Haziq Jamil
Department of Statistics
London School of Economics and Political Science
PhD thesis: 'Regression modelling using priors depending on Fisher information covariance kernels'
11 October 2018 (`v1.1@3045d93`)

Chapter 1

Preceding chapters

Supplementary S1

Basic estimation concepts

Statistics concerns what can be learned from data (Davison, 2003). A statistical model comprises of a probabilistic component which drives the data generative process, in addition to a systematic or deterministic component, which sets it apart from pure mathematical models. Real-valued observations $\mathbf{y} := \{y_1, \ldots, y_n\}$ are treated as realisations from an assumed probability distribution with parameters $\theta = (\theta_1, \ldots, \theta_p)^\top \in \Theta \subseteq \mathbb{R}^p$. The crux of statistical inference is to estimate θ given the observed values, so that this optimised value may be used in the model to make deductions. We describe the *frequentist* and *Bayesian* paradigms for parameter estimation.

S1.1 Maximum likelihood estimation

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 which provides the largest likelihood value,

$$\hat{\theta}_{\text{ML}} = \arg\max_{\theta} p(\mathbf{y}|\theta). \tag{S1.1}$$

The value $\hat{\theta}_{\text{ML}}$ is referred to as the *maximum likelihood estimate* for θ. For convenience, 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.

When ML estimates are unable to be found in closed-form, the maximisation problem of (S1.1) requires iterative, numerical methods to find the maximum. These methods are often *gradient based*, i.e. algorithms that make use of the gradient of the objective function to be optimised. Examples include Newton's method, Fisher's scoring, quasi-Newton methods, gradient descent, and conjugate gradient methods. As the name suggests, these methods require evaluation of gradients or approximate gradients, and in some cases, the Hessian. Depending on the situation, gradients or Hessians can be expensive or inconvenient to compute or approximate. In cases of multi-modality of the objective function, the algorithms can potentially converge to a local optima, as it is known that the algorithms are quite sensitive to starting locations.

Besides invariance, the ML estimate comes with the attractive limiting property $\sqrt{n}(\hat{\theta}_{\text{ML}} - \theta_{\text{true}}) \xrightarrow{\text{dist.}} \text{N}_p\left(0, \mathcal{I}(\theta)^{-1}\right)$ (Casella and R. L. Berger, 2002) as sample size $n \to \infty$, where $\mathcal{I}(\theta)$ is the Fisher information for θ. Other asymptotic properties of the ML estimate include consistency, i.e. $\text{P}(\|\hat{\theta}_{\text{ML}} - \theta_{\text{true}}\| > \epsilon) \xrightarrow{\text{prob.}} 0$ for any $\epsilon > 0$, and efficiency, i.e. it achieves the Cramér-Rao lower bound $\text{Var}(\theta_{\text{ML}}) \geq \mathcal{I}(\theta)^{-1}$.

As the likelihood measures the plausibility of a parameter value given the data, it can be used to compare two competing models. Let $\Theta_0 = \{\theta \,|\, \theta_{d+1} = \theta_{d+1,0}, \ldots, \theta_p = \theta_{p,0}\}$ be the set of parameters with restrictions on the last d components of θ. The *likelihood ratio test* statistic for testing the null hypothesis $\text{H}_0 : \theta \in \Theta_0$ against the alternative $\text{H}_1 : \theta \notin \Theta_0$ is

$$\lambda = -2 \log \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)} = -2\big(\log L(\hat{\theta}_0) - \log L(\hat{\theta})\big), \tag{S1.2}$$

where $\hat{\theta}_0 = \arg\max_{\theta \in \Theta_0} \log p(\mathbf{y}|\theta)$. Wilks' theorem states that λ has an asymptotic chi-squared distribution with degrees of freedom equal to the number of restrictions imposed (or rather, the difference in dimensionality of Θ and Θ_0). This gives a convenient way of comparing nested models.

As a remark, models with more parameters will always have higher, or similar, log-likelihood, than models with fewer parameters, because the model has a better ability to fit the data with more free parameters. In a linear regression setting, this relates to overfitting: a linear model with as many explanatory variables as there are data points ($n = p$) will extrapolate every point in the data set. Overfitting is an oft cited problem of maximum likelihood.

S1.2 Bayesian estimation

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. Broadly speaking, priors, and hence Bayesian analysis (Kadane, 2011; Robert, 2007), 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, objective 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*

$$p(\theta|\mathbf{y}) \propto \overbrace{p(\mathbf{y}|\theta)}^{\text{likelihood}} \times \overbrace{p(\theta)}^{\text{prior}}, \tag{S1.3}$$

rather than through a single point estimate such as the ML estimate in the frequentist case. The posterior density encapsulates the uncertainty surrounding the parameters θ after observing the data $\mathbf{y}$. The *posterior mean*

$$\tilde{\theta} = \int \theta p(\theta|\mathbf{y}) \, \mathrm{d}\theta \tag{S1.4}$$

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, 1985, Sec. 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,

$$\hat{\mathrm{E}}(\theta|\mathbf{y}) = \frac{1}{T} \sum_{i=1}^{T} \theta^{(t)} \tag{S1.5}$$

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.

The normalising constant is the marginal likelihood over the distribution of the parameters, $p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta) \, \mathrm{d}\theta$. The quantity $p(\mathbf{y})$ is also known as the *model evidence*,

or simply, *evidence*. As its name suggests, model evidence is used as a measure of how much support there is for a particular model. As such, it is used as a basis for model comparison. Let $p(\mathbf{y}|M_0)$ and $p(\mathbf{y}|M_1)$ be the model evidence for two competing models M_0 and M_1 respectively. Define the *Bayes factor* for comparing model M_0 against an alternative model M_1 as

$$\mathrm{BF}(M_0, M_1) = \frac{p(\mathbf{y}|M_0)}{p(\mathbf{y}|M_1)}. \tag{S1.6}$$

Values of $\mathrm{BF}(M_0, M_1) < 1$ would suggest that the data provides more evidence for model M_1 over M_0.

Note that the model evidence is free of θ because all of 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. Thus, model comparison using Bayes factors differs from the frequentist likelihood ratio comparison in that it does not depend on any one particular set of values for the parameters.

S1.3 Maximum a posteriori estimation

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

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\mathbf{y}|\theta)p(\theta), \tag{S1.7}$$

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 estimate with no apparent way of quantifying its uncertainty. Furthermore, unlike ML estimators, MAP estimators are not 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

$$p_\xi(\xi) = p_\theta\big(g^{-1}(\xi)\big) \left| \frac{\mathrm{d}}{\mathrm{d}\xi} g^{-1}(\xi) \right|. \tag{S1.8}$$

The second term in (S1.8) 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 (S1.7).

S1.4 Empirical Bayes

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

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta|\eta)}{p(\mathbf{y}|\eta)} \tag{S1.9}$$

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

$$\hat{\eta} = \arg\max_{\eta} p(\mathbf{y}|\eta) \tag{S1.10}$$

in place of η in the equation of (S1.9). This procedure is also coined *maximum likelihood type-II* (Bishop, 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 hyperprior for η could be parameterised by a hyper-hyperparameter, 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 (Beal, 2003).

Supplementary S2

The EM algorithm

Often times, there are unobserved, random variables $\mathbf{w} = \{w_1, \ldots, w_n\}$ that are assumed to make up the data generative process, prescribed in the statistical model through the *joint pdf* $p(\mathbf{y}, \mathbf{w}|\theta)$. Examples of models that include latent variables are plentiful: Gaussian mixture models, latent class analysis, factor models, random coefficient models, and so on. In order to obtain maximum likelihood (ML) estimates through a direct maximisation of the likelihood, it is necessary to first marginalise out the latent variables,

$$p(\mathbf{y}|\theta) = \int \overbrace{p(\mathbf{y}|\mathbf{w}, \theta)p(\mathbf{w}|\theta)}^{p(\mathbf{y},\mathbf{w}|\theta)} \mathrm{d}\mathbf{w}, \tag{S2.1}$$

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 $\mathbf{w}$.

Direct maximisation of the marginal (log-)likelihood might not be favourable due to intractability in obtaining ML solutions. The form of the marginal likelihood might might not be conducive for closed-form estimates to be found, necessitating the use of numerical, gradient-based methods which is subject to its own undesirable quirks. Moreover, when the evaluation of the (log-)likelihood, gradient and/or Hessian are expensive to compute, then numerical methods are burdensome to execute.

