To-do list

Contents

Haziq Jamil
Department of Statistics
London School of Economics and Political Science
PhD thesis: 'Regression modelling using Fisher information covariance kernels (I-priors)'

Chapter 5

I-priors for categorical responses

chapter5

In a regression setting such as (1.1), consider polytomous response variables $y_1, \ldots, y_n$, where each y_i takes on exactly one of the values from the set of m possible choices $\mathcal{M} = \{1, \ldots, m\}$. Modelling categorical response variables is of profound interest in statistics, econometrics and machine learning, with applications aplenty. In the social sciences, categorical variables often arise from survey responses, and one may be interested in studying correlations between explanatory variables and the categorical response of interest. Economists are often interested in discrete choice models to explain and predict choices between several alternatives, such as consumer choice of goods or modes of transport. In this age of big data, machine learning algorithms are used for classification of observations based on what is usually a large set of variables or features.

The normality assumption (1.2) is not entirely appropriate anymore. As an extension to the I-prior methodology, we propose a flexible modelling framework suitable for regression of categorical response variables. In the spirit of generalised linear models (McCullagh and Nelder, 1989), we relate class probabilities of the observations to a normal I-prior regression model via a link function. Perhaps though, it is more intuitive to view it as machine learners do: since the regression function is ranged on the entire real line, it is necessary to "squash" it through some sigmoid function to conform it to the interval $[0, 1]$ suitable for probability ranges.

Expanding on this idea further, assume that the y_i's follow a categorical distribution, denoted by

$$y_i \sim \text{Cat}(p_{i1}, \ldots, p_{im}),$$

with the class probabilities satisfying $p_{ij} \geq 0, \forall j = 1, \ldots, m$ and $\sum_{j=1}^{m} p_{ij} = 1$. The probability mass function (pmf) of y_i is given by

$$p(y_i) = p_{i1}^{[y_i=1]} \cdots p_{im}^{[y_i=m]}$$

where the notation $[\cdot]$ refers to the Iverson bracket[1]. The dependence of the class probabilities on the covariates is specified through the relationship

$$g(p_{i1}, \ldots, p_{im}) = \big(\alpha_1 + f_1(x_i), \ldots, \alpha_m + f_m(x_i)\big)$$

where $g : [0,1]^m \to \mathbb{R}^m$ is some specified link function. As we will see later, a normal regression model as in (1.1) subject to (1.2) naturally implies a *probit* link function. With an I-prior assumed on the f_j's, we call this method of probit regression using I-priors the *I-probit* regression model.

Due to the nature of the model, unfortunately, the posterior distribution of the regression functions cannot be found in closed form. In particular, marginalising the I-prior from the joint likelihood involves a high-dimensional intractable integral. We explore a fully Bayesian approach to estimate I-probit models using *variational inference*. The main idea is to replace the difficult posterior distribution with an approximation that is tractable. Working in a Bayesian setting together with variational inference allows us to estimate the model much faster than traditional MCMC sampling methods, yet provides us with the conveniences that come with Bayesian machinery. For example, inferences around log-odds is usually cumbersome for probit models, but a credibility interval can easily be obtained by resampling methods from the relevant posteriors, which are typically made up of densities which are familiar and readily available in software.

By choosing appropriate RKHSs/RKKSs for the regression functions, we are able to fit a multitude of binary and multinomial models, including multilevel or random-effects models, linear and non-linear classification models, and even spatio-temporal models. Examples of these models applied to real-world data is shown in Section 5.7. We find that the many advantages of the normal I-prior methodology transfer over quite well to the I-probit model for binary and multinomial regression.

[1] $[A]$ returns 1 if the proposition A is true, and 0 otherwise. The Iverson bracket is a generalisation of the Kronecker delta.

5.1 A latent variable motivation: the I-probit model

5.2 Identifiability and IIA

sec:iia

5.3 Estimation

5.4 The variational EM algorithm for I-probit models

sec:iprobit
var

We present an EM algorithm to estimate the I-probit latent variables $\mathbf{y}^*$ and $\mathbf{w}$, in which the E-step consists of a mean-field variational approximation of the conditional density $p(\mathbf{y}^*, \mathbf{w}|\mathbf{y}, \theta) = q(\mathbf{y}^*)q(\mathbf{w})$. As per assumptions **??**, **??** and **??**, the parameters of the I-probit model consists of $\theta = \{\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\top, \eta\}$.

The algorithm cycles through a variational inference E-step, in which the variational density $q(\mathbf{y}^*, \mathbf{w}) = \prod_{i=1}^n q(\mathbf{y}_{i\cdot}^*)q(\mathbf{w})$ is optimised with respect to the Kullbeck-Leibler divergence KL $\big[q(\mathbf{y}^*, \mathbf{w})\|p(\mathbf{y}^*, \mathbf{w}|\mathbf{y})\big]$, and an M-step, in which the approximate expected joint density **??** is maximised with respect to the parameters θ. Convergence is assessed by monitoring the ELBO. Apart from the fact that the variational EM algorithm uses approximate conditional distributions and involves matrices $\mathbf{y}^*$ and $\mathbf{w}$, it is very similar to the EM described in Chapter 4, and as such, the efficient computational work derived there is applicable.

5.4.1 The variational E-step

Let $\tilde{q}(\mathbf{y}^*, \mathbf{w})$ be the pdf that minimises the Kullbeck-Leibler divergence KL $\big[q\|p\big]$ subject to the mean-field constraint $q(\mathbf{y}^*, \mathbf{w}) = q(\mathbf{y}^*)q(\mathbf{w})$. By appealing to Bishop (2006, equation 10.9, p. 466), the optimal mean-field variational density $\tilde{q}$ for the latent variables $\mathbf{y}^*$ and $\mathbf{w}$ satisfy

$$\log \tilde{q}(\mathbf{y}^*) = \mathrm{E}_{\mathbf{w}\sim\tilde{q}}\big[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w})\big] + \text{const.} \tag{5.1}$$

{eq:logqyst
ar}

$$\log \tilde{q}(\mathbf{w}) = \mathrm{E}_{\mathbf{y}^*\sim\tilde{q}}\big[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w})\big] + \text{const.} \tag{5.2}$$

{eq:logqw}

where $p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}) = p(\mathbf{y}|\mathbf{y}^*)p(\mathbf{y}^*|\mathbf{w})p(\mathbf{w})$ is as per **??**. We now present the variational densities $\tilde{q}(\mathbf{y}^*)$ and $\tilde{q}(\mathbf{w})$. For further details on the derivation of these densities, please refer to the appendix.

Variational distribution for the latent propensities $\mathbf{y}^*$

The fact that the rows $\mathbf{y}^*_{i\cdot} \in \mathbb{R}^m$, $i = 1, \ldots, n$ of $\mathbf{y}^* \in \mathbb{R}^{n \times m}$ are independent can be exploited, and this results in a further induced factorisation $q(\mathbf{y}^*) = \prod_{i=1}^n q(\mathbf{y}^*_i)$. Define the set $\mathcal{C}_j = \{y^*_{ij} > y^*_{ik} \,|\, \forall k \neq j\}$. Then $q(\mathbf{y}^*_{i\cdot})$ is the density of a multivariate normal distribution with mean $\tilde{\boldsymbol{\mu}}_{i\cdot} = \boldsymbol{\alpha} + \tilde{\mathbf{w}}^\top \mathbf{h}_\eta(x_i)$, where $\tilde{\mathbf{w}} = \mathrm{E}_{\mathbf{w} \sim \tilde{q}} \mathbf{w}$, and variance $\boldsymbol{\Psi}^{-1}$, subject to a truncation of its components to the set $\mathcal{C}_{y_i}$. That is, for each $i = 1, \ldots, n$ and noting the observed categorical response $y_i \in \{1, \ldots, m\}$ for the i'th observation, the $\mathbf{y}^*_i$'s are distributed according to

$$\mathbf{y}^*_{i\cdot} \overset{\text{iid}}{\sim} \begin{cases} \mathrm{N}_m(\tilde{\boldsymbol{\mu}}_{i\cdot}, \boldsymbol{\Psi}^{-1}) & \text{if } y^*_{iy_i} > y^*_{ik}, \forall k \neq y_i \\ 0 & \text{otherwise.} \end{cases} \tag{5.3}$$

{eq:ystardist}

We denote this by $\mathbf{y}^*_{i\cdot} \overset{\text{iid}}{\sim} {}^{\mathrm{t}}\mathrm{N}(\tilde{\boldsymbol{\mu}}_{i\cdot}, \boldsymbol{\Psi}^{-1}, \mathcal{C}_{y_i})$, and the important properties of this distribution are explored in the appendix.

