, it can prove to be unfeasible to explore the set of all possible functions, in which case some restrictions have to be made. For example, we could consider only a certain family of functions, or as we will see later, that the function factorises easily.

2.1.1 The KL divergence

Let us consider an inferential problem for which we have n (assumed) iid observations $y = (y_1, \ldots, y_n)$, and perhaps also some latent variables $z = (z_1, \ldots, z_n)$ that requires taking care of. In a fully Bayesian model, we can think of the z as containing the parameters to be estimated as well. The goal is to find an approximation for the posterior distribution $p(z|y)$ as well as for the likelihood $p(y)$ (the model evidence, in Bayesian terminology).

Consider the Kullback–Leibler divergence between any distribution q of the latent variables z, and the posterior $p(z|y)$

$$\mathrm{KL}(q\|p) = \mathrm{E}_q\left[\log \frac{q(z)}{p(z|y)}\right] = \int q(z) \log \frac{q(z)}{p(z|y)} \, \mathrm{d}z.$$

It is interesting to note that the log-likelihood $\log p(y)$ can be decomposed into a term which involves a KL divergence between some distribution q and the posterior, and a linear functional of q:

$$\begin{aligned}
\log p(y) &= \log p(y, z) - \log p(z|y) \\
\log p(y) &= \log p(y, z) - \log p(z|y) - \log q(z) + \log q(z) \\
\int \log p(y) q(z) \, \mathrm{d}z &= \int \left\{ \log \frac{p(y, z)}{q(z)} - \log \frac{p(z|y)}{q(z)} \right\} q(z) \, \mathrm{d}z \\
\log p(y) &= \mathcal{L}(q) + \mathrm{KL}(q\|p) \\
&\geq \mathcal{L}(q)
\end{aligned}$$

From the properties of the KL divergence, we know that it is a positive quantity. Thus, the functional $\mathcal{L}$ is typically referred to as the *lower bound*, and this serves as the proxy objective function in the likelihood maximisation problem. Note that maximising the lower bound is equivalent to minimising the KL-divergence. Of course $\mathrm{KL}(q\|p) \geq 0$ and achieves equality if and only if $q \equiv p$, but for whatever reason we cannot work with the posterior distribution $p(z|y)$ and instead must make some approximation to it in the form of $q(z)$. Incidentally, the EM algorithm is a special case of the variational inference in which $q \equiv p$. In such cases, one can either get closed-form estimates of the E-step involving the posterior distribution, or find ways around it by other estimation techniques.

2.1.2 Factorised distributions

In order to proceed with variational inference, we first make some assumptions about the distribution q. Our goal really is to restrict the form of q such that computations become tractable. Suppose we partition the elements of z into m disjoint groups $z = (z^{(1)}, \ldots, z^{(m)})$. We consider a restriction on q such that

$$q(z) = \prod_{i=1}^{m} q_i(z^{(i)}), \tag{1}$$

i.e., the distribution q factorises with respect to the m groups. This type of approximation has also been studied in Physics under mean-field theory. Among all distributions q which have the form (1), we seek to find one which maximises the lower bound $\mathcal{L}(q)$. Consider first the impact of this mean field assumption on the functional $\mathcal{L}(q)$:

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_{i=1}^{m} (q_i \,\mathrm{d}z_i) \log\left(\frac{p(y,z)}{\prod_{i=1}^{m} q_i}\right) \\
&= \int \prod_{i=1}^{m} (q_i \,\mathrm{d}z_i) \left(\log p(y,z) - \sum_{i=1}^{m} \log q_i\right) \\
&= \int q_k \prod_{i\neq k} (q_i \,\mathrm{d}z_i) \left(\log p(y,z) - \log q_k - \sum_{i\neq k} \log q_i\right) \mathrm{d}z_k \\
&= \int q_k \left(\int \prod_{i\neq k} (q_i \,\mathrm{d}z_i) \log p(y,z)\right) \mathrm{d}z_k - \int q_k \log q_k \,\mathrm{d}z_k + \text{const.} \\
&= \int q_k \log \tilde{p}(y,z_k) \,\mathrm{d}z_k - \int q_k \log q_k \,\mathrm{d}z_k + \text{const.} \\
&= -\,\mathrm{KL}(q_k\|\tilde{p}) + \text{const.}
\end{aligned}$$