It is usually the case that if the latent variables $\mathbf{w}$ were somehow known, estimation would be made simpler. That is, the solution to $\arg\max_\theta \log p(\mathbf{y}, \mathbf{w}|\theta)$ can be obtained in a simple manner. The expectation-maximisation algorithm (Dempster et al., 1977), commonly known as the EM algorithm, is an iterative procedure which exploits the fact that the so-called *complete data likelihood* is easier to work with. Correspondingly, in EM terminology, the marginal likelihood is referred to as the *incomplete data likelihood*.

We describe a derivation of both a general EM algorithm and an EM algorithm for models whose data generative pdf belongs to an exponential family of pdfs. Interestingly, the EM algorithm can be modified to obtain maximum a posteriori estimates or penalised log-likelihood solutions. As a note, the EM algorithm is not an algorithm per se, in that it does not provide exact instructions as to what the E- and M-steps should comprise of. Rather, it is a generic device to obtain parameter estimates (McLachlan and Krishnan, 2007).

S2.1 Derivation of the EM algorithm

For want of an iterative procedure to obtain maximum likelihood estimates, we seek a solution to

$$\arg\max_{\theta}\{L(\theta|\mathbf{y}) - L(\theta^{(t)}|\mathbf{y}) \geq 0\}, \tag{S2.2}$$

where the solution to (S2.2) yields an improvement to the current t'th iteration of the log-likelihood value $L(\theta^{(t)}|\mathbf{y})$. Note that the objective function in (S2.2) forms an upper bound for the quantity $Q(\theta|\theta^{(t)})$, as shown below:

$$\begin{aligned}
L(\theta|\mathbf{y}) - L(\theta^{(t)}|\mathbf{y}) &= \log \int p(\mathbf{y}|\mathbf{w},\theta)p(\mathbf{w}|\theta)\,\frac{p(\mathbf{w}|\mathbf{y},\theta^{(t)})}{p(\mathbf{w}|\mathbf{y},\theta^{(t)})}\,\mathrm{d}\mathbf{w} - \log p(\mathbf{y}|\theta^{(t)}) \\
&\geq \int p(\mathbf{w}|\mathbf{y},\theta^{(t)}) \log \frac{p(\mathbf{y}|\mathbf{w},\theta)p(\mathbf{w}|\theta)}{p(\mathbf{w}|\mathbf{y},\theta^{(t)})}\,\mathrm{d}\mathbf{w} \quad \text{(Jensen's inequality)} \\
&\quad - \log p(\mathbf{y}|\theta^{(t)}) \int p(\mathbf{w}|\mathbf{y},\theta^{(t)})\,\mathrm{d}\mathbf{w} \\
&= \int p(\mathbf{w}|\mathbf{y},\theta^{(t)}) \log \frac{p(\mathbf{y}|\mathbf{w},\theta)p(\mathbf{w}|\theta)}{p(\mathbf{w}|\mathbf{y},\theta^{(t)})p(\mathbf{y}|\theta^{(t)})}\,\mathrm{d}\mathbf{w} \\
&=: Q(\theta|\theta^{(t)}).
\end{aligned}$$

Evidently, to maximise $L(\theta|\mathbf{y})$, we can't do any worse than maximising $Q(\theta|\theta^{(t)})$ in θ. Denote by $\theta^{(t+1)}$ as the maximiser of $Q(\theta|\theta^{(t)})$. Then,

$$\begin{aligned}
\theta^{(t+1)} &= \arg\max_{\theta} \int p(\mathbf{w}|\mathbf{y},\theta^{(t)}) \log \frac{p(\mathbf{y}|\mathbf{w},\theta)p(\mathbf{w}|\theta)}{p(\mathbf{w}|\mathbf{y},\theta^{(t)})p(\mathbf{y}|\theta^{(t)})}\,\mathrm{d}\mathbf{w} \\
&= \arg\max_{\theta} \int p(\mathbf{w}|\mathbf{y},\theta^{(t)}) \log p(\mathbf{y}|\mathbf{w},\theta)p(\mathbf{w}|\theta)\,\mathrm{d}\mathbf{w} \\
&= \arg\max_{\theta} \int p(\mathbf{w}|\mathbf{y},\theta^{(t)}) \log p(\mathbf{y},\mathbf{w}|\theta)\,\mathrm{d}\mathbf{w} \\
&= \arg\max_{\theta} \mathrm{E}_{\mathbf{w}}\left[\log p(\mathbf{w},\mathbf{y}|\theta)|\mathbf{y},\theta^{(t)}\right]
\end{aligned}$$

We arrive at an iterative procedure summarised succintly as the following:

Algorithm S1 EM algorithm

1: **initialise** $\boldsymbol{\theta}^{(0)}$ and $t \leftarrow 0$
2: **while** not converged **do**
3: E-step: compute $Q(\theta|\theta^{(t)}) = \mathrm{E}_{\mathbf{w}}\left[\log p(\mathbf{w}, \mathbf{y}|\theta)|\mathbf{y}, \theta^{(t)}\right]$
4: M-step: $\theta^{(t+1)} \leftarrow \arg\max_{\theta} Q(\theta|\theta^{(t)})$
5: $t \leftarrow t + 1$
6: **end while**

Notice that the log-likelihood function satisfies

$$L(\theta|\mathbf{y}) \geq L(\theta^{(t)}|\mathbf{y}) + Q(\theta|\theta^{(t)}), \tag{S2.3}$$

for which equality is achieved when $\theta = \theta^{(t)}$, since

$$\begin{aligned} Q(\theta^{(t)}|\theta^{(t)}) &= \int p(\mathbf{w}|\mathbf{y}, \theta^{(t)}) \log \frac{p(\mathbf{y}|\mathbf{w}, \theta^{(t)}) p(\mathbf{w}|\theta^{(t)})}{p(\mathbf{w}|\mathbf{y}, \theta^{(t)}) p(\mathbf{y}|\theta^{(t)})} \, \mathrm{d}\mathbf{w} \\ &= \int p(\mathbf{w}|\mathbf{y}, \theta^{(t)}) \log \frac{p(\mathbf{y}, \mathbf{w}|\theta^{(t)})}{p(\mathbf{y}, \mathbf{w}|\theta^{(t)})} \, \mathrm{d}\mathbf{w} \\ &= 0. \end{aligned}$$

This implies that the EM algorithm improves the log-likelihood values at each iteration, since

$$L(\theta^{(t+1)}|\mathbf{y}) - L(\theta^{(t)}|\mathbf{y}) \geq Q(\theta^{(t+1)}|\theta^{(t)}) \geq 0$$

and $Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)}) = 0$ since $\theta^{(t+1)}$ maximises $Q(\cdot|\theta^{(t)})$.

The expectation in the E-step involves the conditional pdf $p(\mathbf{w}|\mathbf{y}, \theta^{(t)})$. Viewed through a Bayesian lens, this is the posterior density of the latent variables using the t'th iteration parameter values. The success of the E-step is predicated on the availability of the conditional pdf for the expectation. If not, approximations to the E-step can be explored, for example using Monte Carlo methods (Wei and Tanner, 1990) or a variational approximation (Beal, 2003).

The solution to the M-step usually, but not always, exists in closed form. Maximising the Q function over all possible values of θ may not be feasible (McLachlan and Krishnan, 2007). In such situations, the generalised EM algorithm (as defined by Dempster et al., 1977) requires only that $\theta^{(t+1)}$ be chosen in a way that

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)}).$$

That is, $\theta^{(t+1)}$ is chosen so as to increase the value of the Q function at its current parameter value. As seen in the argument above, this requirement is sufficient for a guaranteed increase in the log-likelihood function at each iteration.