The required expectation $\tilde{\mathbf{y}}^* := \mathrm{E}_{\mathbf{y}^* \sim \tilde{q}} \mathbf{y}^*_{i\cdot} = \mathrm{E}_{\mathbf{y}^* \sim \tilde{q}} (y^*_{i1}, \ldots, y^*_{im})^\top$ in the M-step can be tricky to obtain. One strategy that can be considered is Monte Carlo integration: using samples from $\mathrm{N}_m(\tilde{\boldsymbol{\mu}}_{i\cdot}, \boldsymbol{\Psi}^{-1})$, disregard those that do not satisfy the condition $y^*_{iy_i} > y^*_{ik}, \forall k \neq j$, and then take the sample average. This works reasonably well so long as the truncation region does not fall into the extreme tails of the multivariate normal. Alternatively, a fast, Gibbs-based approach (C. P. Robert, 1995) for sampling from a truncated multivariate normal can be implemented, and this is detailed in the appendix.

If the independent I-probit model is under consideration, whereby the covariance matrix has the independent structure $\boldsymbol{\Psi} = \mathrm{diag}(\sigma_1^{-2}, \ldots, \sigma_m^{-2})$, then the first moment can be considered component-wise. Each component of this expectation is given by

$$\tilde{y}^*_{ik} = \begin{cases} \tilde{\mu}_{ik} - \sigma_k C_i^{-1} \displaystyle\int \phi_{ik}(z) \prod_{l \neq k, y_i} \Phi_{il}(z) \phi(z) \, \mathrm{d}z & \text{if } k \neq y_i \\ \tilde{\mu}_{iy_i} - \sigma_{y_i} \sum_{k \neq y_i} \left(\tilde{y}^*_{ik} - \tilde{\mu}_{ik} \right) & \text{if } k = y_i \end{cases} \tag{5.4}$$

{eq:ystarupdate}

with

$$\phi_{ik}(Z) = \phi\left(\frac{\sigma_{y_i}}{\sigma_k}Z + \frac{\tilde{\mu}_{iy_i} - \tilde{\mu}_{ik}}{\sigma_k}\right)$$
$$\Phi_{ik}(Z) = \Phi\left(\frac{\sigma_{y_i}}{\sigma_k}Z + \frac{\tilde{\mu}_{iy_i} - \tilde{\mu}_{ik}}{\sigma_k}\right)$$
$$C_i = \int \prod_{l \neq j} \Phi_{il}(z)\phi(z)\,\mathrm{d}z$$

and $Z \sim \mathrm{N}(0,1)$ with pdf and cdf $\phi(\cdot)$ and $\Phi(\cdot)$ respectively. The integrals that appear above are functions of a unidimensional Gaussian pdf, and these can be computed rather efficiently using quadrature methods.

Variational distribution for the I-prior random effects w

Given that both $\operatorname{vec}\mathbf{y}^*|\operatorname{vec}\mathbf{w}$ and $\operatorname{vec}\mathbf{w}$ are normally distributed as per the model **??**, we find that the full conditional distribution $p(\mathbf{w}|\mathbf{y}^*,\mathbf{y}) \propto p(\mathbf{y}^*,\mathbf{y},\mathbf{w}) \propto p(\mathbf{y}^*|\mathbf{w})p(\mathbf{w})$ is also normal. The variational density q for $\operatorname{vec}\mathbf{w} \in \mathbb{R}^{nm}$ is found to be Gaussian with mean and precision given by

$$\operatorname{vec}\tilde{\mathbf{w}} = \tilde{\mathbf{V}}_w(\boldsymbol{\Psi} \otimes \mathbf{H}_\eta)\operatorname{vec}(\tilde{\mathbf{y}}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top) \quad \text{and} \quad \tilde{\mathbf{V}}_w^{-1} = \boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2 + \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n = \mathbf{V}_{y^*}. \tag{5.5}$$

{eq:varipostw}

As a computational remark, computing the inverse $\tilde{\mathbf{V}}_w^{-1}$ presents a challenge, as this takes $O(n^3m^3)$ time if computed naïvely. By exploiting the Kronecker product structure in $\tilde{\mathbf{V}}_w$, we are able to efficiently compute the required inverse in roughly $O(n^3m)$ time—see the Section X for details. Storage requirement is $O(n^2m^2)$, as a result of the covariance matrix in (5.5).

If the independent I-probit model is assumed, i.e. $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1,\ldots,\psi_m)$, then the posterior covariance matrix $\tilde{\mathbf{V}}_w$ has a simpler structure which implies column independence in the matrix $\mathbf{w}$. By writing $\mathbf{w}_{\cdot j} = (w_{1j},\ldots,w_{nj})^\top \in \mathbb{R}^n$, $j = 1,\ldots,m$, to denote the column vectors of $\mathbf{w}$, and with a slight abuse of notation, we have that

$$\mathrm{N}_{nm}(\operatorname{vec}\mathbf{w}|\operatorname{vec}\tilde{\mathbf{w}},\tilde{\mathbf{V}}_w) = \prod_{j=1}^{m}\mathrm{N}_n(\mathbf{w}_{\cdot j}|\tilde{\mathbf{w}}_{\cdot j},\tilde{\mathbf{V}}_{w_j}),$$

where

$$\tilde{\mathbf{w}}_{\cdot j} = \psi_j \tilde{\mathbf{V}}_{w_j} \mathbf{H}_\eta (\tilde{\mathbf{y}}_j^* - \alpha_j \mathbf{1}_n) \quad \text{and} \quad \tilde{\mathbf{V}}_{w_j} = \left(\psi_j \mathbf{H}_\eta^2 + \psi_j^{-1} \mathbf{I}_n\right)^{-1}.$$

We note the similarity between (5.5) above and the posterior distribution for the I-prior random effects in a normal model (4.7) seen in the previous chapter, with the difference being (5.5) uses the continuous latent propensities $\mathbf{y}^*$ instead of the the observations $\mathbf{y}$. The consequence of this is that the posterior regression functions are class independent, the exact intended effect by specifying a diagonal precision matrix $\boldsymbol{\Psi}$. Storage requirement is $O(n^2 m)$, since we need $\mathbf{V}_{w_1}, \ldots, \mathbf{V}_{w_m}$.

Remark 5.1. The variational distribution $q(\mathbf{w})$ which approximates $p(\mathbf{w}|\mathbf{y})$ is in fact exactly $p(\mathbf{w}|\mathbf{y}^*)$, the conditional density of the I-prior random effects given the latent propensities. By the law of total expectations,

$$\mathrm{E}[r(\mathbf{w})|\mathbf{y}] = \mathrm{E}_{\mathbf{y}^*}\left[\,\mathrm{E}[r(\mathbf{w})|\mathbf{y}^*]\,|\,\mathbf{y}\right],$$

where $r(\cdot)$ is some function of $\mathbf{w}$, and expectations are taken under the posterior distribution of $\mathbf{y}^*$. Hypothetically, if the true pdf $p(\mathbf{y}^*|\mathbf{y})$ were tractable, then the E-step can be computed using the true conditional distribution. Since it is not tractable, we resort to an approximation, and in the case of a variational approximation, (5.5) is obtained.

5.4.2 The M-step

sec:varupdeta

From **??**, the function to be maximised in the M-step is

$$\begin{aligned} Q(\theta) = \text{const.} &- \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Psi}\,\mathrm{E}[\mathbf{w}^\top \mathbf{H}_\eta^2 \mathbf{w}] + \boldsymbol{\Psi}^{-1}\,\mathrm{E}[\mathbf{w}^\top \mathbf{w}]\right) \\ &- \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Psi}\left(\mathrm{E}[\mathbf{y}^{*\top}\mathbf{y}^*] + n\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - 2\tilde{\mathbf{y}}^{*\top}\mathbf{1}_n\boldsymbol{\alpha}^\top - 2\tilde{\mathbf{w}}^\top \mathbf{H}_\eta(\tilde{\mathbf{y}}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top)\right)\right), \end{aligned}$$

where expectations are taken with respect to the variational distributions of $\mathbf{y}^*$ and $\mathbf{w}$. Note that since $\boldsymbol{\Psi}$ is treated as fixed, the term $\mathrm{E}[\mathbf{y}^{*\top}\mathbf{y}^*]$ is absorbed into the constant. On closer inspection, the trace involving the second moments of $\mathbf{w}$ is found to be

$$\operatorname{tr}\left(\boldsymbol{\Psi}\,\mathrm{E}[\mathbf{w}^\top \mathbf{H}_\eta^2 \mathbf{w}] + \boldsymbol{\Psi}^{-1}\,\mathrm{E}[\mathbf{w}^\top \mathbf{w}]\right) = \sum_{i,j=1}^{m}\left\{\psi_{ij}\operatorname{tr}(\mathbf{H}_\eta^2 \tilde{\mathbf{W}}_{ij}) + \psi_{ij}^{-}\operatorname{tr}(\tilde{\mathbf{W}}_{ij})\right\}$$

