To-do list

1. Exponential family for y not really necessary, it just follows nicely from the latent variable motivation. 2
2. Section X 12
3. Section X 13

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

In a regression setting, consider polytomous response variables $y_1, \dots, y_n$, where each y_i takes on exactly one of the values $\{1, \dots, m\}$ from a set of m possible choices. 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.

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 the class probabilities of the observations to the 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 measures. As in GLMs, the y_i's are assumed to follow an exponential family distribution, and in this case, the categorical distribution. We denote this by

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

1. Exponential family for y not really necessary, it just follows nicely from the latent variable motivation.

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

$$p(y_i) = p_{i1}^{[y_i=1]} \cdots p_{im}^{[y_i=m]} \tag{5.1}$$ {eq:catdist}

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

$$g(p_{ij}) = \big(\alpha_j + f_j(x_i)\big)_{j=1}^{m}$$

where $g : [0, 1] \to \mathbb{R}^m$ is some specified link function. As we will see later, the normality assumption of the errors naturally implies a *probit* link function, i.e., g is the inverse cumulative distribution function (CDF) of a standard normal distribution (or more precisely, a function that *involves* the standard normal CDF). Normality is also a required assumption for I-priors to be specified on the regression functions. We call this method of probit regression using I-priors the *I-probit* regression model.

Note that the probabilities are modelled per class $j \in \{1, \ldots, m\}$ by individual regression curves f_j, and in the most general setting, m sets of intercepts α_j and kernel hyperparameters η_j must be estimated. The dependence of these m curves are specified through covariances $\sigma_{jk} := \text{Cov}[\epsilon_{ij}, \epsilon_{ik}]$, for each $j, k \in \{1, \ldots, m\}$ and $j \neq k$. While it may be of interest to estimate these covariances, this paper considers cases where the regression functions are class independent, i.e. $\sigma_{jk} = 0, \forall j \neq k$. This violates the independence of irrelevant alternatives (IIA) assumption (see Section **??** for details) crucial in choice models, but not so much necessary for classification when the alternatives are distinctively different.

The many advantages of the I-prior methodology of Jamil and Bergsma, 2017 transfer over quite well to the I-probit model for classification and inference. In particular, by choosing appropriate RKHSs 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 **??**. 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 posterior estimates. For example, inferences around log-odds is usually cumbersome

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

for probit models, but a credibility interval can easily be obtained by resampling methods from the relevant posteriors, which are normally exponential family distributions in the I-probit model.

5.1 Miscellanea

5.1.1 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

$$\mathcal{I} = \int p(\mathbf{y}|\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}. \tag{5.2}$$

eq:varint

Models that include latent variables are plenty, for example: Gaussian mixture models, latent class analysis, factor models, random coefficient models, and so on. Marginalising out the latent variables in (5.2) 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 (5.2) 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 (5.2)) holds the following bound:

$$\begin{aligned}
\log p(y) &= \log p(\mathbf{y}, \mathbf{z}) - \log p(\mathbf{z}|\mathbf{y}) \quad \text{(Bayes' theorem)} \\
&= \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{z})}{q(\mathbf{z})} - \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} \right\} q(\mathbf{z}) \, \mathrm{d}z \quad \text{(expectations both sides)} \\
&= \mathcal{L}(q) + \mathrm{KL}(q \| p) \\
&\geq \mathcal{L}(q) \qquad (5.3)
\end{aligned}$$

{eq:varbound}

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

$$\begin{aligned}
\mathcal{L}(q) &= \int \log \frac{p(\mathbf{y}, \mathbf{z})}{q(\mathbf{z})} q(\mathbf{z}) \, \mathrm{d}\mathbf{z} \\
&= \mathrm{E}_{\mathbf{z} \sim q} \log p(\mathbf{y}, \mathbf{z}) + H(q), \qquad (5.4)
\end{aligned}$$

{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 (5.3) 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 5.1.2 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.