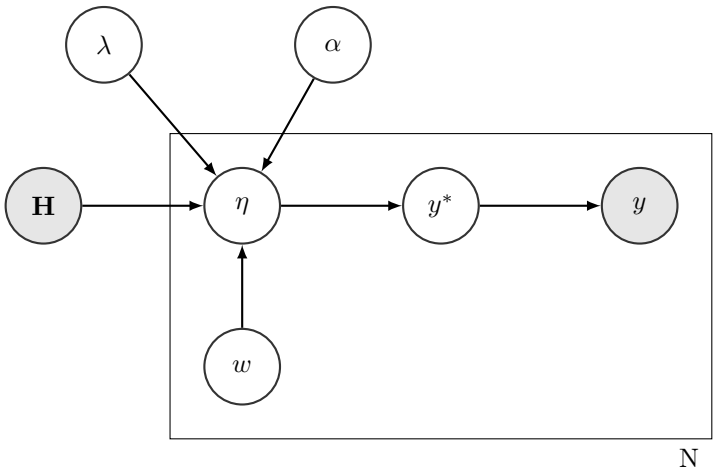
where we have defined a new distribution $\tilde{p}(y, z_k)$ by the relation

$$\begin{aligned}
\log \tilde{p}(y,z_k) &= \int \prod_{i\neq k} (q_i \,\mathrm{d}z_i) \log p(y,z) + \text{const.} \\
&= \mathrm{E}_{-k}\left[\log p(y,z)\right] + \text{const.}
\end{aligned}$$

That is, $\tilde{p}(y, z_k)$ proportional to the exponent of the expectation of the joint distribution $p(y, z)$ under the factorised distribution q as in (1), but excluding factor k. The task of maximising $\mathcal{L}$ is then equivalent to minimising the KL divergence $\mathrm{KL}(q_k \| \tilde{p})$ – for which the solution is $q_j \equiv \tilde{p}(y, z_j)$ for all $j \in \{1, \ldots, m\}$.

In practice the normalising constant does not need to be calculated explicitly, because it can be found by inspection (if the kernel of the density $\tilde{p}$ is a recognisable form). It should be emphasised that the factorisation is the only assumption made to restrict the family of q, and that no explicit assumption about the functional form of q is made.

2.2 DAG for the probit I-prior model



2.3 Distributions

2.3.1 Priors

$$p(w_1, \ldots, w_n) \equiv [\mathrm{N}(0, 1)]^n$$
$$p(\lambda, \alpha) \propto \text{const.}$$

2.3.2 Joint data and latent

$$\begin{aligned} p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \alpha, \lambda) &= p(\mathbf{y}|\mathbf{y}^*, \mathbf{w}, \alpha, \lambda) p(\mathbf{y}^*, \boldsymbol{\eta}, \mathbf{w}, \alpha, \lambda) \\ &= p(\mathbf{y}|\mathbf{y}^*) p(\mathbf{y}^*|\boldsymbol{\eta}) p(\mathbf{w}) p(\lambda) p(\alpha) \end{aligned}$$

where

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \lambda \mathbf{H} \mathbf{w}$$

2.3.3 pdf/pmf

$$p(y_i|y_i^*) = \mathbb{1}[y_i^* \geq 0]^{y_i}\, \mathbb{1}[y_i^* < 0]^{1-y_i}$$

$$\begin{aligned}
\log p(\mathbf{y}^*|\boldsymbol{\eta}) &= \log \mathrm{N}(\boldsymbol{\eta}, \mathbf{I}_n) \\
&= -\frac{n}{2}\log 2\pi - \frac{1}{2}\|\mathbf{y}^* - \boldsymbol{\eta}\|^2 \\
&= -\frac{n}{2}\log 2\pi - \frac{1}{2}\|\mathbf{y}^* - \boldsymbol{\alpha} - \lambda\mathbf{H}\mathbf{w}\|^2 \\
&= -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}(y_i^* - \alpha - \lambda\mathbf{H}_i\mathbf{w})^2 \\
&= -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}\left(y_i^* - \alpha - \lambda\textstyle\sum_{j=1}^{n} h(x_i, x_j)w_j\right)^2
\end{aligned}$$

$$\begin{aligned}
\log p(\mathbf{w}) &= \log \mathrm{N}(\mathbf{0}, \mathbf{I}_n) \\
&= -\frac{n}{2}\log 2\pi - \frac{1}{2}\|\mathbf{w}\|^2 \\
&= -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n} w_i^2
\end{aligned}$$