S2.2 Exponential family EM algorithm

Consider the density function $p(\cdot|\boldsymbol{\theta})$ of the complete data $\mathbf{z} = \{\mathbf{y}, \mathbf{w}\}$, which depends on parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)^\top \in \Theta \subseteq \mathbb{R}^s$, belonging to an exponential family of distributions. This density takes the form $p(\mathbf{z}|\boldsymbol{\theta}) = B(\mathbf{z}) \exp\big(\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \mathbf{T}(\mathbf{z})\rangle - A(\boldsymbol{\theta})\big)$, where $\boldsymbol{\eta} : \mathbb{R}^s \mapsto \mathbb{R}$ is a link function, $\mathbf{T}(\mathbf{z}) = \big(T_1(\mathbf{z}), \ldots, T_s(\mathbf{z})\big)^\top \in \mathbb{R}^s$ are the sufficient statistics of the distribution, and $\langle \cdot, \cdot \rangle$ is the usual Euclidean dot product. It is often easier to work in the *natural parameterisation* of the exponential family distribution

$$p(\mathbf{z}|\boldsymbol{\eta}) = B(\mathbf{z}) \exp\big(\langle \boldsymbol{\eta}, \mathbf{T}(\mathbf{z})\rangle - A^*(\boldsymbol{\eta})\big) \tag{S2.4}$$

by defining $\boldsymbol{\eta} := \big(\eta_1(\boldsymbol{\theta}), \ldots, \eta_r(\boldsymbol{\theta})\big) \in \mathcal{E}$, and $\exp A^*(\boldsymbol{\eta}) = \int B(\mathbf{z}) \exp\langle \boldsymbol{\eta}, \mathbf{T}(\mathbf{z})\rangle \, \mathrm{d}\mathbf{z}$ to ensure the density function normalises to one. As an aside, the set $\mathcal{E} := \big\{\boldsymbol{\eta} = (\eta_1, \ldots, \eta_s) \,|\, \int \exp A^*(\boldsymbol{\eta}) < \infty\big\}$ is called the *natural parameter space*. If $\dim \mathcal{E} = r < s = \dim \Theta$, then the the pdf belongs to the *curved exponential family* of distributions. If $\dim \mathcal{E} = r = s = \dim \Theta$, then the family is a *full exponential family*.

Assuming the latent $\mathbf{w}$ variables are observed and working with the natural parameterisation, then the complete maximum likelihood (ML) estimate for $\boldsymbol{\eta}$ is obtained by solving

$$\frac{\partial}{\partial \boldsymbol{\eta}} \log p(\mathbf{z}|\boldsymbol{\eta}) = \mathbf{T}(\mathbf{z}) - \frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta}) = 0. \tag{S2.5}$$

Of course, the variable $\mathbf{w}$ are never observed, so the ML estimate for $\boldsymbol{\eta}$ can only be informed from what is observed. Let $p(\mathbf{y}|\boldsymbol{\eta}) = \int p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \, \mathrm{d}\mathbf{w}$ represent the marginal density of the observations $\mathbf{y}$. Now, the ML estimate for $\boldsymbol{\eta}$ is obtained by solving

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\eta}} \log p(\mathbf{y}|\boldsymbol{\eta}) &= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \cdot \frac{\partial}{\partial \boldsymbol{\eta}} p(\mathbf{y}|\boldsymbol{\eta}) \\
&= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \cdot \frac{\partial}{\partial \boldsymbol{\eta}} \left(\int p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \, \mathrm{d}\mathbf{w} \right) \\
&= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \int \left(\frac{\partial}{\partial \boldsymbol{\eta}} p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \right) \mathrm{d}\mathbf{w} \\
&= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \int \left(p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \frac{\partial}{\partial \boldsymbol{\eta}} \log p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \right) \mathrm{d}\mathbf{w} \\
&= \int \left(\mathbf{T}(\mathbf{y}, \mathbf{w}) - \frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta}) \right) p(\mathbf{w}|\mathbf{y}, \boldsymbol{\eta}) \, \mathrm{d}\mathbf{w} \\
&= \mathrm{E}_{\mathbf{w}} \left[\mathbf{T}(\mathbf{y}, \mathbf{w}) | \mathbf{y} \right] - \frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta})
\end{aligned} \tag{S2.6}$$

equated to zero. Note that we are allowed to change the order of integration and differentation provided the integrand is continuously differentiable. So the only difference

between the first order condition of (S2.5) and that of (S2.6) is that the sufficient statistics involving the unknown $\mathbf{w}$ are replaced by their conditional or posterior expectations.

A useful identity to know is that $\frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta}) = \mathrm{E}_{\mathbf{z}}[\mathbf{T}(\mathbf{z})]$ (Casella and R. L. Berger, 2002, Thm. 3.4.2 & Exer. 3.32(a)), which can be expressed in terms of the original parameters $\boldsymbol{\theta}$. As a consequence, solving for the ML estimate for $\boldsymbol{\theta}$ from the FOC equations (S2.6) is possible without having to deal with the derivative of A^* with respect to the natural parameters. Having said this, an analytical solution in $\boldsymbol{\theta}$ may not exist, because the relationship of $\boldsymbol{\theta}$ could be implicit in the set of equations $\mathrm{E}_{\mathbf{w}}\left[\mathbf{T}(\mathbf{w}, \mathbf{y})|\mathbf{y}, \boldsymbol{\theta}\right] = \mathrm{E}_{\mathbf{y},\mathbf{w}}\left[\mathbf{T}(\mathbf{y}, \mathbf{w})|\boldsymbol{\theta}\right]$. One way around this is to employ an iterative procedure, as detailed in Algorithm S2.

Algorithm S2 Exponential family EM

1: **initialise** $\boldsymbol{\theta}^{(0)}$ and $t \leftarrow 0$
2: **while** not converged **do**
3: E-step: $\tilde{\mathbf{T}}^{(t+1)}(\mathbf{y}, \mathbf{w}) \leftarrow \mathrm{E}_{\mathbf{w}}\left[\mathbf{T}(\mathbf{w}, \mathbf{y})|\mathbf{y}, \boldsymbol{\theta}^{(t)}\right]$
4: M-step: $\boldsymbol{\theta}^{(t+1)} \leftarrow$ solution to $\tilde{\mathbf{T}}^{(t+1)}(\mathbf{y}, \mathbf{w}) = \mathrm{E}_{\mathbf{y},\mathbf{w}}\left[\mathbf{T}(\mathbf{y}, \mathbf{w})|\boldsymbol{\theta}\right]$
5: $t \leftarrow t + 1$
6: **end while**

To see how Algorithm S2 motivates the EM algorithm, consider the following argument. Recall that for the EM algorithm, the function $Q_t(\boldsymbol{\eta}) = \mathrm{E}_{\mathbf{w}}[\log p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta})|\mathbf{y}, \boldsymbol{\eta}^{(t)}]$ is maximised at each iteration t. For exponential families of the form (S2.4), the Q_t function turns out to be

$$Q_t(\boldsymbol{\eta}) = \mathrm{E}_{\mathbf{w}}\left[\langle \boldsymbol{\eta}, \mathbf{T}(\mathbf{z})\rangle|\mathbf{y}, \boldsymbol{\eta}^{(t)}\right] - A^*(\boldsymbol{\eta}) + \log B(\mathbf{z}),$$

and this is maximised at the value of $\boldsymbol{\eta}$ satisfying

$$\frac{\partial}{\partial \boldsymbol{\eta}} Q_t(\boldsymbol{\eta}) = \mathrm{E}_{\mathbf{w}}\left[\mathbf{T}(\mathbf{y}, \mathbf{w})|\mathbf{y}, \boldsymbol{\eta}^{(t)}\right] - \frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta}) = 0,$$

a similar condition to (S2.6) when obtaining ML estimate of $\boldsymbol{\eta}$. Thus, Q_t is maximised by the solution to line 4 in Algorithm S2.

S2.3 Bayesian EM algorithm

A simple modification of the EM algorithm can be done to obtain maximum a posteriori estimates, or maximum penalised likelihood estimates. Under a Bayesian framework, a prior is assigned on the model parameters, $\theta \sim p(\theta)$. Recall that the MAP estimate is obtained as the maximiser of the log-density $\log p(\mathbf{y}|\theta) + \log p(\theta)$.

The EM algorithm works as before, but replaces the E-step with

$$\mathrm{E}_{\mathbf{w}}\left[\log p(\mathbf{w},\mathbf{y}|\theta)+\log p(\theta)|\mathbf{y},\theta^{(t)}\right]=Q(\theta|\theta^{(t)})+\log p(\theta) \tag{S2.7}$$

since $\log p(\theta)$ has no terms involving the latent variables $\mathbf{w}$. The M-step now maximises (S2.7) with respect to θ, which includes the log prior density (or a penalty term). It would seem that the regular EM algorithm maximises (S2.7) such that $p(\theta) \propto$ const. is a diffuse prior for θ. Beal and Ghahramani (2003) discuss a more Bayesian extension of EM, in which the output of the so-called *variational Bayes EM* algorithm are (approximate) posterior distributions of the parameters, rather than MAP estimates discussed here.