by the results of equation derived in the appendix. In the above, we had defined ψ_{ij}^{-} to

be the (i,j)'th element of $\boldsymbol{\Psi}^{-1}$, and

$$\tilde{\mathbf{W}}_{ij} = \mathrm{E}[\mathbf{w}_{\cdot i}\mathbf{w}_{\cdot j}^\top] = \mathbf{V}_w[i,j] + \tilde{\mathbf{w}}_{\cdot i}\tilde{\mathbf{w}}_{\cdot j}^\top,$$

where $\mathbf{V}_w[i,j] \in \mathbb{R}^{n\times n}$ refers to the (i,j)'th submatrix block of $\mathbf{V}_w$, and the n-vector $\tilde{\mathbf{w}}_{\cdot j} = \big(\mathrm{E}[w_{ij}]\big)_{i=1}^n$ is the expected value of the random effects for class j. Specifically, when the error precision is of the form $\boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \dots, \psi_m)$, this trace reduces to

$$\begin{aligned}\mathrm{tr}\left(\boldsymbol{\Psi}\,\mathrm{E}[\mathbf{w}^\top\mathbf{H}_\eta^2\mathbf{w}] + \boldsymbol{\Psi}^{-1}\,\mathrm{E}[\mathbf{w}^\top\mathbf{w}]\right) &= \sum_{j=1}^m \left\{\psi_j\,\mathrm{tr}(\mathbf{H}_\eta^2\tilde{\mathbf{W}}_{jj}) + \psi_j^{-1}\,\mathrm{tr}(\tilde{\mathbf{W}}_{jj})\right\} \\ &= \sum_{j=1}^m \mathrm{tr}\left(\overbrace{(\psi_j\mathbf{H}_\eta^2 + \psi_j^{-1}\mathbf{I}_n)}^{\boldsymbol{\Sigma}_{\theta,j}}\tilde{\mathbf{W}}_{jj}\right)\end{aligned}$$

The bulk of the computational effort required to evaluate $Q(\theta)$ stems from the trace involving the second moments of $\mathbf{w}$, and the fact that $\mathbf{H}_\eta^2$ needs to be reevaluated each time $\theta = \{\boldsymbol{\alpha}, \eta\}$ changes. As discussed previously, each E-step takes $O(n^3m)$ time to compute the required first and second (approximate) posterior moments of $\mathbf{w}$. Once this is done, we can use the 'front-loading of the kernel matrices' trick described in Section 4.3.2, which effectively renders the evaluation of Q to be linear in θ (after an initial $O(n^2)$ procedure at the beginning).

As in the normal linear model, we employ a sequential update of the parameters (à la expectation conditional maximisation algorithm) by solving the first order conditions

$$\frac{\partial}{\partial\eta}Q(\eta|\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i,j=1}^m \psi_{ij}\,\mathrm{tr}\left(\frac{\partial\mathbf{H}_\eta^2}{\partial\eta}\tilde{\mathbf{W}}_{ij}\right) + \mathrm{tr}\left(\boldsymbol{\Psi}\tilde{\mathbf{w}}^\top\frac{\partial\mathbf{H}_\eta}{\partial\eta}(\tilde{\mathbf{y}}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top)\right) \tag{5.6}$$ {eq:vemeta}

$$\frac{\partial}{\partial\boldsymbol{\alpha}}Q(\boldsymbol{\alpha}|\eta) = 2n\boldsymbol{\Psi}\boldsymbol{\alpha} - 2\sum_{i=1}^n \boldsymbol{\Psi}\big(\mathbf{y}_{i\cdot}^* - \tilde{\mathbf{w}}^\top\mathbf{h}_\eta(x_i)\big) \tag{5.7}$$ {eq:vemalpha}

equated to zero, where $\mathbf{h}_\eta(x_i) \in \mathbb{R}^n$ is the i'th row of the kernel matrix $\mathbf{H}_\eta$. We now present the update equations for the parameters.

Update for kernel parameters η

When only ANOVA RKHS scale parameters are involved, then the conditional solution of η to (5.6) can be found in closed-form, much like in the exponential family EM algorithm

described in Section 4.3.3. Under the same setting as in that subsection, assume that only $\eta = \{\lambda_1, \ldots, \lambda_p\}$ need be estimated, and for each $k = 1, \ldots, p$, we can decompose the kernel matrix as $\mathbf{H}_\eta = \lambda_k \mathbf{R}_k + \mathbf{S}_k$ and its square as $\mathbf{H}_\eta^2 = \lambda_k^2 \mathbf{R}_k^2 + \lambda_k \mathbf{U}_k + \mathbf{S}_k^2$. As a follow-on from (5.6), the conditional solution for λ_k given the rest of the parameters is obtained by solving

$$\begin{aligned}
\frac{\partial}{\partial \lambda_k} Q(\lambda_k | \boldsymbol{\alpha}, \boldsymbol{\lambda}_{-k}) &= -\frac{1}{2} \sum_{i,j=1}^{m} \psi_{ij} \operatorname{tr}\left((2\lambda_k \mathbf{R}_k^2 + \mathbf{U}_k) \tilde{\mathbf{W}}_{ij} \right) + \operatorname{tr}\left(\boldsymbol{\Psi} \tilde{\mathbf{w}}^\top \mathbf{R}_k (\tilde{\mathbf{y}}^* - \mathbf{1}_n \boldsymbol{\alpha}^\top) \right) \\
&= -\lambda_k \sum_{i,j=1}^{m} \psi_{ij} \operatorname{tr}(\mathbf{R}_k^2 \tilde{\mathbf{W}}_{ij}) - \frac{1}{2} \sum_{i,j=1}^{m} \psi_{ij} \operatorname{tr}(\mathbf{U}_k \tilde{\mathbf{W}}_{ij}) \\
&\quad + \operatorname{tr}\left(\boldsymbol{\Psi} \tilde{\mathbf{w}}^\top \mathbf{R}_k (\tilde{\mathbf{y}}^* - \mathbf{1}_n \boldsymbol{\alpha}^\top) \right) \\
&= 0.
\end{aligned}$$

This yields the solution

$$\hat{\lambda}_k = \frac{\operatorname{tr}\left(\boldsymbol{\Psi} \tilde{\mathbf{w}}^\top \mathbf{R}_k (\tilde{\mathbf{y}}^* - \mathbf{1}_n \boldsymbol{\alpha}^\top) \right) - \frac{1}{2} \sum_{i,j=1}^{m} \psi_{ij} \operatorname{tr}(\mathbf{U}_k \tilde{\mathbf{W}}_{ij})}{\sum_{i,j=1}^{m} \psi_{ij} \operatorname{tr}(\mathbf{R}_k^2 \tilde{\mathbf{W}}_{ij})}$$

In the case of the independent I-probit model, where $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1, \ldots, \psi_m)$, $\hat{\lambda}_k$ has the form

$$\hat{\lambda}_k = \frac{\sum_{j=1}^{m} \psi_j \left(\tilde{\mathbf{w}}_{\cdot j}^\top \mathbf{R}_k (\tilde{\mathbf{y}}_{\cdot j}^* - \alpha_j \mathbf{1}_n) - \frac{1}{2} \operatorname{tr}(\mathbf{U}_k \tilde{\mathbf{W}}_{jj}) \right)}{\sum_{j=1}^{m} \psi_j \operatorname{tr}(\mathbf{R}_k^2 \tilde{\mathbf{W}}_{jj})}.$$

Remark 5.2. There is no closed-form solution for η when the polynomial kernel is used, or when there are kernel parameters to optimise (e.g. Hurst coefficient or SE kernel lengthscale). In these situations, solutions for η are obtained using numerical methods (i.e. employ quasi-Newton methods such as L-BFGS algorithm for optimising $Q(\eta | \boldsymbol{\alpha})$).

Update for intercepts $\boldsymbol{\alpha}$

It is easy to see that the unique solution to (5.7) is

$$\hat{\boldsymbol{\alpha}} = \frac{1}{n} \boldsymbol{\Psi}^{-1} \left(\sum_{i=1}^{n} \boldsymbol{\Psi} \left(\mathbf{y}_{i\cdot}^* - \tilde{\mathbf{w}}^\top \mathbf{h}_\eta(x_i) \right) \right) = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{y}_{i\cdot}^* - \tilde{\mathbf{w}}^\top \mathbf{h}_\eta(x_i) \right) \in \mathbb{R}^m.$$

Being free of $\boldsymbol{\Psi}$, the solution is the same whether the full or independent I-probit model is assumed. Furthermore, we must have that $\sum_{j=1}^{m} \alpha_j = 0$ for identifiability, so as an additional step to satisfy this condition, the solution $\boldsymbol{\alpha}$ is centred.