2.4 Mean field approximation

$$\begin{aligned}
q(\mathbf{y}^*, \mathbf{w}, \alpha, \lambda) &\equiv q(\mathbf{y}^*)q(\mathbf{w})q(\alpha)q(\lambda) \\
&\equiv \prod_{i=1}^{n} q(y_i^*)q(\mathbf{w})q(\alpha)q(\lambda)
\end{aligned}$$

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

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

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

Case: $y_i = 1$

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

where

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

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

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

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

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

2.4.2 Distribution of $\tilde{q}(\mathbf{w})$

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

Let $\mathbf{A} = \mathrm{E}(\lambda^2)\mathbf{H}^2 + \mathbf{I}_n$ and $\mathbf{a} = \mathrm{E}\,\lambda(\mathrm{E}\,\mathbf{y}^* - \mathrm{E}\,\boldsymbol{\alpha})$. Then, using the fact that

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

we see the $\tilde{q}(\mathbf{w})$ is quadratic in $\mathbf{w}$, and we recognise this as the kernel of a multivariate normal density. Therefore,

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

For convenience later in deriving the lower bound, we note that the second moment of $\tilde{q}(\mathbf{w})$ is equal to $\mathrm{E}[\mathbf{w}\mathbf{w}^\top] = \mathbf{A}^{-1}(\mathbf{I}_n + \mathbf{a}\mathbf{a}^\top \mathbf{A}^{-1}) =: \widetilde{\mathbf{W}}$.

2.4.3 Distribution of $\tilde{q}(\lambda)$

$$\begin{aligned}
\log \tilde{q}(\lambda) &= \mathrm{E}_{\mathbf{y}^*,\mathbf{w},\alpha}\left[-\frac{1}{2}\|\mathbf{y}^* - \boldsymbol{\alpha} - \lambda\mathbf{H}\mathbf{w}\|^2\right] + \text{const.} \\
&= -\frac{1}{2}\,\mathrm{E}_{\mathbf{y}^*,\mathbf{w},\alpha}\left[\lambda^2\,\mathrm{tr}(\mathbf{H}^2\mathbf{w}\mathbf{w}^\top) - 2\lambda(\mathbf{y}^* - \boldsymbol{\alpha})^\top\mathbf{H}\mathbf{w}\right] + \text{const.} \\
&= -\frac{1}{2}\Big[\lambda^2\,\mathrm{tr}\left(\mathbf{H}^2\,\mathrm{E}[\mathbf{w}\mathbf{w}^\top]\right) - 2\lambda(\mathrm{E}\,\mathbf{y}^* - \mathrm{E}\,\boldsymbol{\alpha})^\top\mathbf{H}\,\mathrm{E}(\mathbf{w})\Big] + \text{const.}
\end{aligned}$$

By completing the square, we get that $\tilde{q}(\lambda) \equiv \mathrm{N}(d/c, 1/c)$, where

$$c = \mathrm{tr}\left(\mathbf{H}^2\,\mathrm{E}[\mathbf{w}\mathbf{w}^\top]\right) \quad \text{and} \quad d = (\mathrm{E}\,\mathbf{y}^* - \mathrm{E}\,\boldsymbol{\alpha})^\top\mathbf{H}\,\mathrm{E}(\mathbf{w})$$

2.4.4 Distribution of $\tilde{q}(\alpha)$

$$\begin{aligned}
\log \tilde{q}(\alpha) &= \mathrm{E}_{\mathbf{y}^*,\mathbf{w},\lambda}\left[-\frac{1}{2}\|\mathbf{y}^* - \boldsymbol{\alpha} - \lambda\mathbf{H}\mathbf{w}\|^2\right] + \text{const.} \\
&= -\frac{1}{2}\,\mathrm{E}_{\mathbf{y}^*,\mathbf{w},\alpha}\left[n\alpha^2 - 2\sum_{i=1}^{n}(y_i^* - \lambda\mathbf{H}_i\mathbf{w})\alpha\right] + \text{const.} \\
&= -\frac{n}{2}\left[\alpha^2 - \frac{2\alpha}{n}\sum_{i=1}^{n}\left(\mathrm{E}\,y_i^* - \mathrm{E}(\lambda)\mathbf{H}_i\,\mathrm{E}(\mathbf{w})\right)\right] + \text{const.} \\
&\equiv \mathrm{N}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{E}\,y_i^* - \mathrm{E}(\lambda)\mathbf{H}_i\,\mathrm{E}(\mathbf{w})\right), 1/n\right)
\end{aligned}$$

2.5 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

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

With the exception of $q(\mathbf{y}^*)$, all of the distributions are Gaussian. The following results will be helpful.

