Computational Statistics II

Unit D.1: Laplace approximation, Variational Bayes, and Expectation Propagation

Tommaso Rigon

University of Milano-Bicocca

Ph.D. in Economics, Statistics and Data Science

Unit D.1

Main concepts

- Laplace approximation;
- Variational Bayes;
- Expectation propagation.

Main references

- Bishop, C. M. (2006). Pattern Recognition and Machine Learning (Chapter 9-10). Springer.
- Blei, D. M., Kucukelbirb A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *JASA*, **112**(518), 859–877.
- Tierney, L. and Kadane, J. (1987). Accurate approximations for posterior moments and marginal densities. *JASA*, **81**(393), 82–86.

Motivations

- MCMC methods could be expensive to compute, especially for large sample sizes n.
- Moreover, many MCMC algorithms require a rough estimate of some key posterior quantities, such as the posterior variance. Recall e.g. the MALA example of **unit B.2**.
- These issues motivates the development of **deterministic approximations** of the posterior distribution.
- Compared to MCMC methods, the accuracy of this class of approximations can not be reduced by running the algorithm longer.
- On the other hand, deterministic approximations are typically **very fast** to compute and sufficiently reliable in several applied contexts.

The Laplace approximation

- Let $\pi(\boldsymbol{\theta} \mid \boldsymbol{X})$ be a **continuous** and **differentiable** posterior density in $\Theta \subseteq \mathbb{R}^p$.

- The **Laplace** approximation is one of the first approximation methods that has been proposed. It was known even before the advent of MCMC.

- The key idea is approximating the log-posterior density $\log \pi(\boldsymbol{\theta} \mid \boldsymbol{X})$ using a **Taylor expansion** around the mode $\hat{\boldsymbol{\theta}}_{\text{MAP}}$, yielding

$$\log \pi(\boldsymbol{\theta} \mid \boldsymbol{X}) \approx \log \pi(\hat{\boldsymbol{\theta}}_{\text{MAP}} \mid \boldsymbol{X}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{MAP}})^\intercal \hat{\boldsymbol{M}}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{MAP}}) + \text{const},$$

  where $\hat{\boldsymbol{M}}$ is the **negative Hessian** of $\log \pi(\boldsymbol{\theta} \mid \boldsymbol{X})$ evaluated at $\hat{\boldsymbol{\theta}}_{\text{MAP}}$, that is

$$\hat{\boldsymbol{M}} = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal} \log \pi(\boldsymbol{\theta} \mid \boldsymbol{X})\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\text{MAP}}}.$$

- Hence, the above quadratic expansion leads to the following **multivariate Gaussian** approximate posterior

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) \approx \mathsf{N}_p\left(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}_{\text{MAP}}, \hat{\boldsymbol{M}}^{-1}\right).$$

Bernstein–von Mises theorem (a rough intuition)

- A fairly strong **asymptotic** justification of the Laplace approximation is based on the **Bernstein–von Mises** theorem.

- Suppose the data $X_1, \dots, X_n$ are iid from a "true" model P_{θ_0}.

- Very **roughly speaking**, under suitable regularity and sampling conditions

$$||\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) - \mathrm{N}_p\left(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}_{\mathrm{MAP}}, \hat{\boldsymbol{M}}^{-1}\right)|| \xrightarrow{P_{\theta_0}} 0, \qquad n \to \infty,$$

  meaning that the total variation distance between the posterior and the Laplace approximation weakly converges to 0 w.r.t. to the law of the sampling process P_{θ_0}.

- Here we are also assuming that $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ and $n\hat{\boldsymbol{M}}^{-1}$ are consistent estimators for the "true" parameter value $\boldsymbol{\theta}_0$ and for the inverse Fisher information matrix, respectively.

- Hence, in several cases and for n large enough, the law $\pi(\boldsymbol{\theta} \mid \boldsymbol{X})$ is roughly a Gaussian centered at the mode and with variance depending on the Fisher information.

Main reference

- van der Vaart, A. W. (1998). Asymptotic Statistics. Cambridge University Press.

Laplace approximation: considerations

- The Laplace approximation is an old and simple method that has appealing asymptotic guarantees. Moreover, it only requires the computation of the Hessian and the MAP.
- Refined **higher order improvements** of expected posterior functionals can be obtained as in Tierney and Kadane (1987).
- On the other hand, especially when the sample size n is relatively small, the quadratic approximation of $\log \pi(\boldsymbol{\theta} \mid \boldsymbol{X})$ may perform poorly.
- For example, if the posterior is not symmetric and unimodal, the MAP is not a good estimate for the posterior mean, thus leading to inaccurate Gaussian approximations.
- Furthermore, if the **parameter space** Θ is **bounded**, a Gaussian approximation could be quite problematic $\implies$ a reparametrization should be considered.
- Finally, it is unclear how to handle **discrete parameter** spaces.