5.4.3 Summary

A summary of the variational EM algorithm is presented. Notice that the evaluation of each component of the posterior depends on knowing the posterior distribution of the other, i.e. $q(\mathbf{y}^*)$ depends on $q(\mathbf{w})$ and vice-versa. Similarly, each parameter update is obtained conditional upon the value of the rest of the parameters. These circular dependencies are dealt with by way of an iterative updating scheme: with arbitrary starting values for the distributions $q^{(0)}(\mathbf{y}^*)$ and $q^{(0)}(\mathbf{w})$, and for the parameters $\theta^{(0)}$, each are updated in turn according to the above derivations.

The updating sequence is repeated until no significant increase in the convergence criterion, the ELBO, is observed. The ELBO for the I-probit model is given by the quantity

$$\mathcal{L}_q(\theta) = \frac{nm}{2} + \sum_{i=1}^{n} \log C_i(\theta) + \frac{1}{2}\log|\tilde{\mathbf{V}}_w| - \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\sum_{i,j=1}^{m} \psi_{ij}^{-} \operatorname{tr}(\tilde{\mathbf{W}}_{ij}), \tag{5.8}$$

where $C_i(\theta)$ is the normalising constant of the distribution ${}^{\text{t}}\text{N}_m(\boldsymbol{\alpha}+\mathbf{w}^\top \mathbf{h}_\eta(x_i), \boldsymbol{\Psi}^{-1}, \mathcal{C}_{y_i})$, with $\mathcal{C}_{y_i} = \{y^*_{iy_i} > y^*_{ik} | \forall k \neq y_i\}$. That is,

$$C_i(\theta) = \int_{\{y^*_{iy_i} > y^*_{ik} | \forall k \neq y_i\}} \cdots \int \phi(y^*_{i1}, \ldots, y^*_{im} | \boldsymbol{\alpha} + \mathbf{w}^\top \mathbf{h}_\eta(x_i), \boldsymbol{\Psi}^{-1}) \, \mathrm{d}y^*_{i1} \cdots \mathrm{d}y^*_{im}.$$

Similar to the EM algorithm, each iteration of the algorithm increases the ELBO to a stationary point (Blei et al., 2017). Unlike the EM algorithm though, the variational EM algorithm does *not* guarantee an increase in the marginal log-likelihood at each step, nor does it guarantee convergence to the global maxima of the log-likelihood.

Further, the ELBO expression to be maximised is often not convex, which means the CAVI algorithm may terminate at local modes, for which they may be many. Note that the variational distribution with the higher ELBO value is the distribution that is closer, in terms of the KL divergence, to the true posterior distribution. In our experience, multiple random starts alleviates this issue for the I-probit model.

alg:varemip
robit

Algorithm 1 Variational EM for the I-probit model (fixed $\boldsymbol{\Psi}$)

procedure INITIALISATION
  Initialise $\theta^{(0)} \leftarrow \{\boldsymbol{\alpha}^{(0)}, \eta^{(0)}\}$
  $\tilde{q}^{(0)}(\mathbf{w}) \leftarrow \mathrm{MN}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$
  $\tilde{q}^{(0)}(\mathbf{y}^*_{i\cdot}) \leftarrow {}^{\mathrm{t}}\mathrm{N}_m(\tilde{\boldsymbol{\alpha}}^{(0)}, \boldsymbol{\Psi}^{-1}, \mathcal{C}_{y_i})$
  $t \leftarrow 0$
end procedure

while not converged **do**
  procedure VARIATIONAL E-STEP
    for $i = 1, \ldots, n$ **do** ▷ Update $\mathbf{y}^*$
      $\tilde{q}^{(t+1)}(\mathbf{y}^*_{i\cdot}) \leftarrow {}^{\mathrm{t}}\mathrm{N}_m\left(\tilde{\boldsymbol{\alpha}}^{(t)} + \tilde{\mathbf{w}}^{(t)\top}\mathbf{h}_{\eta^{(t)}}(x_i), \boldsymbol{\Psi}, \mathcal{C}_{y_i}\right)$
      $\tilde{\mathbf{y}}^{*(t+1)}_{i\cdot} \leftarrow \mathrm{E}_{q^{(t+1)}}[\mathbf{y}^*_{i\cdot}]$
    end for

    $\tilde{\mathbf{V}}_w^{(t+1)} \leftarrow \left((\boldsymbol{\Psi} \otimes \mathbf{H}^2_{\eta^{(t)}}) + (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)\right)^{-1}$ ▷ Update $\mathbf{w}$
    $\operatorname{vec} \tilde{\mathbf{w}}^{(t+1)} \leftarrow \tilde{\mathbf{V}}_w^{(t+1)}(\boldsymbol{\Psi} \otimes \mathbf{H}_{\eta^{(t)}}) \operatorname{vec}(\tilde{\mathbf{y}}^{*(t+1)} - \mathbf{1}_n \boldsymbol{\alpha}^{(t)\top})$
    $\tilde{q}^{(t+1)}(\mathbf{w}) \leftarrow \mathrm{N}_{nm}\left(\operatorname{vec} \tilde{\mathbf{w}}^{(t+1)}, \tilde{\mathbf{V}}_w^{(t+1)}\right)$
  end procedure

  procedure M-STEP
    if ANOVA kernel (closed-form updates) **then** ▷ Update η
      for $k = 1, \ldots, p$ **do**
        $T_{1k} \leftarrow \sum_{i,j=1}^m \psi_{ij} \operatorname{tr}(\mathbf{R}_k^2 \tilde{\mathbf{W}}_{ij})$
        $T_{2k} \leftarrow \operatorname{tr}\left(\boldsymbol{\Psi}\tilde{\mathbf{w}}^\top \mathbf{R}_k(\tilde{\mathbf{y}}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top)\right) - \frac{1}{2}\sum_{i,j=1}^m \psi_{ij} \operatorname{tr}(\mathbf{U}_k \tilde{\mathbf{W}}_{ij})$
        $\lambda_k^{(t+1)} \leftarrow T_{2k}/T_{1k}$
      end for
    else
      $\eta^{(t+1)} \leftarrow \arg\max_\eta Q(\eta|\boldsymbol{\alpha}^{(t)})$ by L-BFGS algorithm
    end if

    $\mathbf{a} \leftarrow \frac{1}{n}\sum_{i=1}^n \left(\tilde{\mathbf{y}}^{*(t+1)}_{i\cdot} - \tilde{\mathbf{w}}^{(t+1)\top}\tilde{\mathbf{h}}_{\eta^{(t+1)}}(x_i)\right)$ ▷ Update $\boldsymbol{\alpha}$
    $\boldsymbol{\alpha}^{(t+1)} \leftarrow \mathbf{a} - \frac{1}{m}\sum_{j=1}^m a_j$
  end procedure

  Calculate ELBO $\mathcal{L}^{(t+1)}$
  $t \leftarrow t + 1$
end while

5.5 Post-estimation

5.6 Computational considerations

5.7 Examples

sec:iprobit
eg

5.8 Conclusion

This work presents an extension of the normal I-prior methodology to fit categorical response models using probit link functions—a methodology we call the I-probit. The main motivation behind this work is to overcome the drawbacks of modelling probabilities using the normal I-prior model. We assumed latent variables that represent 'class propensities' exist, modelled these using a normal I-prior, and simply transformed them via a probit link function. In this way, all of the advantages of the I-prior methodology seen for the normal model are preserved for binary and multinomial regression as well.

The core of this work explores ways in which to overcome the intractable integral presented by the I-probit model in **??**. Techniques such as quadrature methods, Laplace approximation and MCMC tend to fail, or are unsatisfactorily slow to accomplish. The main reason for this is the dimension of this integral, which is nm, and thus for large sample sizes and/or number of classes, is unfeasible with such methods. We turned to variational inference in the face of an intractable posterior density that hampers an EM algorithm, and the result is a sequential updating scheme, similar in time and storage requirements to the EM algorithm.

In terms of similarity to other works, the generalised additive models (GAMs) of Hastie and Tibshirani (1986) comes close. The setup of GAMs is near identical to the I-probit model, although estimation is done differently. GAMs do not assume smooth functions from any RKHS, but instead estimates the f's using a local scoring method or a local likelihood method. Kernel methods for classification are extremely popular in computer science and machine learning; examples include support vector machines (Schölkopf and Smola, 2002) and Gaussian process classification (Rasmussen and Williams, 2006), with the latter being more closely related to the I-probit method. However, Gaussian process classification typically uses the logistic sigmoid function, and estimation most commonly performed using Laplace approximation, but other methods such as expectation propagation (Minka, 2001) and MCMC (Neal, 1999) have been

explored as well. Variational inference for Gaussian process probit models have been studied by Girolami and Rogers (2006), with their work providing a close reference to the variational algorithm employed by us.