Definition 1 (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 1. *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,*

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

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

$$\begin{aligned}
\mathrm{E}\left[\log p(\mathbf{y}^*|\boldsymbol{\eta})\right] - \mathrm{E}\left[\log q(\mathbf{y}^*)\right] &= \sum_{i=1}^{n}\mathrm{E}\left[\log p(y_i^*|\eta_i)\right] + \sum_{i=1}^{n}\mathcal{H}\big(q(y_i^*)\big) \\
&= \cancel{\sum_{i=1}^{n}\mathrm{E}\left(-\frac{1}{2}\log 2\pi - \frac{1}{2}\mathrm{E}(y_i^*-\eta_i)^2\right)} \\
&\quad \cancel{+\sum_{i=1}^{n}\left(\frac{1}{2}\log 2\pi + \frac{1}{2}\left(\mathrm{E}\,y_i^{*2} + \tilde{\eta}_i^2 - 2\tilde{\eta}_i\,\mathrm{E}\,y_i^*\right)\right)} \\
&\quad + \sum_{i=1}^{n}\left(\mathbb{1}[y_i^*\geq 0]\log\Phi(\tilde{\eta}_i) + \mathbb{1}[y_i^*<0]\log\left(1-\Phi(\tilde{\eta}_i)\right)\right) \\
&= \sum_{i=1}^{n}\Big(y_i\log\Phi(\tilde{\eta}_i) + (1-y_i)\log\left(1-\Phi(\tilde{\eta}_i)\right)\Big)
\end{aligned}$$

2.5.2 Terms involving distributions of w

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

2.5.3 Terms involving distribution of $q(\lambda)$

$$\begin{aligned}
-\mathrm{E}\left[\log q(\lambda)\right] &= \mathcal{H}\big(q(\lambda)\big) \\
&= \frac{1}{2}(1 + \log 2\pi) - \frac{1}{2}\log\left[\operatorname{tr}\left(\mathbf{H}^2\widetilde{\mathbf{W}}\right)\right]
\end{aligned}$$

2.5.4 Terms involving distribution of $q(\alpha)$

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

2.6 Prediction

Upon obtaining estimates for the parameters and latent variables $(\hat{\mathbf{y}}^*, \hat{\mathbf{w}}, \hat{\lambda}, \hat{\alpha})$, we are interested in the fitted values $\hat{y}_i, \ldots, \hat{y}_n$ and also the fitted probabilities $(\hat{p}_1, \ldots, \hat{p}_n)$. These can be obtained as follows.

$$(\hat{y}_i, \ldots, \hat{y}_n) = \big(\mathbb{1}[\hat{y}_1^* \geq 0], \ldots, \mathbb{1}[\hat{y}_n^* \geq 0]\big)$$

$$(\hat{p}_1, \ldots, \hat{p}_n) = \Phi(\hat{\alpha}\mathbf{1} + \hat{\lambda}\mathbf{H}\hat{\mathbf{w}})$$

Note the slight abuse of notation: the standard normal cdf Φ as a function of the vector $\hat{\alpha}\mathbf{1}+\hat{\lambda}\mathbf{H}\hat{\mathbf{w}}$ is simply Φ applied element-wise to the input vector.

Suppose instead we wish to predict the classes of a new data set $(\tilde{x}_1, \ldots, \tilde{x}_m)$. Firstly, we calculate the predicted $m \times n$ kernel matrix $\tilde{\mathbf{H}}$ whose (i, j) elements consist of $h(\tilde{x}_i, x_j)$. In this case, set

$$(\hat{y}_i^*, \ldots, \hat{y}_m^*) = \hat{\alpha}\mathbf{1} + \hat{\lambda}\tilde{\mathbf{H}}\hat{\mathbf{w}}$$

and then the predicted values $\hat{y}_i, \ldots, \hat{y}_m$ and probabilities $(\hat{p}_1, \ldots, \hat{p}_m)$ are as above, replacing the kernel matrix with the predicted kernel matrix $\tilde{\mathbf{H}}$.