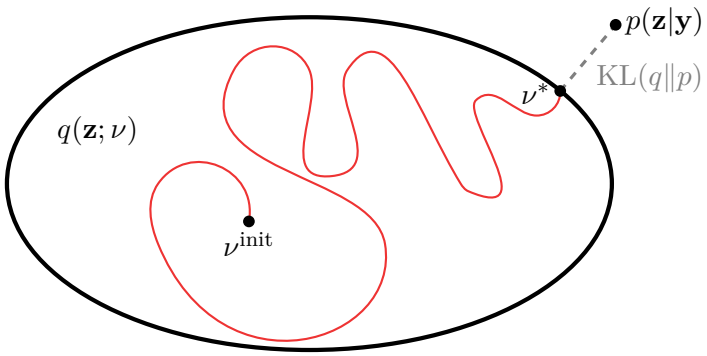

Figure 5.1: Schematic view of variational inference[2]. The aim is to find the closest distribution q (parameterised by a variational parameter ν) to p in terms of KL divergence within the set of variational distributions, represented by the ellipse.

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

$$\log \tilde{q}_k(z_{[k]}) = \mathrm{E}_{-k} \log p(\mathbf{y}, \mathbf{z}) + \text{const.} \tag{5.5}$$ {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 (5.5) 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 (5.5),

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

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 1 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 (5.5)
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.

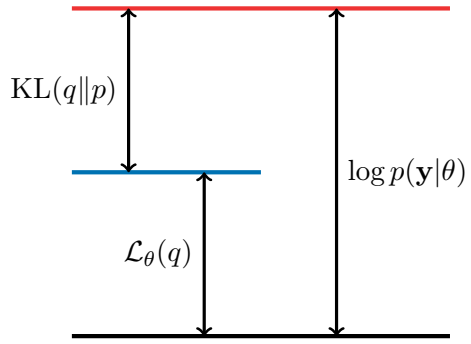

Figure 5.2: Illustration[3] of the decomposition of the log-likelihood into $\mathcal{L}_\theta(q)$ and $\mathrm{KL}[q(\mathbf{z})\|p(\mathbf{z}|\mathbf{y})]$. The quantity $\mathcal{L}_\theta(q)$ is a lower bound for the log-likelihood.

fig:loglikdecomp

5.1.2 Variational methods and the EM algorithm

sec:varEM

Consider again the latent variable setup described in Section 5.1.1, 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 (5.3), 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 5.2. 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 5.1.1, 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.

5.1.3 The EM algorithm for I-probit models is intractable—variational Bayes EM?

Consider employing an EM algorithm, similar to the one seen in the previous chapter, to estimate I-probit models. This time, treat both the latent propensities $\mathbf{y}^*$ and the I-prior random effects $\mathbf{w}$ as 'missing', so the complete data is $\{\mathbf{y}, \mathbf{y}^*, \mathbf{w}\}$. Now, due to