Supplementary S3

Hamiltonian Monte Carlo

Hamiltonian Monte Carlo had its beginnings in statistical physics, with the 1987 paper by Duane et al., using what they called 'Hybrid Monte Carlo' in lattice models of quantum theory. Their work merged the approaches of molecular dynamics and Markov chain Monte Carlo methods. As an interesting side note, their method abbreviates also to 'HMC', but throughout the statistical literature, it is more commonly referred to by its more descriptive name Hamiltonian Monte Carlo. Incidentally, the use of HMC started with applications to neural networks as early as 1996 (see Neal, 2011 for an excellent review of the subject matter). It was not until 2011 when active development of the method, and in particular, software for for statistical applications began. The Stan initiative (Carpenter et al., 2017) began in response to difficulties faced when performing full Bayesian inference on multilevel generalised linear models. These difficulties mainly involved poor efficiency in usual MCMC samplers, particularly due to high autocorrelations in the posterior chains, which meant that many chains and many iterations were required to get an adequate sample. It was a case of exhausting all possible algorithmic remedies for existing samplers (Gibbs samplers, Metropolis samplers, etc.), and realising that fundamentally not much improvement can be had unless a novel sampling technique was discovered.

The basic idea behind HMC is to use Hamiltonian dynamics to propose new states in the posterior sampling, rather than relying on random walks. If one were to understand and use the geometry of the posterior density to one's benefit, then it should be possible to generate new proposal states with high probabilities of acceptance and move far away from the current state. Hamiltonian dynamics, like classical Newtonian mechanics, provides a framework for modelling the motion of a body in space across time t. Additionally, Hamiltonian dynamics concatenates the position vector x with its momentum z, and the motion of x in d-dimensional space is then described through

Hamilton's equations

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\partial H}{\partial z} \quad \text{and} \quad \frac{\mathrm{d}z}{\mathrm{d}t} = -\frac{\partial H}{\partial x}, \tag{S3.1}$$

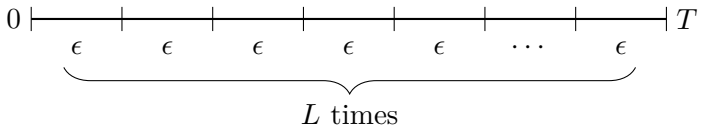
where $H = H(x, z)$ is called the Hamiltonian of the system. The Hamiltonian is an operator which encapsulates the total energy of the system. In a closed system, one can express the sum of operators corresponding to the kinetic energy $K(p)$ and the potential energy $U(z)$ of the system

$$H(x, z) = K(z) + U(x). \tag{S3.2}$$

Substituing (S3.2) into (S3.1), we get the system of partial differential equations (PDEs)

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\partial}{\partial z}K(z) \quad \text{and} \quad \frac{\mathrm{d}z}{\mathrm{d}t} = -\frac{\partial}{\partial x}U(x). \tag{S3.3}$$

To describe the evolution of $\big(x(t), z(t)\big)$ from time t to $t+T$, it is necessary to discretise time, and split $T = L\epsilon$. The quantity L is known as the number of *leapfrogs*, and ϵ the *step size*.



The system of PDEs is solved using Euler's method, or the more commonly used leapfrog integration, which is a three-step process:

1. **Half-step momentum.** $z(t + \epsilon/2) = z(t) - \frac{\epsilon}{2}\frac{\partial}{\partial x}U\big(x(t)\big)$
2. **Full-step position.** $x(t + \epsilon) = x(t) + \epsilon\frac{\partial}{\partial z}K\big(z(t + \epsilon/2)\big)$
3. **Half-step momentum.** $z(t + \epsilon) = z(t + \epsilon/2) = z(t) - \frac{\epsilon}{2}\frac{\partial}{\partial x}U\big(x(t)\big)$

in which steps 1–3 are repeated L times.

Having knowing the formula for how particles move in space, we can use this information to treat random points drawn from some probability density as "particles". Randomness of position and momentum are prescribed through probability densities on each. Given some energy function $E(\theta)$ over states θ, the *canonical distribution* of the states θ (otherwise known as the *canonical ensemble*) is given by the probability density function

$$p(\theta) \propto \exp\left(-\frac{E(\theta)}{k\tau}\right),$$

where k is Boltzmann's constant, τ is the absolute temperature of the system. The Hamiltonian is one such energy function over states (x, z). By replacing $E(\theta)$ by (S3.2) in the pdf above, we realise that the distribution for x and z are independent. The system can be manipulated such that $k\tau = 1$—in any case, these are constants which can be absorbed into one of the terms in the pdf anyway.

Using a *quadratic kinetic energy* function $K(z) = z^\top M^{-1} z/2$[1], we find that the probability density function for z is

$$p(z) \propto \exp\left(-\frac{1}{2} z^\top M^{-1} z\right),$$

implying $z \sim \mathrm{N}_d(0, M)$. Here, $M = \mathrm{diag}(m_1, \ldots, m_d)$ is called the *mass matrix*, which obviously serves as the variance for the randomly distributed z. As for the potential energy, choose a function such that $U(x) = -\log p(x)$, implying $p(x) \propto \exp\big(-U(x)\big)$. Here, $p(x)$ represents the target density from which we wish to sample, for instance, a posterior density of interest. Thus, to sample variables x from $p(x)$, one artificially introduces momentum variables z and sample jointly instead from $p(x, z) = p(z)p(z)$, and discarding z thereafter. The HMC algorithm is summarised in Algorithm S3.

Algorithm S3 Hamiltonian Monte Carlo

1: **initialise** $x^{(0)}$, $z^{(0)}$ and choose values for L, ϵ and M
2: **for** $t = 1, \ldots, T$ **do**
3: Draw $z \sim \mathrm{N}_d(0, M)$ ▷ Perturb momentum
4: Move $(x^{(t)}, z^{(t)}) \mapsto (x^*, z^*)$ using Hamiltonian dynamics ▷ Proposal state
5: Accept/reject proposal state, i.e. ▷ Metropolis update

$$(x^{(t+1)}, z^{(t+1)}) \leftarrow \begin{cases} (x^*, z^*) & \text{w.p. } \min(1, A) \\ (x^{(t)}, z^{(t)}) & \text{otherwise} \end{cases}$$

where

$$A = \frac{p(x^*, z^*)}{(x^{(t)}, z^{(t)})} = \exp\Big(H(x, z) - H(x^{(t)}, z^{(t)})\Big)$$

6: **end for**
7: **return** Samples $\big\{x^{(1)}, \ldots, x^{(T)}\big\}$

HMC is often times superior to standard Gibbs sampling, for a variety of reasons. For one, conjugacy does not play any role in the efficiency of the HMC sampler, thus freeing the modeller to choose more appropriate and more intuitive prior densities for the parameters of the model. For another, the HMC sampler is designed to incite little autocorrelations between samples, and thus increasing efficiency.

Several drawbacks do exist with the HMC sampler. Firstly, it is impossible to directly sample from discrete distributions $p(x)$. More concretely, HMC requires that the domain of $p(x)$ is continuous and that $\partial \log p(x)/\partial x$ is inexpensive to compute. To work around this, one must reformulate the model by marginalising out the discrete variables, and obtain them back later by separately sampling from their posteriors. Alternatively, a Gibbs sampler specifically for the discrete variables could be augmented with the HMC sampler. The other drawback of HMC is that there are many tuning parameters (leapfrog L, step-size ϵ, mass matrix M, etc.) that is not immediately easy to perfect, at least not to the novice user.