2.7 New findings 8/4/2017

Suppose we know the posterior predictive distribution $p(f_{new}|y)$. Interested in the posterior predictive distribution of the latent variables $p(y^*_{new}|y)$.

$$\begin{aligned}
p(y^*_{new}|y) &= \int p(y^*_{new}|f_{new}, y)p(f_{new}|y)\,\mathrm{d}f_{new} \\
&= \int \mathrm{N}(f_{new}, 1)\,\mathrm{N}(\mu, \sigma^2)\,\mathrm{d}f_{new} \\
&= \int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y^*_{new} - f_{new})^2\right) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(f_{new} - \mu)^2\right) \mathrm{d}f_{new}
\end{aligned}$$

In the exponent:

$$\begin{aligned}
&\left(y^2 + f^2 - 2fy + \psi(f^2 + \mu^2 - 2\mu f)\right) \\
&= \left((1+\psi)f^2 - 2(y + \psi\mu)f + y^2 + \psi\mu^2\right) \\
&= (1+\psi)\left(f^2 - 2\frac{y+\psi\mu}{1+\psi}f\right) + \left(y^2 + \psi\mu^2\right) \\
&= (1+\psi)\left(f - \frac{y+\psi\mu}{1+\psi}\right)^2 - \frac{(y+\psi\mu)^2}{1+\psi} + \left(y^2 + \psi\mu^2\right) \\
&= (1+\psi)\left(f - \frac{y+\psi\mu}{1+\psi}\right)^2 + \frac{(1+\psi)(y^2 + \psi\mu^2) - (y+\psi\mu)^2}{1+\psi} \\
&= (1+\psi)\left(f - \frac{y+\psi\mu}{1+\psi}\right)^2 + \frac{\cancel{y^2} + \psi\mu^2 + \psi y^2 + \cancel{\psi^2\mu^2} - \cancel{y^2} - \cancel{\psi^2\mu^2} - 2\psi\mu y}{1+\psi} \\
&= (1+\psi)\left(f - \frac{y+\psi\mu}{1+\psi}\right)^2 + \frac{\psi(y^2 + \mu^2 - 2\mu y)}{1+\psi} \\
&= (1+\psi)\left(f - \frac{y+\psi\mu}{1+\psi}\right)^2 + \frac{\psi(y-\mu)^2}{1+\psi}
\end{aligned}$$

Thus

$$
\begin{aligned}
p(y^*_{new}|y) &= \int p(y^*_{new}|f_{new}, y) p(f_{new}|y)\,\mathrm{d}f_{new} \\
&= \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\psi}}{\sqrt{2\pi}} \sqrt{\frac{1+\psi}{1+\psi}} \int \exp\left(-\frac{1+\psi}{2}\left(f - \frac{y+\psi\mu}{1+\psi}\right)^2\right) \exp\left(-\frac{\psi/(1+\psi)}{2}(y-\mu)^2\right) \mathrm{d}f \\
&= \frac{\sqrt{\psi/(1+\psi)}}{\sqrt{2\pi}} \exp\left(-\frac{\psi/(1+\psi)}{2}(y-\mu)^2\right) \int \frac{\sqrt{1+\psi}}{\sqrt{2\pi}} \exp\left(-\frac{1+\psi}{2}\left(f - \frac{y+\psi\mu}{1+\psi}\right)^2\right) \mathrm{d}f \\
&\equiv \mathrm{N}(\mu, \sigma^2 + 1)
\end{aligned}
$$

The fitted probabilities are $\Phi(\mu/\sqrt{\sigma^2+1})$.

An alternative derivation is by studying $p(y_{new}|y)$ (straight to the Bernoullis)

$$
\begin{aligned}
p(y_{new}|y) &= \int p(y_{new}|y, f_{new}) p(f_{new}|y)\,\mathrm{d}f_{new} \\
&= \int \Phi(f_{new})\,\mathrm{N}(\mu, \sigma^2)\,\mathrm{d}f_{new} \\
&= \Phi(\mu/\sqrt{\sigma^2+1})
\end{aligned}
$$

and we get the same results.

Replace $\mu = \mathrm{E}[f|y] = \mathrm{E}_q[f]$ and $\sigma^2 = \mathrm{Var}[f|y] = \mathrm{Var}_q[f]$ the approximated posterior for f.