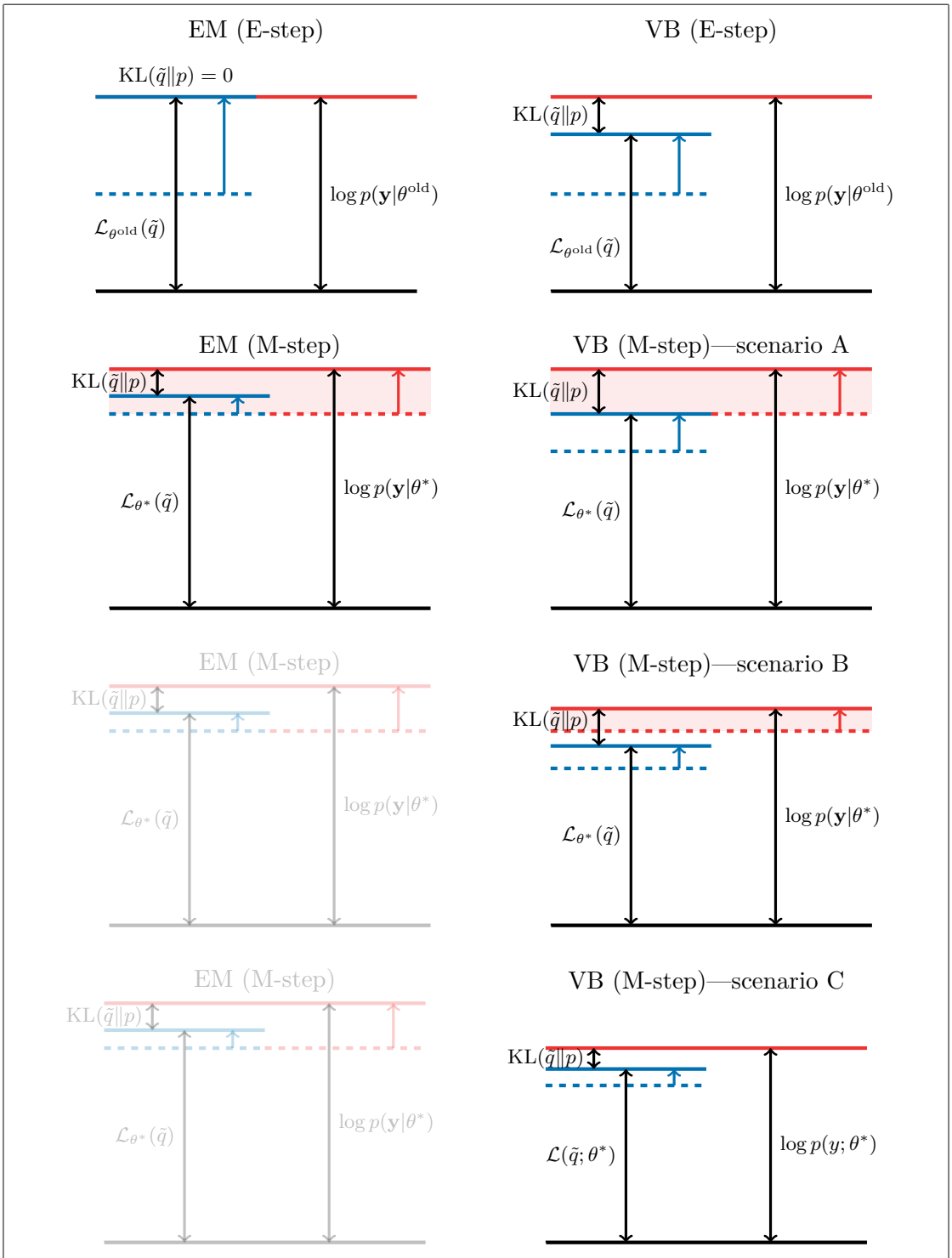

Figure 5.3: Illustration of EM vs VB-EM. Whereas the EM guarantees an increase in log-likelihood value (red shaded region), the VB-EM does not.

the independence of the observations $i = 1, \dots, n$, the complete data log-likelihood is

$$\begin{aligned}
&\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}) \\
&\quad = \sum_{i=1}^{n} \Big\{ \log p(y_i | \mathbf{y}_{i\cdot}^*) + \log p(\mathbf{y}_{i\cdot}^* | \mathbf{w}_{i\cdot}) + \log p(\mathbf{w}_{i\cdot}) \Big\} \\
&\quad = -\frac{1}{2} \sum_{i=1}^{n} \mathbb{1}[y_{ij}^* = \max_k y_{ik}^*] \Big[(\mathbf{y}_{i\cdot}^* - \boldsymbol{\alpha} - \mathbf{w}_{i\cdot}^\top \mathbf{h}_\eta(x_i))^\top \boldsymbol{\Psi} (\mathbf{y}_{i\cdot}^* - \boldsymbol{\alpha} - \mathbf{w}_{i\cdot}^\top \mathbf{h}_\eta(x_i)) \\
&\qquad\qquad + \mathbf{w}_{i\cdot}^\top \boldsymbol{\Psi}^{-1} \mathbf{w}_{i\cdot} \Big] + \text{const.}
\end{aligned}$$

which looks like the complete data log-likelihood seen previously in (4.9), except that here, together with the $\mathbf{w}_{i\cdot}$'s, the $\mathbf{y}_{i\cdot}^*$'s are never observed.