Suggestions for future work include:

1. **Estimation of $\boldsymbol{\Psi}$**. A limitation we had to face in this work was to treat $\boldsymbol{\Psi}$ as fixed. This limitation was in part due to the non-conjugate nature of the variational density for $\boldsymbol{\Psi}$. We believe the variational Bayes EM algorithm, which estimates maximum a posteriori values for the parameters, could alleviate this issue. This would bring the estimation procedure on par with the frequentist objective of maximum likelihood via the EM algorithm, albeit with the use of approximate posterior densities (see Section 5.9.4 and **??** for further discussions).

2. **Inclusion of class-specific covariates**. Throughout the chapter, we assumed that covariates were unit-specific, rather than class-specific. One such example is modelling the choice of travel mode between two destinations (car, coach, train or aeroplane) as a function of travel time. Clearly, travel time depends on the mode of transport. This would require a careful rethink of the appropriate RKHS/RKKS to which the regression function belongs: the regression on the latent propensities could be extended as such:

$$y_{ij}^* = \alpha_j + f_j(x_i) + e(z_{ij})$$

and $f_j \in \mathcal{F}_\mathcal{X}$, the RKHS with kernel $h : (\mathcal{X} \times \mathcal{M})^2 \to \mathbb{R}$ defined by $\delta_{jj'}h(x, x')$, and $e \in \mathcal{F}_\mathcal{Z}$, the RKHS of functions of the form $e : \{z_{ij} | i = 1, \dots, n,\ j = 1, \dots, m\} \to \mathbb{R}$. An I-prior would then be applied as usual, but the implications on the estimation would need to be considered as well.

3. **Improving computational efficiency**. The $O(n^3m)$ time requirement for estimating I-probit models hinder its use towards large-data applications. In a limited study, we did not obtain reliable improvements using low-rank approximations of the kernel matrix such as the Nyström method. The key to improving computational efficiency could lie in sparse variational methods, a suggestion that was made to improve normal I-prior models as well.

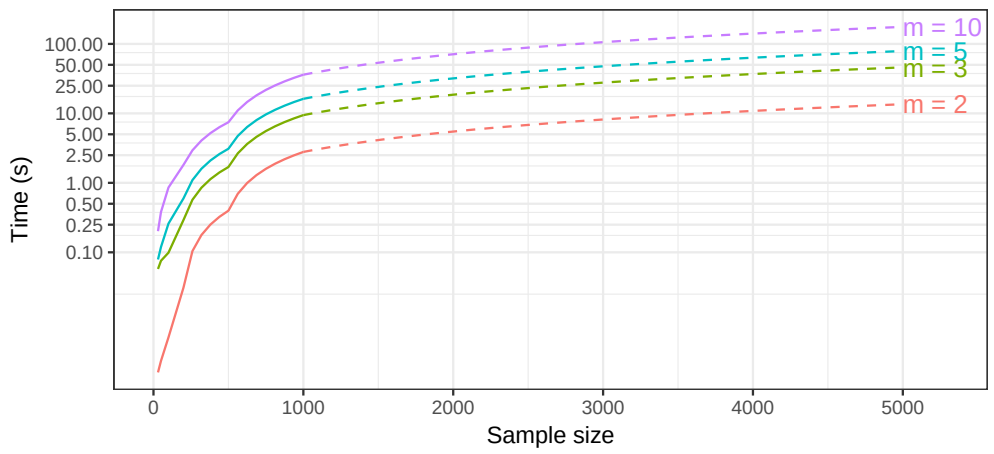

Figure 5.1: Time taken to complete a single variational inference iteration for varying sample sizes and number of classes m. The solid line represents actual timings, while the dotted lines are linear extrapolations.

5.9 Estimation concepts

Consider a statistical model for which we have real-valued observations $\mathbf{y} := \{y_1, \ldots, y_n\}$, which are treated as realisations from an assumed probability distribution with parameters θ. The crux of statistical inference is to estimate θ given the observed values. In the *frequentist setting*, the *likelihood* function, or simply likelihood, is a function of the parameters θ which measures the plausibility of the parameter value given the observed data to fit a statistical model. It is defined as the mapping $\theta \mapsto p(\mathbf{y}|\theta)$, where $p(\mathbf{y}|\theta)$ is the probability density function (or in the case of discrete observations, the probability mass function) of the modelled distribution of the observations.

It is logical to consider the parameter set which provides the largest likelihood value,

$$\hat{\theta}_{\text{ML}} = \arg\max_{\theta} p(\mathbf{y}|\theta). \tag{5.9}$$

The value $\hat{\theta}$ is referred to as the *maximum likelihood estimate* for θ. For convenience, it is often the cases that the *log-likelihood* function $L(\theta) = \log p(\mathbf{y}|\theta)$ is maximised instead. As the logarithm is a monotonically increasing function, the maximiser of the log-likelihood function is exactly the maximiser of the likelihood function itself. Besides invariance, the ML estimate comes with the attractive limiting property $\sqrt{n}(\theta_{\text{ML}} - \theta_{\text{true}}) \xrightarrow{\text{dist.}} \text{N}(0, \mathcal{I})$ (Casella and R. L. Berger, 2002) as sample size $n \to \infty$, where $\mathcal{I}$ is the Fisher information (consistent, efficient, and asymptotically normal).

The *Bayesian* approach to estimating θ takes a different outlook, in that it supplements what is already known from the data with additional information in the form of prior beliefs about the parameters. This usually means treating the parameters as random, following some distribution dictated by a *prior density* $p(\theta)$. There are many ways of categorising different types of priors, but broadly speaking, priors, and hence Bayesian analysis (C. Robert, 2007; Kadane, 2011), can be either *subjective* or *objective*, with the demonyms 'subjectivists' and 'objectivists' used to refer to those subscribing to each respective principle. Subjectivists assert that probabilities are merely opinions, while objectivists, in contrast, view probabilities as an extension of logic. In this regard, objectives Bayes seek to minimise the statistician's contribution to inference and 'let data speak for itself', while subjective Bayes does the opposite.

In either case, inference about the parameters are then performed using the *posterior density*

$$p(\theta|\mathbf{y}) \propto \overbrace{p(\mathbf{y}|\theta)}^{\text{likelihood}} \times \overbrace{p(\theta)}^{\text{prior}}, \tag{5.10}$$

The normalising constant is the marginal likelihood over the distribution of the parameters, $p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta)\,\mathrm{d}\theta$. Note that this quantity is free of θ because the parameters have been marginalised out, or put another way, considered in entirety and averaged over all possible values of θ drawn from its prior density. The quantity $p(\mathbf{y})$ is also known as the *model evidence*, or simply, *evidence*.

The posterior density $p(\theta|\mathbf{y})$ encapsulates the uncertainty surrounding the parameters θ after observing the data $\mathbf{y}$. The *posterior mean*

$$\tilde{\theta} = \int \theta p(\theta|\mathbf{y})\,\mathrm{d}\theta \tag{5.11}$$

{eq:postmeanest}

is normally taken to be the point estimate for θ, with its uncertainty usually reported in the form of a *credible interval*: if θ_k is the k'th component of θ, then a $(1-\alpha)\times 100\%$ credible interval for θ_k is (θ_k^l, θ_k^u), where $\mathrm{P}(\theta_k^l \leq \theta_k \leq \theta_k^u) = (1-\alpha)\times 100\%$. Under a quadratic loss function, $\tilde{\theta}$ minimises the expected loss $\mathrm{E}[(\theta - \theta_{\text{true}})^2]$ (J. O. Berger, 2013, §4.4.2, Result 3), and is hence also viewed as the *minimum mean squared error* (MMSE) estimator.

On a practical note, integration over the parameter space may be intractable, for instance, the model consists of a large number of parameters for which we would like the posterior mean of, or the marginalising integral cannot be found in closed form. Markov

chain Monte Carlo (MCMC) methods are the standard way of approximating such integrals, by way of random sampling from the posterior. The sample $\{\theta^{(1)}, \ldots, \theta^{(T)}\}$ is then manipulated in a way to derive its approximation. In the case of the posterior mean,

$$\mathrm{E}[\hat{\theta}|\mathbf{y}] = \frac{1}{T}\sum_{i=1}^{T}\theta^{(t)} \tag{5.12}$$

gives an approximation, and its $(1-\alpha)\times 100\%$ credible interval can be approximated using the lower $\alpha/2 \times 100\%$ and upper $(1-\alpha/2)\times 100\%$ quantile of the sample.

One may also find the value of θ which maximises the posterior,

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\mathbf{y}|\theta)p(\theta), \tag{5.13}$$

{eq:mapest}

which is the mode of the posterior distribution. This quantity is known as the *maximum a posteriori* (MAP) estimate. It is different from the ML estimate in that the maximisation objective is augmented with with the prior density for θ. In this sense, MAP estimation can be seen as regularisation of the ML estimation procedure, whereby a 'penalty' term is added to avoid overfitting.