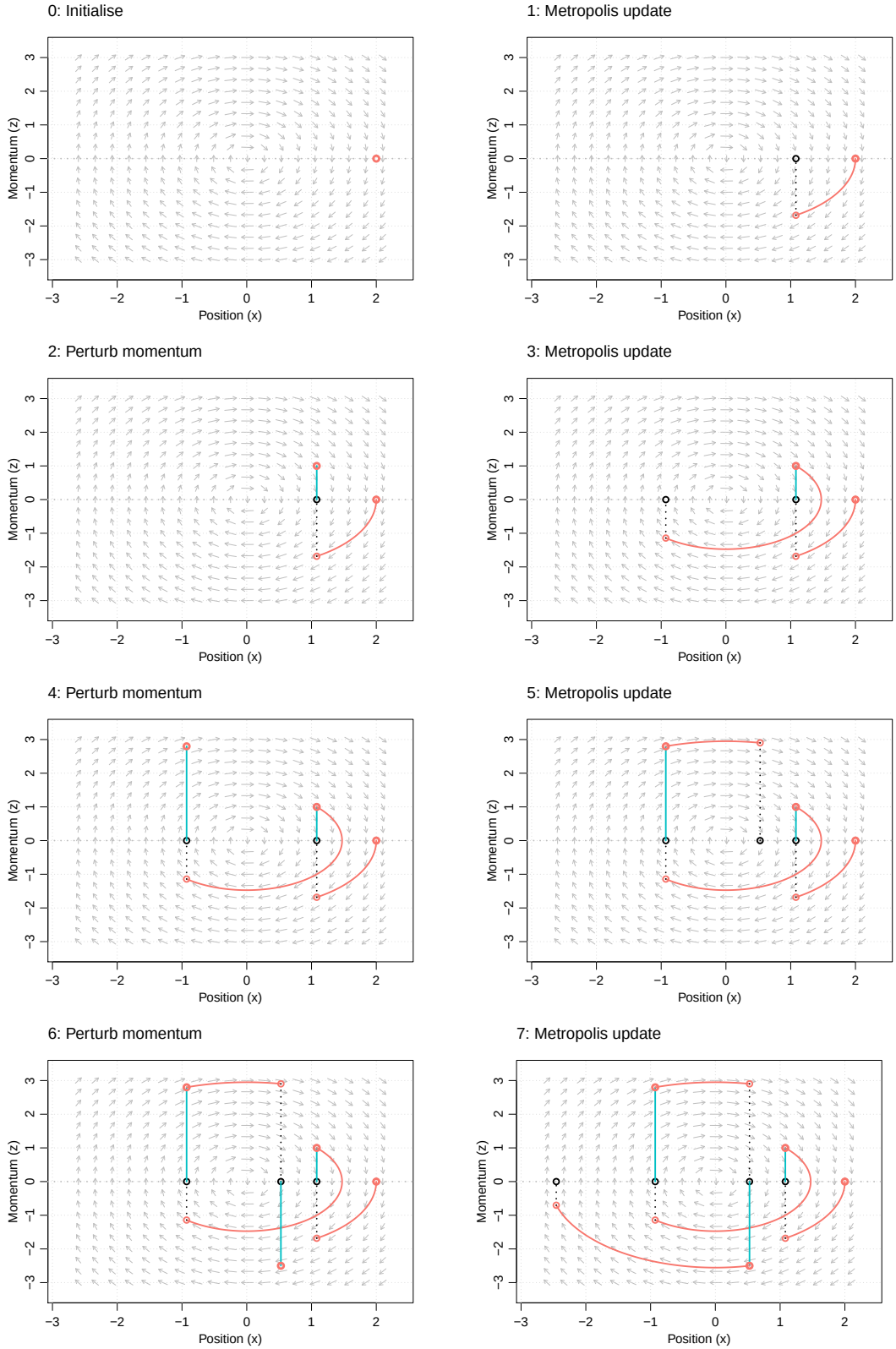


Figure S3.1: A phase diagram schematic of HMC sampling in one-dimension. At step 0, initialise values for momentum and position. At step 1, simulate movement using Hamiltonian dynamics, accept position and discard momentum. At step 2, perturb momentum using a normal density, then repeat.

The implementation of HMC by the programming language Stan, which interfaces many other programming languages including R, Python, MATLAB, Julia, Stata and Mathematica, is a huge step forward in computational Bayesian analysis. Stan takes the liberty of performing all the tuning necessary, and the practitioner is left with simply specifying the model. A vast library of differentiable probability functions are available, with the ability to bring your own code as well. Development is very active and many improvements and optimisations have been made since its inception.

[1]Thinking back to elementary mechanics, this is the familiar $\frac{1}{2}mv^2$ formula for kinetic energy and substituting in the identity $z = mv$, where m is the mass of the object, and v is its velocity.

Supplementary S4

Variational inference

Consider a statistical model parameterised by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ for which we have observations $\mathbf{y} := \{y_1, \ldots, y_n\}$, but also some latent variables $\mathbf{w}$. Typically, in such models, there is a want to to evaluate the integral

$$I = \int p(\mathbf{y}|\mathbf{w})p(\mathbf{w})\,\mathrm{d}\mathbf{w}, \tag{S4.1}$$

Marginalising out the latent variables in (S4.1) is usually a precursor to obtaining a log-likelihood function to be maximised in a frequentist setting, whereby there is an implicit dependence on the model parameters in the evaluation of I. In Bayesian analysis, priors are specified on the model parameters $\theta \sim p(\theta)$. By concatenating the latent variables and model parameters to form $\mathbf{w}$, the I corresponds to the marginal density for $\mathbf{y}$, on which the posterior depends.

In many instances, for one reason or another, evaluation of (S4.1) or is difficult, in which case inference is halted unless a way of overcoming the intractability is found. In this chapter, we discus *variational inference* (VI) as a means of approximating the integral. The literature on variational inference is typically presented in a Bayesian light (Bishop, 2006; Blei et al., 2017; Jordan et al., 1999), and as such, it is commonly known as *variational Bayes* method. The main attraction from a Bayesian point of view is that it provides a deterministic way of obtaining (approximate) posteriors, i.e. it does not involve sampling from posteriors.

Variational inference can be used in conjunction with an EM algorithm, in which the E-step is replaced with a variational E-step. This *variational EM algorithm* is used for maximum likelihood learning, but can modified to obtain maximum a posteriori estimates. In the works of (Beal, 2003; Beal and Ghahramani, 2003), the authors realised that the EM algorithm can be extended easily to obtain posterior densities of the latent variables and parameters if the statistical model is conjugate exponential family. They

refer to this as the *variational Bayes EM algorithm*, but in fact this is really just variational inference in which the algorithm resembles an EM algorithm with clear E- and M-steps.

We first briefly introduce variational methods for approximating the intractable integral, and this is usually considered a fully Bayesian treatment of the model. We then describe variational EM, and provide a comparison of the two methods.

S4.1 A brief introduction to variational inference

The crux of variational inference is this: find a suitably close distribution function $q(\mathbf{w})$ that approximates the true posterior $p(\mathbf{w}|\mathbf{y})$, where closeness here is defined in the Kullback-Leibler divergence sense,

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

Posterior inference is then conducted using $q(\mathbf{w})$ in lieu of $p(\mathbf{w}|\mathbf{y})$. 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 $\mathbf{w}$. One may show that log marginal density, i.e. the log of the intractable integral (S2.1), holds the following bound:

$$\begin{aligned}
\log p(y) &= \log p(\mathbf{y}, \mathbf{w}) - \log p(\mathbf{w}|\mathbf{y}) \quad \text{(Bayes' theorem)} \\
&= \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{w})}{q(\mathbf{w})} - \log \frac{p(\mathbf{w}|\mathbf{y})}{q(\mathbf{w})} \right\} q(\mathbf{w})\,\mathrm{d}\mathbf{w} \quad \text{(expectation both sides)} \\
&= \mathcal{L}(q) + \mathrm{D}_{\mathrm{KL}}(q\|p) \\
&\geq \mathcal{L}(q) \qquad \text{(S4.2)}
\end{aligned}$$

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

$$\begin{aligned}
\mathcal{L}(q) &= \int \log \frac{p(\mathbf{y}, \mathbf{w})}{q(\mathbf{w})} q(\mathbf{w})\,\mathrm{d}\mathbf{w} \\
&= \mathrm{E}_{\mathbf{w}\sim q}[\log p(\mathbf{y}, \mathbf{w})] + H(q), \qquad \text{(S4.3)}
\end{aligned}$$

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

(S4.2) achieves equality if and only if $q(\mathbf{w}) \equiv p(\mathbf{w}|\mathbf{y})$, but of course the true form of the posterior is unknown to us—see Section S4.2 for a discussion. Maximising $\mathcal{L}(q)$ or minimising $\mathrm{D}_{\mathrm{KL}}(q\|p)$ 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 $\mathrm{D}_{\mathrm{KL}}(q||p)$ is impossible to compute, since one does not know the true distribution $p(\mathbf{w}|\mathbf{y})$. 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 $p(\mathbf{w}|\mathbf{y})$ in terms of KL divergence. In this case, the task is to find optimal mean and variance parameters of a normal distribution.