Approximation methods I

- Let $\pi(\boldsymbol{\theta} \mid \boldsymbol{X})$ be the intractable posterior distribution and let $q(\boldsymbol{\theta})$ be a density belonging to $\mathbb{Q}$, where $\mathbb{Q}$ is a general **class** of **tractable densities**.

- An optimal approximation $\hat{q}(\boldsymbol{\theta}) \in \mathbb{Q}$ of the posterior distribution is defined as

$$\hat{q}(\boldsymbol{\theta}) = \arg\min_{q \in \mathbb{Q}} \mathcal{D}\{q(\boldsymbol{\theta}), \pi(\boldsymbol{\theta} \mid \boldsymbol{X})\},$$

  where $\mathcal{D}(\cdot, \cdot)$ is some **divergence** or **metric** over the space of probability distributions.

- An example is the Kullback-Leibler divergence $\mathcal{D}(\cdot, \cdot) = \mathrm{KL}(\cdot \mid\mid \cdot)$.

- Depending on the choice of the divergence $\mathcal{D}(\cdot, \cdot)$ and of the space of approximating densities $\mathbb{Q}$, the problem could be computationally feasible or not.

- Clearly, the posterior $\pi(\boldsymbol{\theta} \mid \boldsymbol{X})$ should not be included in the space of tractable densities $\mathbb{Q}$, otherwise we would get $\hat{q}(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta} \mid \boldsymbol{X})$ for any reasonable divergence $\mathcal{D}(\cdot, \cdot)$.

Approximation methods II

- As for the choice of $\mathcal{D}(\cdot,\cdot)$, it would be theoretically appealing to consider metrics such as the Hellinger distance, the total variation distance or the Wasserstein distance.

- Unfortunately, even when we let $\mathbb{Q}$ be the space of multivariate Gaussians, finding the optimal density $\hat{q}(\boldsymbol{\theta})$ could be problematic.

- A basic requirement is that the optimization procedure should not depend on the intractable normalizing constant of the posterior.

- We will consider two different though quite related divergences.

- The $\textsc{kl}\{q(\boldsymbol{\theta}) \mid\mid \pi(\boldsymbol{\theta} \mid \boldsymbol{X})\}$ divergence, leading to the **variational Bayes** method.

- The $\textsc{kl}\{\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) \mid\mid q(\boldsymbol{\theta})\}$ divergence, leading to the **expectation propagation** method.

The evidence lower bound (ELBO)

- In first place, let us note that the following decomposition hold

$$\log \pi(\boldsymbol{X}) = \textsc{kl}\{q(\boldsymbol{\theta}) \mid\mid \pi(\boldsymbol{\theta} \mid \boldsymbol{X})\} + \textsc{elbo}\{q(\boldsymbol{\theta})\},$$

- Recall that the **Kullback-Leibler divergence** is

$$\textsc{kl}\{q(\boldsymbol{\theta}) \mid\mid \pi(\boldsymbol{\theta} \mid \boldsymbol{X})\} = -\int_{\Theta} q(\boldsymbol{\theta}) \log \frac{\pi(\boldsymbol{\theta} \mid \boldsymbol{X})}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}.$$

- The **evidence lower bound** $\textsc{elbo}\{q(\boldsymbol{\theta})\}$ is instead defined as

$$\textsc{elbo}\{q(\boldsymbol{\theta})\} = \int_{\Theta} q(\boldsymbol{\theta}) \log \frac{\pi(\boldsymbol{\theta}, \boldsymbol{X})}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}.$$

- **Key property**. Since $\log \pi(\boldsymbol{X})$ does not depend on $\boldsymbol{\theta}$, we obtain that

$$\hat{q}(\boldsymbol{\theta}) = \arg\min_{q \in \mathbb{Q}} \textsc{kl}\{q(\boldsymbol{\theta}) \mid\mid \pi(\boldsymbol{\theta} \mid \boldsymbol{X})\} = \arg\max_{q \in \mathbb{Q}} \textsc{elbo}\{q(\boldsymbol{\theta})\},$$

therefore the optimization does not depend on the intractable normalizing constant.

Evidence lower bound (ELBO)

- The ELBO is indeed a **lower bound** of the marginal likelihood, because the divergence $\text{KL}\{q(\boldsymbol{\theta}) \,||\, \pi(\boldsymbol{\theta} \mid \boldsymbol{X})\} \geq 0$, implying that

$$\text{ELBO}\{q(\boldsymbol{\theta})\} \leq \log \pi(\boldsymbol{X}).$$

- This property of the ELBO has led to using the variational bound as a model selection criterion, on the assumption that the ELBO is a good approximation of the marginal.

- **Remark**. Even when the optimal distribution $\hat{q}(\boldsymbol{\theta})$ can found, there is no guarantee that the minimized KL

$$\text{KL}\{\hat{q}(\boldsymbol{\theta}) \,||\, \pi(\boldsymbol{\theta} \mid \boldsymbol{X})\} \geq 0$$

  will be small in absolute terms.