MAP estimation is often criticised for not being representative of Bayesian methods. That is, MAP estimation returns a point estimation with no apparent way of quantifying uncertainty of this point estimate. Furthermore, unlike ML estimators, MAP estimators are invariant under reparameterisation. If θ is a random variable with density $p(\theta)$, then the pdf of $\xi := g(\theta)$, where $g : \theta \mapsto g(\theta)$ is a one-to-one transformation, is

$$p_\xi(\xi) = p_\theta\big(g^{-1}(\xi)\big)\left|\frac{\mathrm{d}}{\mathrm{d}\xi}g^{-1}(\xi)\right|. \tag{5.14}$$

{eq:pdftransform}

The second term in (5.14) is called the *Jacobian (determinant)*. Therefore, a different parameterisation of θ will impact the location of the maximum because of the introduction of the Jacobian into the optimisation objective (5.13).

The term *empirical Bayes* (Robbins, 1956; Casella, 1985) refers to procedure in which features of the prior is informed by the data. This is realised by parameterising the prior by a hyper-parameter η, i.e. $\theta \sim p(\theta|\eta)$. Values for the hyper-parameter are clearly important, because they appear in the posterior for θ:

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta|\eta)}{p(\mathbf{y}|\eta)} \tag{5.15}$$

{eq:empbayes1}

To avoid the subjectivist's approach of specifying values for η a priori, one instead turns to the data for guidance. Information concerning η is contained in the marginal likelihood $p(\mathbf{y}|\eta) = \int p(\mathbf{y}|\theta)p(\theta|\eta)\,\mathrm{d}\theta$. This paves the way for using the *maximum marginal likelihood* estimate

$$\hat{\eta} = \arg\max_{\eta} p(\mathbf{y}|\eta) \tag{5.16}$$

in place of η in the equation of (5.15). This procedure is coined *maximum likelihood type-II* by Rasmussen and Williams (2006), and is commonly referred to as such in the machine learning literature. It is also commonplace in statistics, especially in random-effects or latent variable models which employ a maximum likelihood procedure such as EM algorithm.

As a remark, estimation of η itself can be made to conform to Bayesian philosophy, i.e., by placing priors on it and inferring η through its posterior. Such a procedure is referred to as *Bayesian hierarchical modelling*. A motivation for doing this is because the ML estimate of η ignores any uncertainty in it. Of course, the hyper-prior for η could parameterised by a hyper-hyper-parameter, and itself have a prior, and so on and so forth. Evidently the model is specified until such a point where there are parameters of the model which are left 'unoptimised' and must be specified in subjective manner.

5.9.1 The EM algorithm for ML estimation

direct ml "difficult", meaning no closed form estimates and requires numerical methods. gradient-based newton or quasinewton need derivatives, if not readily available then approximate numerical methods. if the z were known, then it is easy => EM algorithm.

Often times, there are unobserved, random variables $\mathbf{z} = \{z_1, \ldots, z_n\}$ that are assumed to make up the data generative process, prescribed in the statistical model through the *joint pdf* $p(\mathbf{y}, \mathbf{z}|\theta)$. Examples of models that include latent variables are plenty: Gaussian mixture models, latent class analysis, factor models, random coefficient models, and so on. In order to obtain the ML estimates through a direct maximisation of the likelihood, it is necessary to first marginalise out the latent variables via

$$p(\mathbf{y}|\theta) = \int \overbrace{p(\mathbf{y}|\mathbf{z},\theta)p(\mathbf{z}|\theta)}^{p(\mathbf{y},\mathbf{z}|\theta)}\,\mathrm{d}\mathbf{z} \tag{5.17}$$

{eq:varint}

and obtain the *marginal likelihood*. Note that the integral is replaced by a summation over all possible values in the case of discrete latent variables $\mathbf{z}$.

5.9.2 A functional view of EM

5.9.3 A brief introduction to variational inference

sec:varintro

Consider a statistical model for which we have observations $\mathbf{y} := \{y_1, \ldots, y_n\}$, but also some latent variables $\mathbf{z} := \{z_1, \ldots, z_n\}$. Typically, in such models, there is a want to to evaluate the integral

$$\mathcal{I} = \int p(\mathbf{y}|\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}. \tag{5.18}$$

Marginalising out the latent variables in (5.17) is usually a precursor to obtaining a log-likelihood function to be maximised, in a frequentist setting. In Bayesian analysis, the $\mathbf{z}$'s are parameters which are treated as random, and the integral corresponds to the marginal density for $\mathbf{y}$, on which the posterior depends.

In many instances, for one reason or another, evaluation of $\mathcal{I}$ is difficult, in which case inference is halted unless a way of overcoming the intractable integral (5.17) is found. Here, we discus *variational inference* (VI), a fully Bayesian treatment of the statistical model with a deterministic algorithm, i.e. does not involve sampling from posteriors. The crux of variational inference is this: find a suitably close distribution function $q(z)$ that approximates the true posterior $p(\mathbf{z}|\mathbf{y})$, where closeness here is defined in the Kullback-Leibler divergence sense,

$$\mathrm{KL}(q\|p) = \int \log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{y})} q(\mathbf{z})\,\mathrm{d}\mathbf{z}.$$

Posterior inference is then conducted using $q(\mathbf{z})$ in lieu of $p(\mathbf{z}|\mathbf{y})$. Advantages of this method are that 1) it is fast to implement computationally (compared to MCMC); 2) convergence is assessed simply by monitoring a single convergence criterion; and 3) it works well in practice, as attested to by the many studies implementing VI.

Briefly, we present the motivation behind variational inference and the minimisation of the KL divergence. Denote by $q(\cdot)$ some density function of $\mathbf{z}$. One may show that log

marginal density (the log of the intractable integral (5.17)) holds the following bound:

$$\begin{aligned}
\log p(y) &= \log p(\mathbf{y}, \mathbf{z}) - \log p(\mathbf{z}|\mathbf{y}) \quad \text{(Bayes' theorem)} \\
&= \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{z})}{q(\mathbf{z})} - \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} \right\} q(\mathbf{z}) \, \mathrm{d}z \quad \text{(expectations both sides)} \\
&= \mathcal{L}(q) + \mathrm{KL}(q \| p) \\
&\geq \mathcal{L}(q) \qquad (5.19)
\end{aligned}$$

{eq:varbound}

since the KL divergence is a non-negative quantity. The functional $\mathcal{L}(q)$ given by

$$\begin{aligned}
\mathcal{L}(q) &= \int \log \frac{p(\mathbf{y}, \mathbf{z})}{q(\mathbf{z})} q(\mathbf{z}) \, \mathrm{d}\mathbf{z} \\
&= \mathrm{E}_{\mathbf{z} \sim q} \log p(\mathbf{y}, \mathbf{z}) + H(q), \qquad (5.20)
\end{aligned}$$

{eq:elbo1}

where H is the entropy functional, is known as the *evidence lower bound* (ELBO). Evidently, the closer q is to the true p, the better, and this is achieved by maximising $\mathcal{L}$, or equivalently, minimising the KL divergence from p to q. Note that the bound (5.19) achieves equality if and only if $q(\mathbf{z}) \equiv p(\mathbf{z}|\mathbf{y})$, but of course the true form of the posterior is unknown to us—see Section 5.9.4 for a discussion. Maximising $\mathcal{L}(q)$ or minimising $\mathrm{KL}(q\|p)$ with respect to the density q is a problem of calculus of variations, which incidentally, is where variational inference takes its name. The astute reader will realise that $\mathrm{KL}(q||p)$ is impossible to compute, since one does not know the true distribution $p(\mathbf{z}|\mathbf{y})$. Efforts are concentrated on maximising the ELBO instead.

Maximising $\mathcal{L}$ over all possible density functions q is not possible without considering certain constraints. Two such constraints are described. The first, is to make a distributional assumption regarding q, for which it is parameterised by ν. For instance, we might choose the closest normal distribution to the posterior $p(\mathbf{z}|\mathbf{y})$ in terms of KL divergence. In this case, the task is to find optimal mean and variance parameters of a normal distribution.

The second type of constraint, and the one considered in this thesis, is simply an assumption that the approximate posterior q factorises into M disjoint factors. Partition $\mathbf{z}$ into M disjoint groups $\mathbf{z} = (z_{[1]}, \ldots, z_{[M]})$. Note that each factor $z_{[k]}$ may be

[2]Reproduced from the talk by David Blei entitled 'Variational Inference: Foundations and Innovations', 2017. URL: https://simons.berkeley.edu/talks/david-blei-2017-5-1.