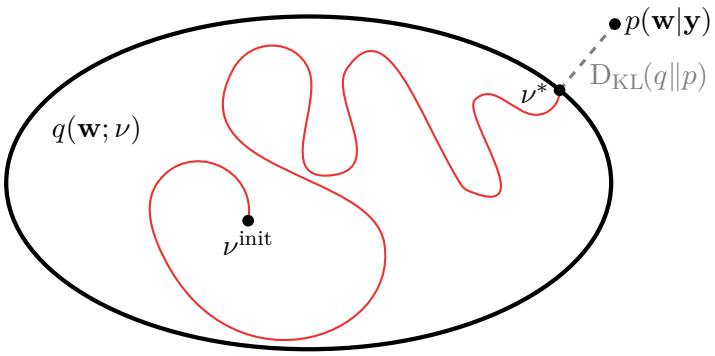


Figure S4.1: Schematic view of variational inference[1]. The aim is to find the closest distribution q (parameterised by a variational parameter ν) to p in terms of KL divergence within the set of variational distributions, represented by the ellipse.

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 $\mathbf{w}$ into M disjoint groups $\mathbf{w} = (w_{[1]}, \ldots, w_{[M]})$. Note that each factor $w_{[k]}$ may be multidimensional. Then, the structure

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

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).

Remark S4.1. The choice of factorisation is completely arbitrary, although forcing a factorisation also induces independence between the factors in the posterior, and this may or may not be suitable for the problem at hand. Landing the correct choice of

[1]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.

factorisation is rather experimental, as the aim is to balance tractability and model misspeficication. In a model with both latent variables and random parameters (in a Bayesian setting), then a good starting point would be to factorise the latent variables and parameters.

Let us denote the distributions which minimise the Kullback-Leibler divergence (maximise the variational lower bound) by the use of tildes. The impact of the mean-field factorisation on the ELBO is inspected:

$$\begin{aligned}\mathcal{L}(q) &= \int \cdots \int \log \frac{p(\mathbf{y}, \mathbf{w})}{\prod_{k=1}^{M} q_k(\mathbf{w})} \prod_{k=1}^{m} \left\{ q_k(w_{[k]}) \, \mathrm{d}w_{[k]} \right\} \\ &= \int \cdots \int \left(\log p(\mathbf{y}, \mathbf{w}) - \sum_{k=1}^{M} \log q_k(\mathbf{w}) \right) \prod_{k=1}^{M} \left\{ q_k(w_{[k]}) \, \mathrm{d}w_{[k]} \right\}\end{aligned}$$

and rearranging slightly for terms involving the j'th component only, we get

$$\begin{aligned}\mathcal{L}(q) = & \int \cdots \int \left(\log p(\mathbf{y}, \mathbf{w}) - \log q_j(w_{[j]}) + \text{const.} \right) q_j(w_{[j]}) \, \mathrm{d}w_{[j]} \prod_{k \neq j} \left\{ q_k(w_{[k]}) \, \mathrm{d}w_{[k]} \right\} \\ = & \int \left(\overbrace{\int \cdots \int \log p(\mathbf{y}, \mathbf{w}) \prod_{k \neq j} \left\{ q_k(w_{[k]}) \, \mathrm{d}w_{[k]} \right\}}^{\log \tilde{p}(\mathbf{y}, w_{[j]}) + \text{const.}} \right) q_j(w_{[j]}) \, \mathrm{d}w_{[j]} \\ & - \int \log q_j(w_{[j]}) q_j(w_{[j]}) \, \mathrm{d}w_{[j]} + \text{const.} \\ = & \; - \mathrm{D}_{\mathrm{KL}}(q_{[j]} \| \tilde{p}) + \text{const.}\end{aligned}$$

The task of maximising $\mathcal{L}$ is then equivalent to maximising $-\mathrm{D}_{\mathrm{KL}}(q_{[j]} \| \tilde{p})$, where $\tilde{p}$ is defined in the overbrace of the second line in the equation above. Thus, for each $w_{[k]}$, $k = 1, \ldots, M$, $\tilde{q}_k$ satisfies

$$\log \tilde{q}_k(w_{[k]}) = \mathrm{E}_{-k}[\log p(\mathbf{y}, \mathbf{w})] + \text{const.} \tag{S4.4}$$

where expectation of the joint log density of $\mathbf{y}$ and $\mathbf{w}$ is taken with respect to all of the unknowns $\mathbf{w}$, except the one currently in consideration $w_{[k]}$, under their respective $\tilde{q}_k$ densities. For further details, refer to Bishop (2006, Eq. 10.9, p. 466).

In practice, rather than an explicit calculation of the normalising constant, one simply needs to inspect (S4.4) 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 $p(w_{[k]} | \mathbf{w}_{-k}, \mathbf{y})$, where $\mathbf{w}_{-k} = \{w_{[i]} \,|\, i \neq k\}$, follows an exponential family distribution

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

Then, from (S4.4),

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

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, Sec. 4) 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 $\mathbf{w}_{-k}$. 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 S4 The CAVI algorithm

1: **initialise** Variational factors $q_k(w_{[k]})$
2: **while** ELBO $\mathcal{L}(q)$ not converged **do**
3: **for** $k = 1, \ldots, M$ **do**
4: $\tilde{q}_k(w_{[k]}) \leftarrow$ const. $\times \exp \mathrm{E}_{-k}\left[\log p(\mathbf{y}, \mathbf{w})\right]$ ▷ from (S4.4)
5: **end for**
6: $\mathcal{L}(q) \leftarrow \mathrm{E}_{\mathbf{w}\sim\prod_k \tilde{q}_k} \log p(\mathbf{y}, \mathbf{w}) + \sum_{k=1}^{m} H\left[q_k(w_{[k]})\right]$ ▷ Update ELBO
7: **end while**
8: **return** $\tilde{q}(\mathbf{w}) = \prod_{k=1}^{M} \tilde{q}_j(w_{[k]})$

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.

S4.2 Variational EM algorithm

Consider again the latent variable setup described in Supplementary Chapter S2, in which the goal is to maximise the (marginal) log-likelihood of the parameters θ of the model, after integrating out the latent variables, as given by (S2.1). We will see how the EM algorithm relates to minimising the KL divergence between a density $q(\mathbf{w})$ and the posterior of $\mathbf{w}$, and connect this idea to variational methods.

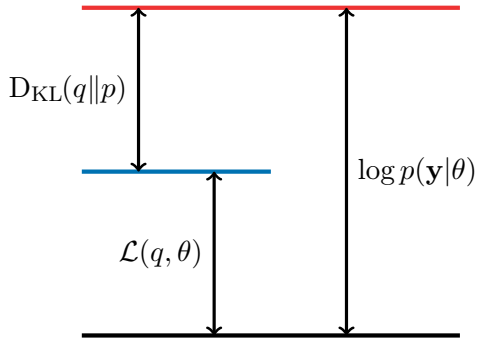


Figure S4.2: Illustration[2] of the decomposition of the log-likelihood into $\mathcal{L}(q,\theta)$ and $\mathrm{D}_{\mathrm{KL}}(q\|p)$. The quantity $\mathcal{L}(q,\theta)$ is a lower bound for the log-likelihood.