2.8 The variational Bayes EM 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.

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
  $\tilde{\lambda}^{(0)} \leftarrow 1$
  $\tilde{\lambda}^{sq(0)} \leftarrow 1$ ▷ this is $\mathrm{E}[\lambda^2]$
  $\tilde{\alpha}^{(0)} \leftarrow 0$
  $\tilde{\mathbf{w}}^{(0)} \leftarrow \mathbf{0}_n$ ▷ or draw $w_i^{(0)} \sim \mathrm{N}(0,1)$ for $i = 1, \ldots, n$.
end procedure

procedure UPDATE FOR $y_1^*, \ldots, y_n^*$ (time t)
  $\tilde{\boldsymbol{\eta}}^{(t+1)} \leftarrow \tilde{\alpha}^{(t)}\mathbf{1}_n + \tilde{\lambda}^{(t)}\mathbf{H}\tilde{\mathbf{w}}^{(t)}$
  for $i = 1, \ldots, n$ **do**
    if $y_i = 1$ **then**
      $\tilde{y}_i^{*(t+1)} \leftarrow \tilde{\eta}_i^{(t+1)} + \frac{\phi\left(\tilde{\eta}_i^{(t+1)}\right)}{\Phi\left(\tilde{\eta}_i^{(t+1)}\right)}$
    end if
    if $y_i = 0$ **then**
      $\tilde{y}_i^{*(t+1)} \leftarrow \tilde{\eta}_i^{(t+1)} + \frac{\phi\left(\tilde{\eta}_i^{(t+1)}\right)}{\Phi\left(\tilde{\eta}_i^{(t+1)}\right)-1}$
    end if
  end for
end procedure

procedure UPDATE FOR $\mathbf{w}$ (time t)
  $\mathbf{A} \leftarrow \tilde{\lambda}^{sq(t)}\mathbf{H}^2 + \mathbf{I}_n$
  $\mathbf{a} \leftarrow \tilde{\lambda}^{(t)}\mathbf{H}(\mathbf{y}^{*(t+1)} - \tilde{\alpha}^{(t)})$
  $\tilde{\mathbf{w}}^{(t+1)} \leftarrow \mathbf{A}^{-1}\mathbf{a}$
  $\widetilde{\mathbf{W}}^{(t+1)} \leftarrow \mathbf{A}^{-1}(\mathbf{I}_n + \mathbf{a}\mathbf{a}^\top\mathbf{A}^{-1})$
  $\mathrm{logdetA}^{(t+1)} \leftarrow \log|\mathbf{A}|$
end procedure

procedure UPDATE FOR λ (time t)
  $c \leftarrow \mathrm{tr}\left(\mathbf{H}^2\widetilde{\mathbf{W}}^{(t+1)}\right)$
  $d \leftarrow (\mathbf{y}^{*(t+1)} - \tilde{\alpha}^{(t)})^\top\mathbf{H}\tilde{\mathbf{w}}^{(t+1)}$
  $\tilde{\lambda}^{(t+1)} \leftarrow d/c$
end procedure

procedure UPDATE FOR α (time t)
  $\tilde{\alpha}^{(t+1)} \leftarrow \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{y}_i^{*(t+1)} - \tilde{\lambda}^{(t+1)} \mathbf{H}_i \tilde{\mathbf{w}}^{(t+1)} \right)$
end procedure

procedure CALCULATE LOWER BOUND (time t)
  $\mathcal{L}^{(t)} \leftarrow \frac{1}{2}(n + 2 - \log n) + \log 2\pi + \sum_{i=1}^{n} \left(y_i \log \Phi(\tilde{\eta}_i^{(t)}) + (1 - y_i) \log \left(1 - \Phi(\tilde{\eta}_i^{(t)})\right) \right) - \frac{1}{2} \left(\operatorname{tr} \widetilde{\mathbf{W}}^{(t)} + \operatorname{logdet} \mathrm{A}^{(t)} + \log \left[\operatorname{tr} \left(\mathbf{H}^2 \widetilde{\mathbf{W}}^{(t)} \right) \right] \right)$
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 $y_1^*, \ldots, y_n^*$
    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 $(\hat{y}_1, \ldots, \hat{y}_n) \leftarrow (\mathbb{1}[\hat{y}_1^* \geq 0], \ldots, \mathbb{1}[\hat{y}_n^* \geq 0])$ ▷ predicted classes
return $(\hat{p}_1, \ldots, \hat{p}_n) \leftarrow \Phi \left(\hat{\alpha} \mathbf{1} + \hat{\lambda} \mathbf{H} \hat{\mathbf{w}} \right)$ ▷ predicted probabilities

A Proof of Lemma 1

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

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

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

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

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

Lemma 2. *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

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

(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 2 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 2 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

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

□