For the E-step, it is of interest to determine the posterior density $p(\mathbf{y}^*, \mathbf{w}|\mathbf{y}) = p(\mathbf{y}^*|\mathbf{w}, \mathbf{y}) p(\mathbf{w}|\mathbf{y})$, which apparently is hard to obtain. We can go as far as determining that the full conditional of the latent propensities is multivariate subject to a conical truncation $\mathcal{C}_j = \{y_{ij}^* > y_{ik}^* \,|\, \forall k \neq j\}$, i.e. $\mathbf{y}_{i\cdot}^* | \mathbf{w}_{i\cdot}, \{y_i = j\} \overset{\text{iid}}{\sim} {}^{\text{t}}\text{N}_m(\boldsymbol{\alpha} + \mathbf{w}_{i\cdot}^\top \mathbf{h}_\eta(x_i), \boldsymbol{\Psi}^{-1}, \mathcal{C}_j)$, for each $i = 1, \dots, n$, and that $\text{vec}\, \mathbf{w}|\mathbf{y}^* \sim \text{N}(\tilde{\mathbf{w}}, \tilde{\mathbf{V}}_w)$ is found to be similar to the distribution in **??**. To obtain the first and second posterior moments for the I-prior random effects, we can use the law of total expectations:

$$\text{E}[\text{vec}\, \mathbf{w}|\mathbf{y}] = \text{E}_{\mathbf{y}^*} \left[\text{E}[\text{vec}\, \mathbf{w}|\mathbf{y}^*] \middle| \mathbf{y} \right] =: \hat{\mathbf{w}}$$

and

$$\text{E}[\text{vec}\, \mathbf{w} (\text{vec}\, \mathbf{w})^\top | \mathbf{y}] = \text{E}_{\mathbf{y}^*} \left[\text{E}[\text{vec}\, \mathbf{w} (\text{vec}\, \mathbf{w})^\top | \mathbf{y}^*] \middle| \mathbf{y} \right] =: \hat{\mathbf{W}},$$

but this requires $p(\mathbf{y}^*|\mathbf{y})$ which does not come by easily. A similar problem has been faced by Chan and Kuk (1997), who analysed binary linear probit models with random effects. The authors ultimately resort to Monte Carlo sampling within an EM framework to overcome the difficult distributions of interest.

Suppose that, instead of the true posterior distribution $p(\mathbf{y}^*, \mathbf{w}|\mathbf{y})$ being used, a mean-field variational approximation $q(\mathbf{y}^*, \mathbf{w}) = q(\mathbf{y}^*) q(\mathbf{w})$ is used instead. As we know from Section X, $q(\mathbf{y}^*)$ is a truncated multivariate normal distribution, and $q(\mathbf{w})$ is multivariate normal, whose means and second moments can be computed with some effort. Let

$\bar{\mathbf{y}}^* = \mathbf{y}^* - \mathbf{1}_n \boldsymbol{\alpha}^\top$. The (approximate) E-step then entails computing

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

In the M-step, this is maximised with respect to θ. This is the VB-EM algorithm described in Section 5.1.2. As per the discussion in Section X, this alleviates the problem of non-conjugacy of the complete conditional for $\boldsymbol{\Psi}$. One downside to VB-EM is that it is not entirely certain how one could obtain standard errors for the parameters, other than by bootstrapping, which for the I-probit model, is likely to be computationally intensive.

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.

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.

chan1997maximum Chan, Jennifer SK and Anthony YC Kuk (1997). "Maximum likelihood estimation for probit-linear mixed models with correlated random effects". In: *Biometrics*, pp. 86–97.

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.

jamil2017 Jamil, Haziq and Wicher Bergsma (2017). "iprior: An R Package for Regression Modelling using I-priors". In: *Manuscript in submission*.

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.