As we did in deriving (S4.2), we decompose the (marginal) log-likelihood as

$$\begin{aligned}\log p(\mathbf{y}|\theta) &= \log p(\mathbf{y},\mathbf{w}|\theta) - \log p(\mathbf{w}|\mathbf{y},\theta) \\ &= \int \left\{ \log \frac{p(\mathbf{y},\mathbf{w}|\theta)}{q(\mathbf{w})} - \log \frac{p(\mathbf{w}|\mathbf{y},\theta)}{q(\mathbf{w})} \right\} q(\mathbf{w})\,\mathrm{d}\mathbf{w} \\ &= \underbrace{\mathrm{E}_{\mathbf{w}\sim q}\left[\log \frac{p(\mathbf{y},\mathbf{w}|\theta)}{q(\mathbf{w})}\right]}_{\mathcal{L}(q,\theta)} - \underbrace{\mathrm{E}_{\mathbf{w}\sim q}\left[\log \frac{p(\mathbf{w}|\mathbf{y},\theta)}{q(\mathbf{w})}\right]}_{-\mathrm{D}_{\mathrm{KL}}(q\|p)},\end{aligned}$$

where $q(\mathbf{w})$ is any density function over the latent variables. This decomposition is shown in Figure S4.2. The interest is then to have a density function $q(\mathbf{w})$ which is as close as possible to the true posterior density $p(\mathbf{y}|\mathbf{w},\theta)$ in the KL divergence sense. Since the KL divergence is non-negative, minimising $\mathrm{D}_{\mathrm{KL}}(q\|p)$ is equivalent to maximising $\mathcal{L}(q,\theta)$.

As a remark, the above line of thought should be familiar as it is the exact same one made for variational inference. The twist here is that we will peruse a distribution which tightens the lower bound $\mathcal{L}(q,\theta)$ to the marginal log-likelihood, and this happens when $\mathrm{D}_{\mathrm{KL}}(q\|p)$ is exactly zero, and this in turn happens when q is exactly the true posterior density. That is, for some parameter value, $\theta = \theta^{(t)}$ say, the solution to

$$\arg\max_{q} \mathcal{L}(q,\theta^{(t)}) \tag{S4.5}$$

[2] Reproduced from Bishop (2006, Fig. 9.11).

is $q^{(t+1)}(\mathbf{w}) = p(\mathbf{w}|\mathbf{y}, \theta^{(t)})$, because

$$\mathrm{D}_{\mathrm{KL}}(q\|p) = \mathrm{E}\left[\log \frac{p(\mathbf{w}|\mathbf{y}, \theta^{(t)})}{p(\mathbf{w}|\mathbf{y}, \theta^{(t)})}\right] = 0.$$

At this stage, we have the equality

$$\begin{aligned}
\log p(\mathbf{y}|\theta) &= \mathcal{L}(q^{(t+1)}, \theta) && \text{(S4.6)}\\
&= \mathrm{E}_{\mathbf{w}\sim q^{(t+1)}}\left[\log \frac{p(\mathbf{y}, \mathbf{w}|\theta)}{p(\mathbf{w}|\mathbf{y}, \theta^{(t)})}\right] && \text{(S4.7)}\\
&= \underbrace{\mathrm{E}_{\mathbf{w}\sim q^{(t+1)}}\left[\log p(\mathbf{y}, \mathbf{w}|\theta)\right]}_{Q(\theta|\theta^{(t)})} - \underbrace{\mathrm{E}_{\mathbf{w}\sim q^{(t+1)}}\left[\log p(\mathbf{w}|\mathbf{y}, \theta^{(t)})\right]}_{-H(q^{(t+1)})}, && \text{(S4.8)}
\end{aligned}$$

The term on the left is recognised as the Q function of the E-step

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

while the term on the left is an entropy term which does not depend on θ. Thus, minimising the KL divergence, or maximising the lower bound $\mathcal{L}$ with respect to q, corresponds to the E-step in the EM algorithm.

Furthermore, since equality between the log-likelihood and the lower bound is achieved after the E-step, increasing $\mathcal{L}(q^{(t+1)}, \theta)$ with respect to θ is sure to bring about an increase in the log-likelihood. That is, 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, given initial values $q^{(0)}$ for the distribution and $\theta^{(0)}$ for the parameters, the EM algorithm is seen as iterating between

- **E-step**: $q^{(t+1)} \leftarrow \arg\max_q \mathcal{L}(q, \theta^{(t)})$, i.e., maximise $\mathcal{L}(q, \theta)$ with respect to q, keeping θ fixed. This is equivalent to minimising the KL divergence $\mathrm{D}_{\mathrm{KL}}(q\|p)$.
- **M-step**. $\theta^{(t+1)} \leftarrow \arg\max_\theta \mathcal{L}(q^{(t+1)}, \theta)$, i.e., maximise $\mathcal{L}(q, \theta)$ with respect to θ, keeping $q(\mathbf{w})$ fixed.

When the true posterior distribution $p(\mathbf{w}|\mathbf{y})$ 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 S4.1, we saw that constraining q to be of a factorised

form, then $\tilde{q}$ is a mean-field density. After a variational E-step, the M-step proceeds as normal. This form of the EM is known as *variational EM algorithm* (VEM) (Beal, 2003). The variational EM algorithm can also be modified to obtain MAP estimates by including the log prior density to the maximisation objective in the M-step.

Due to an approximation to the true posterior being used in the E-step, there is no guarantee that the log-likelihood value will increase at each iteration. This is seen pictorially in Section S4.2: since the bound on the log-likelihood is not tight, increasing this bound will not necessarily cause an increase in log-likelihood value (Scenario C), and even if it did, it may not give as much an increase as it would under the true posterior density (Scenario B). Scenario A depicts an ideal case whereby the increase in log-likelihood is as much as it would be if the true posterior density was used.

On a practical note, if the posterior density is intractable, then so is the marginal likelihood, which means that we're unable to determine convergence of the EM using the log-likelihood. Instead, the lower bound $\mathcal{L}(q, \theta)$ should be used, which monotonically increases to a local optima (as in the CAVI algorithm).

S4.3 Comparing variational inference and variational EM

Variational inference is a fully Bayesian treatment of the model, for which the goal is to obtain approximate posterior densities for all latent variables and parameters. Variational EM algorithm on the other hand has the objective of obtaining ML or MAP estimates of the parameters using an EM algorithm in which the E-step is replaced with a variational E-step. In some cases, the CAVI algorithm can resemble an EM algorithm, especially when there is a distinction between latent variables and parameters, and a conjugate exponential family model is involved (Blei et al., 2017).

Variational inference can yield exactly similar point estimates as variational EM if the approximate posterior is symmetric, e.g. a normal distribution. Under a normal posterior, its mean is used as a point estimate, which coincides with the mode, which is a MAP estimate, or in the case of diffuse priors, a ML estimate. However, since the output of variational inference are posterior densities instead of a single point estimate, one is able to obtain posterior standard deviations or credibility intervals about the parameters, something which is not so straightforward under a variational EM or even EM framework.

Derivation of the CAVI algorithm and ELBO for specific models is certainly more tedious than the derivation of the variational EM algorithm. Often, quantities that are required in the derivation include $\mathrm{E}(\theta)$, $\mathrm{E}(\theta^2)$, $\mathrm{E}(\theta^{-1})$, $\mathrm{E}(\log\theta)$ or any other moment of some function of θ, where expectations are taken under the approximating q posterior