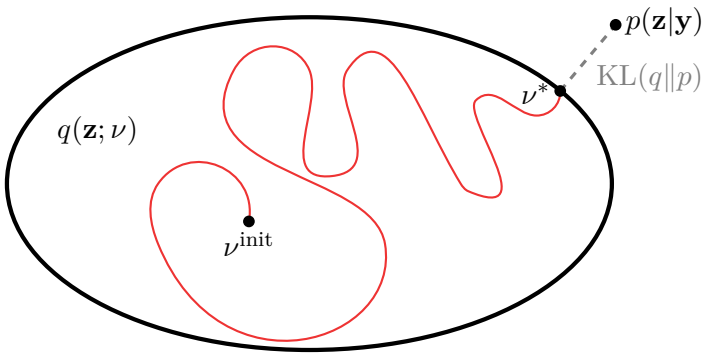

Figure 5.2: Schematic view of variational inference[2]. The aim is to find the closest distribution q (parameterised by a variational parameter ν) to p in terms of KL divergence within the set of variational distributions, represented by the ellipse.

multidimensional. Then, the structure

$$q(\mathbf{z}) = \prod_{k=1}^{M} q_k(z_{[k]})$$

for q is considered. This factorised form of variational inference is known in the statistical physics literature as the *mean-field theory* (Itzykson and Drouffe, 1991).

Let us denote the distributions which minimise the Kullback-Leibler divergence (maximise the variational lower bound) by the use of tildes. By appealing to Bishop (2006, equation 10.9, p. 466), we find that for each $z_{[k]}$, $k = 1, \ldots, M$, $\tilde{q}_k$ satisfies

$$\log \tilde{q}_k(z_{[k]}) = \mathrm{E}_{-k} \log p(\mathbf{y}, \mathbf{z}) + \text{const.} \tag{5.21}$$ {eq:qtilde}

where expectation of the joint log density of $\mathbf{y}$ and $\mathbf{z}$ is taken with respect to all of the unknowns $\mathbf{z}$, except the one currently in consideration $z_{[k]}$, under their respective $\tilde{q}_k$ densities.

In practice, rather than an explicit calculation of the normalising constant, one simply needs to inspect (5.21) to recognise it as a known log-density function, which is the case when exponential family distributions are considered. That is, suppose that each complete conditional $p(z_{[k]}|\mathbf{z}_{-k}, \mathbf{y})$, where $\mathbf{z}_{-k} = \{z_{[i]}|i \neq k\}$, follows an exponential family distribution

$$p(z_{[k]}|\mathbf{z}_{-k}, \mathbf{y}) = B(z_{[k]}) \exp\left(\langle \zeta_k(\mathbf{z}_{-k}, \mathbf{y}), z_{[k]} \rangle - A(\zeta_k)\right).$$

Then, from (5.21),

$$\begin{aligned}\tilde{q}(z_{[k]}) &\propto \exp\left(\mathrm{E}_{-k} \log p(z_{[k]}|\mathbf{z}_{-k}, \mathbf{y})\right) \\ &= \exp\Big(\log B(z_{[k]}) + \mathrm{E}\langle \zeta_k(\mathbf{z}_{-k}, \mathbf{y}), z_{[k]}\rangle - \mathrm{E}[A(\zeta_k)]\Big) \\ &\propto B(z_{[k]}) \exp \mathrm{E}\langle \zeta_\xi(\mathbf{z}_{-k}, \mathbf{y}), z_{[k]}\rangle\end{aligned}$$

is also in the same exponential family. In situations where there is no closed form expression for $\tilde{q}$, then one resorts to sampling methods such as a Metropolis random walk to estimate quantities of interest. This stochastic step within a deterministic algorithm has been explored before in the context of a Monte Carlo EM algorithm—see Meng and Van Dyk (1997, §4, pp. 537–538) and references therein.

One notices that the optimal mean-field variational densities for each component are coupled with one another, in the sense that the distribution $\tilde{q}_k$ depends on the moments of the rest of the components $\mathbf{z}_{-k}$. For very simple problems, an exact solution for each $\tilde{q}_k$ can be found, but usually, the way around this is to employ an iterative procedure. The *coordinate ascent mean-field variational inference* (CAVI) algorithm cycles through each of the distributions in turn, updating them in sequence starting with arbitrary distributions as initial values.

Algorithm 2 The CAVI algorithm

alg:cavi

1: **initialise** Variational factors $q_k(z_{[k]})$
2: **while** ELBO $\mathcal{L}(q)$ not converged **do**
3: **for** $k = 1, \ldots, M$ **do**
4: $\tilde{q}_k(z_{[k]}) \leftarrow$ const. $\times \exp \mathrm{E}_{-k} \log p(\mathbf{y}, \mathbf{z})$ ▷ from (5.21)
5: **end for**
6: $\mathcal{L}(q) \leftarrow \mathrm{E}_{\mathbf{z}\sim\prod_k \tilde{q}_k} \log p(\mathbf{y}, \mathbf{z}) + \sum_{k=1}^{m} H\left[q_k(z_{[k]})\right]$ ▷ Update ELBO
7: **end while**
8: **return** $\tilde{q}(\mathbf{z}) = \prod_{k=1}^{M} \tilde{q}_j(z_{[k]})$

Each iteration of the CAVI brings about an improvement in the ELBO (hence the name coordinate ascent). The algorithm terminates when there is no more significant improvement in the ELBO, indicating a convergence of the CAVI. Blei et al. (2017) notes that the ELBO is typically a non-convex function, in which case convergence may be to (one of possibly many) local optima. A simple solution would be to restart the CAVI at multiple initial values, and the solution giving the highest ELBO is the distribution that is closest to the true posterior.

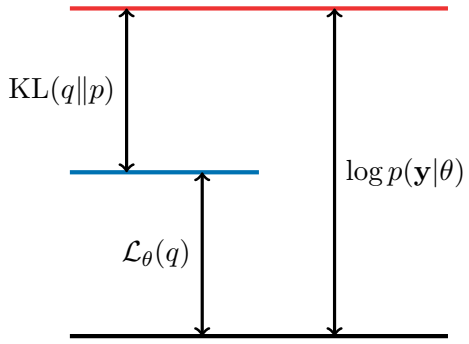

Figure 5.3: Illustration[3] of the decomposition of the log-likelihood into $\mathcal{L}_\theta(q)$ and $\mathrm{KL}[q(\mathbf{z})\|p(\mathbf{z}|\mathbf{y})]$. The quantity $\mathcal{L}_\theta(q)$ is a lower bound for the log-likelihood.

fig:loglikdecomp

5.9.4 Variational methods and the EM algorithm

sec:varEM

Consider again the latent variable setup described in Section 5.9.3, but suppose the goal now is to maximise the (marginal) log-likelihood of the parameters θ of the model. We will see how the EM algorithm relates to minimising the KL divergence between a density $q(\mathbf{z})$ and the posterior of $\mathbf{z}$, and connect this idea to variational methods.

As we did in deriving (5.19), we decompose the marginal log-likelihood as

$$\log p(y|\theta) = \mathrm{E}\left[\log \frac{p(\mathbf{y},\mathbf{z}|\theta)}{q(\mathbf{z})}\right] - \mathrm{E}\left[\log \frac{p(\mathbf{z}|\mathbf{y},\theta)}{q(\mathbf{z})}\right] = \mathcal{L}(q) + \mathrm{KL}(q\|p).$$

This decomposition is shown in Figure 5.3. We realise that the KL divergence non-negative, and is zero exactly when $q(\mathbf{z}) \equiv p(\mathbf{z}|\mathbf{y},\theta)$. Substituting this into the above equation yields the relationship

$$\begin{aligned}\log p(y|\theta) &= \mathrm{E}\left[\log \frac{p(\mathbf{y},\mathbf{z}|\theta)}{p(\mathbf{z}|\mathbf{y},\theta)}\right] - \mathrm{E}\left[\log \frac{p(\mathbf{z}|\mathbf{y},\theta)}{p(\mathbf{z}|\mathbf{y},\theta)}\right] \\ &= \mathrm{E}\log p(\mathbf{y},\mathbf{z}|\theta) - \mathrm{E}\, p(\mathbf{z}|\mathbf{y},\theta).\end{aligned}$$

By taking expectations under the posterior distribution with known parameter values $\theta^{(t)}$, the term on the left becomes the Q function of the E-step

$$Q(\theta) = Q(\theta|\theta^{(t)}) = \mathrm{E}_{\mathbf{z}}\left[\log p(\mathbf{y},\mathbf{z}|\theta)\,\middle|\,\mathbf{y},\theta^{(t)}\right],$$

[3] Reproduced from Bishop (2006, Figure 9.11).

while the term on the left is an entropy term. Thus, minimising the KL divergence corresponds to the E-step in the EM algorithm. As a side fact, for any θ, we find that

$$\begin{aligned}\log p(\mathbf{y}|\theta) - \log p(\mathbf{y}|\theta^{(t)}) &= Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) + \Delta\,\text{entropy} \\ &\geq Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}).\end{aligned}$$

because entropy differences are positive by Gibbs' inequality. We see that maximising Q with respect to θ (the M-step) brings about an improvement to the log-likelihood value. To summarise, the EM algorithm is seen as

- **E-step**. Maximise $\mathcal{L}_\theta\big[q(\mathbf{z})\big]$ with respect to q, keeping θ fixed. This is equivalent to minimising $\mathrm{KL}(q\|p)$.
- **M-step**. Maximise $\mathcal{L}\big[q(\mathbf{z}|\theta)\big]$ with respect to θ, keeping q fixed.