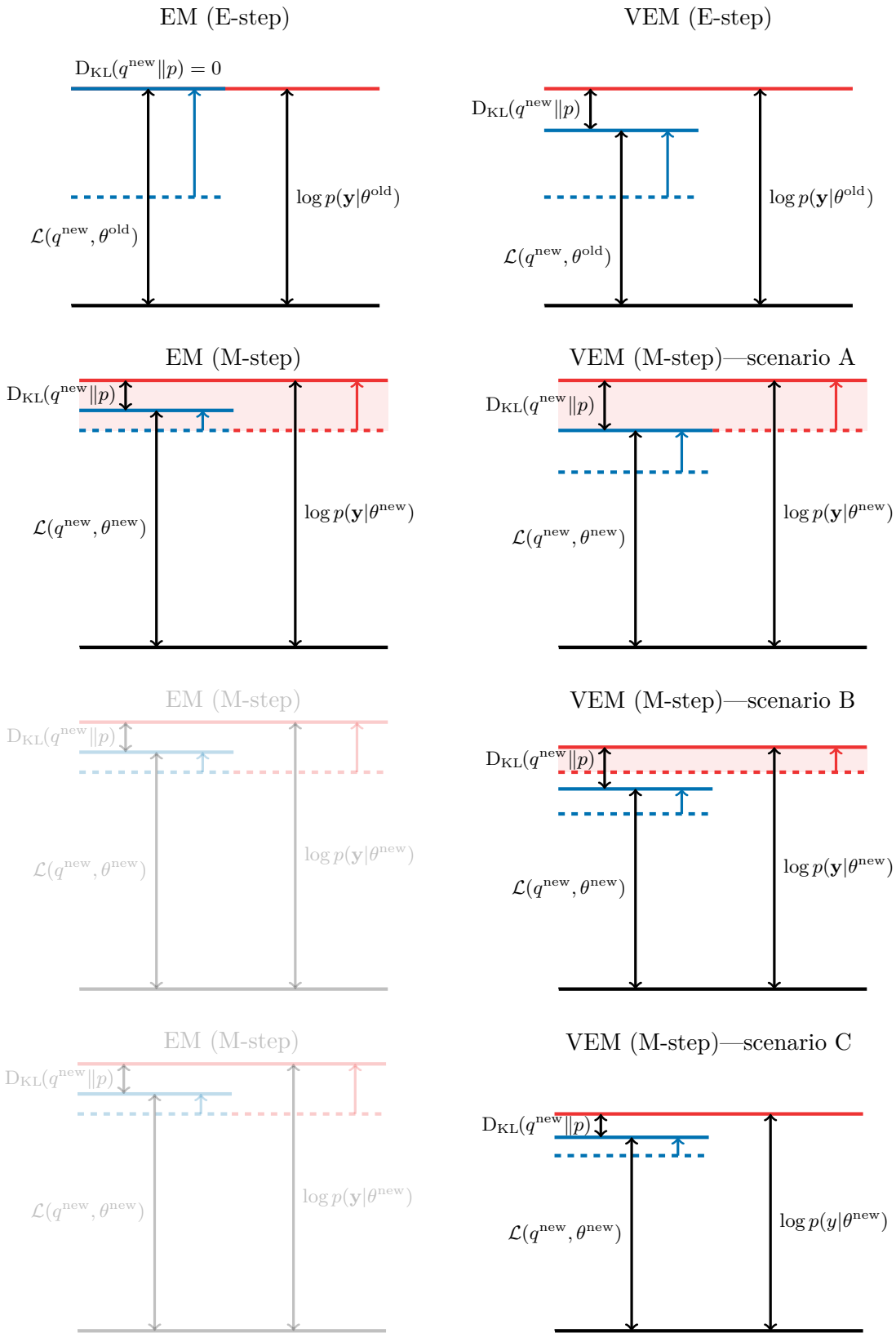


Figure S4.3: Illustration of EM vs Variational EM (VEM) algorithms. Whereas the EM guarantees an increase in log-likelihood value (red shaded region), the VEM does not.

Table S4.1: Comparison between variational inference and variational EM.

Variational inference	Variational EM
GOAL: Posterior densities for $(\mathbf{w}, \theta)$	**GOAL**: ML/MAP estimates for θ
Variational approximation for latent variables and parameters $q(\mathbf{w}, \theta) \approx p(\mathbf{w}, \theta \vert \mathbf{y})$	Variational approximation for latent variables only $q(\mathbf{w}) \approx p(\mathbf{w} \vert \mathbf{y})$
Priors required on θ	Priors not necessary for θ
Derivation can be tedious	Derivation less tedious
Inference on θ through posterior density $q(\theta)$	Asymptotic distribution of θ not well studied; standard errors for θ not easily obtained
Suited to conjugate exponential family models: posteriors will be easily recognisable	Suited to conjugate exponential family models, but not necessary

density. For certain distributions $q(\theta)$ these quantities can be awkward to compute, and may need approximating themselves.

The computational time and storage requirements of variational methods is virtually the same as EM algorithm (Beal, 2003; Blei et al., 2017). Consider the mean-field variational approximation. In variational inference or variational EM, the updating step for the factors involve

$$\tilde{q}_k^{(t+1)}(w_{[k]}) \leftarrow \text{const.} \times \exp\left(\mathrm{E}_{\mathbf{w}_{-k} \sim \tilde{q}^{(t)}}[\log p(\mathbf{y}, \mathbf{w})]\right), \tag{S4.9}$$

for each of the factors of the approximate posterior $q(\mathbf{w}) = \prod_{k=1}^{M} q_k(w_{[k]})$. In the EM algorithm E-step, one obtains the Q function

$$Q(\theta \vert \theta^{(t)}) = \mathrm{E}_{\mathbf{w}}\left(\log p(\mathbf{y}, \mathbf{w}) \vert \mathbf{y}, \theta^{(t)}\right). \tag{S4.10}$$

We can see that in both equations (S4.9) and (S4.10), there is a need to compute the expectation of the joint log density, but the difference between the variational inference and EM or variational EM lies in the M-step. In variational inference one seeks a distribution, while in EM or variational EM one seeks a point estimate (posterior mode) of this distribution.

Bibliography

Beal, Matthew James (2003). "Variational algorithms for approximate Bayesian inference". PhD thesis. Gatsby Computational Neuroscience Unit, University College London.

Beal, Matthew James and Zoubin 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. (Susie) Bayarri, and Adrian F. M. Smith. Oxford University Press, pp. 453–464. ISBN: 978-0-19-852615-5.

Berger, James O. (1985). *Statistical Decision Theory and Bayesian Analysis*. 2nd ed. New York: Springer-Verlag. ISBN: 978-0-387-96098-2. DOI: 10.1007/978-1-4757-4286-2.

Bishop, Christopher (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag. ISBN: 978-0-387-31073-2.

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe (2017). "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518, pp. 859–877. DOI: 10.1080/01621459.2017.1285773.

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell (2017). "Stan: A Probabilistic Programming Language". In: *Journal of Statistical Software* 76.1, pp. 1–32. DOI: 10.18637/jss.v076.i01.

Casella, George (1985). "An Introduction to Empirical Bayes Data Analysis". In: *The American Statistician* 39.2, pp. 83–87. DOI: 10.2307/2682801.

Casella, George and Roger L. Berger (2002). *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury. ISBN: 978-0-534-24312-8.

Davison, Anthony Christopher (2003). *Statistical Models*. Cambridge University Press. ISBN: 978-0-511-81585-0. DOI: 10.1017/CBO9780511815850.

Dempster, Arthur P, Nan M Laird, and Donald B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pp. 1–38.

Duane, Simon, Anthony D Kennedy, Brian J. Pendleton, and Duncan Roweth (1987). "Hybrid Monte Carlo". In: *Physics Letters B* 195.2, pp. 216–222. DOI: 10.1016/0370-2693(87)91197-X.

Itzykson, Claude and Jean-Michel Drouffe (1991). *Statistical Field Theory*. Vol. 2: Strong Coupling, Monte Carlo Methods, Conformal Field Theory and Random Systems.

Cambridge University Press. ISBN: 978-0-511-62278-6. DOI: 10.1017/CBO9780511622786.

Jordan, Michael I., Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul (1999). "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 37.2, pp. 183–233. DOI: 10.1023/A:1007665907178.

Kadane, Joseph B. (2011). *Principles of Uncertainty*. Chapman & Hall/CRC. ISBN: 978-1-4398-6161-5.

McLachlan, Geoffrey and Thriyambakam Krishnan (2007). *The EM Algorithm and Extensions*. 2nd ed. John Wiley & Sons. ISBN: 978-0-471-20170-0. DOI: 10.1002/9780470191613.

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. DOI: 10.1111/1467-9868.00082.

Neal, Radford M. (2011). "MCMC using Hamiltonian dynamics". In: *Handbook of Markov Chain Monte Carlo*. Ed. by Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. Chapman & Hall/CRC. ISBN: 978-1-4200-7941-8. arXiv: 1206.1901 [stat.CO].

Robbins, Herbert (1956). "An Empirical Bayes Approach to Statistics". In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. Ed. by Jerzy Neyman. Vol. 1: Contributions to the Theory of Statistics. Berkeley, CA: University of California Press, pp. 157–163.

Robert, Christian (2007). *The Bayesian Choice*. From Decision-Theoretic Foundations to Computational Implementation. New York: Springer-Verlag. ISBN: 978-0-387-95231-4. DOI: 10.1007/0-387-71599-1.

Wei, Greg C. G. and Martin A. Tanner (1990). "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms". In: *Journal of the American statistical Association* 85.411, pp. 699–704. DOI: 10.2307/2290005.