When the true posterior distribution $p(\mathbf{z}|\mathbf{y})$ is not tractable, then the E-step becomes intractable as well. By constraining the maximisation in the E-step to consider q belonging to a family of tractable densities, the E-step yields a variational approximation $\tilde{q}$ to the true posterior. In Section 5.9.3, we saw that constraining q to be of a factorised form, then $\tilde{q}$ is a mean-field density. This form of the EM is known as *variational Bayes EM algorithm* (VB-EM) (Beal and Ghahramani, 2003).

In variational inference, a fully Bayesian treatment of the parameters is considered, with the aim of obtaining approximation to their posterior distributions. In VB-EM, the variational approximation is only performed on the latent, or 'missing' variables, to use the EM nomenclature. After a variational E-step, the M-step proceeds as usual, and as such, all of the material relating to the EM in the previous chapter is applicable. The VB-EM can also be seen as obtaining (approximate) maximum a posteriori estimates with diffuse priors on the parameters.

variational inference, EM algorithm, variational Bayes EM, differences, pros cons, MAP vs MLE, MAP vs fully Bayes

Appendix

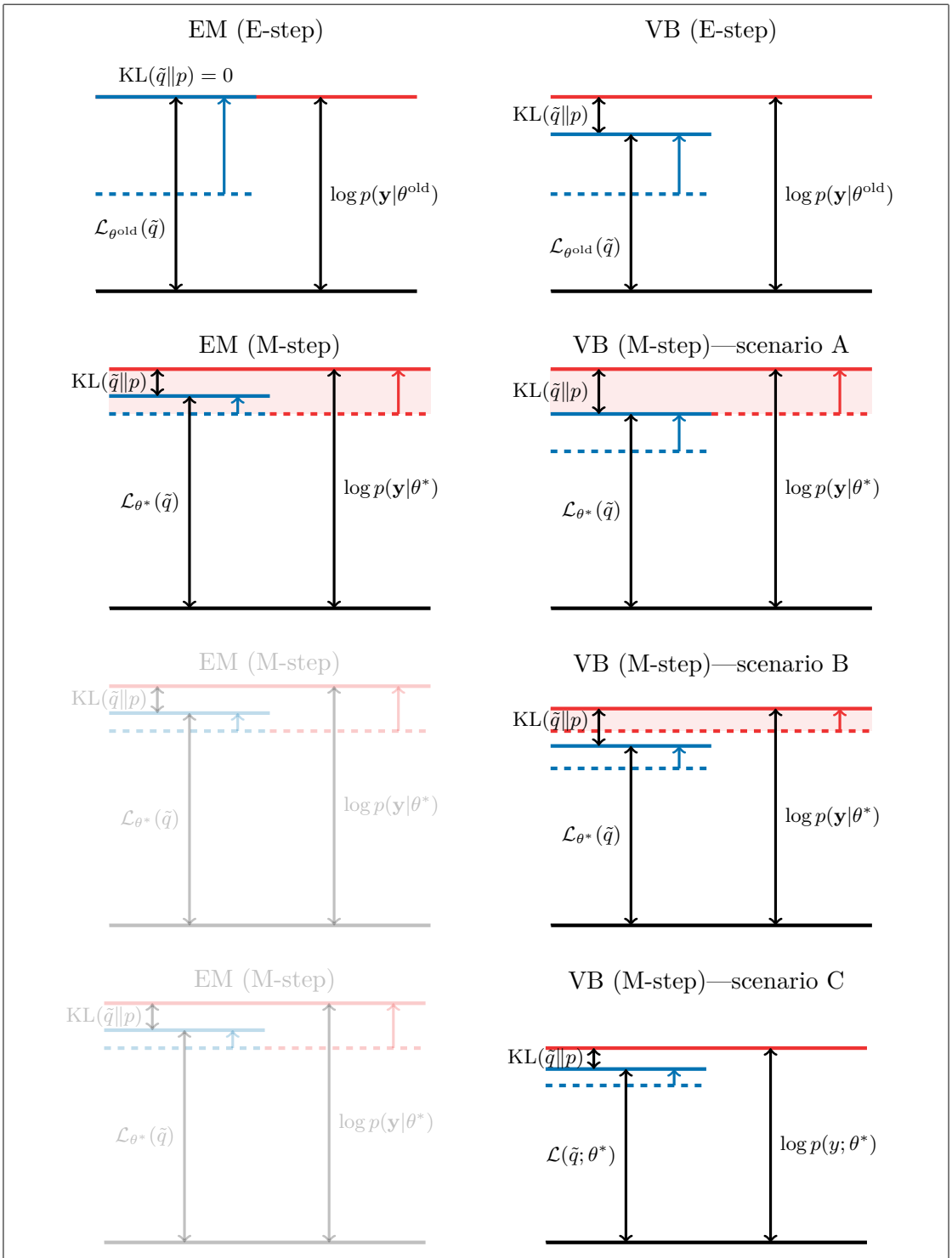

Figure 5.4: Illustration of EM vs VB-EM. Whereas the EM guarantees an increase in log-likelihood value (red shaded region), the VB-EM does not.

Bibliography

beal2003 Beal, M. J. and Z. Ghahramani (2003). "The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures". In: *Bayesian Statistics 7*. Proceedings of the Seventh Valencia International Meeting. Ed. by José M. Bernardo, A. Philip Dawid, James O. Berger, Mike West, David Heckerman, M.J. Bayarri, and Adrian F.M. Smith. Oxford: Oxford University Press, pp. 453–464.

berger2013statistical Berger, James O (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

bishop2006pattern Bishop, Christopher (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag.

blei2017variational Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2017). "Variational inference: A review for statisticians". In: *Journal of the American Statistical Association* just-accepted.

casella1985introduction Casella, George (1985). "An introduction to empirical Bayes data analysis". In: *The American Statistician* 39.2, pp. 83–87.

casella2002statistical Casella, George and Roger L Berger (2002). *Statistical inference*. Vol. 2. Duxbury Pacific Grove, CA.

girolami2006variational Girolami, Mark and Simon Rogers (2006). "Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors". In: *Neural Computation* 18.8, pp. 1790–1817.

hastie1986 Hastie, Trevor and Robert Tibshirani (Aug. 1986). "Generalized Additive Models". In: *Statist. Sci.* 1.3, pp. 297–310. DOI: 10.1214/ss/1177013604. URL: https://doi.org/10.1214/ss/1177013604.

itzykson1991statistical
Itzykson, Claude and Jean Michel Drouffe (1991). *Statistical Field Theory: Volume 2, Strong Coupling, Monte Carlo Methods, Conformal Field Theory and Random Systems*. Cambridge University Press.

kadane2011principles
Kadane, Joseph B (2011). *Principles of uncertainty*. CRC Press.

mccullagh1989
McCullagh, P. and John A. Nelder (1989). *Generalized Linear Models*. 2nd. Chapman & Hall/CRC Press.

meng1997algorithm
Meng, Xiao-Li and David Van Dyk (1997). "The EM Algorithm—an Old Folk-song Sung to a Fast New Tune". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59.3, pp. 511–567.

minka2001expectation
Minka, Thomas P (2001). "Expectation propagation for approximate Bayesian inference". In: *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 362–369.

neal1999
Neal, Radford M. (1999). "Regression and Classification using Gaussian Process Priors". In: *Bayesian Statistics*. Ed. by J M Bernardo, J O Berger, A P Dawid, and A F M Smith. Vol. 6. Oxford University Press. (with discussion), pp. 475–501.

rasmussen2006gaussian
Rasmussen, Carl Edward and Christopher K I Williams (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

robbins1956empirical
Robbins, Herbert (1956). *An empirical Bayes approach to statistics*. Tech. rep. COLUMBIA UNIVERSITY New York City United States.

robert2007bayesian
Robert, Christian (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

robert1995simulation
Robert, Christian P (1995). "Simulation of truncated normal variables". In: *Statistics and computing* 5.2, pp. 121–125.

scholkopf2002learning
Schölkopf, Bernhard and Alexander J Smola (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.