- Moreover, quantifying the value of $\text{KL}\{\hat{q}(\boldsymbol{\theta}) \,||\, \pi(\boldsymbol{\theta} \mid \boldsymbol{X})\} = \log \pi(\boldsymbol{X}) - \text{ELBO}\{q(\boldsymbol{\theta})\}$ would require the knowledge of the normalizing constant, which is intractable.

- Essentially, it is currently hard to **assess the quality** of the obtained **approximation** without comparing it with some "gold standard" such as MCMC.

Mean-field approximation

- The VB optimization problem is ill-posed if we do not specify a tractable class $\mathbb{Q}$.
- For reasons that will become clear later on, a convenient assumption is restricting the focus on the class $\mathbb{Q}$ of **mean-field approximations**, in which we assume

$$q(\boldsymbol{\theta}) = \prod_{b=1}^{B} q(\boldsymbol{\theta}_b),$$

  implying that we are forcing **independence** among B groups of parameters.

- It is important to notice that dependence is preserved within each block of parameters.
- Moreover, note that we are not forcing $q(\boldsymbol{\theta})$ to belong to any known parametric family of distributions. The only assumption we are making is independence.

Derivation of the CAVI algorithm

- Under the mean-field assumption, the optimization of the ELBO can be written as

$$\text{ELBO}\{q(\boldsymbol{\theta})\} = \int_\Theta \prod_{b=1}^B \{q(\boldsymbol{\theta}_b) \log \pi(\boldsymbol{\theta}, \boldsymbol{X})\} \,\mathrm{d}\boldsymbol{\theta} - \int_\Theta \prod_{b=1}^B \{q(\boldsymbol{\theta}_b) \log q(\boldsymbol{\theta}_b)\} \,\mathrm{d}\boldsymbol{\theta}.$$

- We aim at maximizing the bth component $q(\boldsymbol{\theta}_b)$, keeping the others fixed. Thus, we express the ELBO isolating the term $q(\boldsymbol{\theta}_b)$, obtaining

$$\int q(\boldsymbol{\theta}_b) \left\{ \int \log \pi(\boldsymbol{\theta}, \boldsymbol{X}) \prod_{j \neq b} q(\boldsymbol{\theta}_j) \mathrm{d}\boldsymbol{\theta}_{-b} \right\} \mathrm{d}\boldsymbol{\theta}_b - \int q(\boldsymbol{\theta}_b) \log q(\boldsymbol{\theta}_b) \mathrm{d}\boldsymbol{\theta}_b + c_b,$$

where c_b denotes a term not depending on $\boldsymbol{\theta}_b$.

- Defining the density $\log \tilde{\pi}(\boldsymbol{\theta}_b, \boldsymbol{X}) = \mathbb{E}_{-b}\{\log \pi(\boldsymbol{\theta}, \boldsymbol{X})\} + \text{const}$ and re-arranging the terms, we get

$$\text{ELBO}\{q(\boldsymbol{\theta})\} = \int q(\boldsymbol{\theta}_b) \log \frac{\tilde{\pi}(\boldsymbol{\theta}_b, \boldsymbol{X})}{q(\boldsymbol{\theta}_b)} \mathrm{d}\boldsymbol{\theta}_b + \tilde{c}_b = -\text{KL}\{q(\boldsymbol{\theta}_b) \,||\, \tilde{\pi}(\boldsymbol{\theta}_b, \boldsymbol{X})\} + \tilde{c}_b.$$

Derivation of the CAVI

- The above previous chain of equations implies that the **local maximization** of the $\text{ELBO}(q(\boldsymbol{\theta})$ with respect to the bth term of $q(\boldsymbol{\theta}_b)$ is obtained by setting

$$\hat{q}(\boldsymbol{\theta}_b) \propto \exp\left[\mathbb{E}_{-b}\{\log \pi(\boldsymbol{\theta}, \boldsymbol{X})\}\right],$$

  for any $b = 1, \dots, B$.

- In practice, the above expectation is often straightforward to compute and usually some **known kernel** can be recognized (as in the Gibbs sampling).

- In the CAVI algorithm, we iteratively update the factors $q(\boldsymbol{\theta}_b)$ by using the locally maximized terms given the others.

- By construction, this produces a **monotonic sequence** that convergences to a local optimum of the ELBO.

Properties and convergence

- The CAVI is an appealing algorithm for maximizing the ELBO under the mean-field assumption, but in principle one could use any other optimizer.
- The necessary computations and expectations are usually doable if the full conditional distributions belong to some **exponential family**.
- The algorithms stops whenever the ELBO sequence has reached convergence.
- Moreover, checking that the ELBO is indeed monotone is a good practice to verify the correctness of the implementation.
- Although not shown here, a common application of the CAVI algorithm is indeed the case of Bayesian **mixture models**.

The CAVI for a Gaussian example

- As in **unit A.2**, let us assume the observations $(x_1, \dots, x_n)$ are draws from

$$(x_i \mid \mu, \tau) \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \tau^{-1}), \qquad i = 1, \dots, n,$$

  with independent priors $\mu \sim \mathrm{N}(\mu_\mu, \sigma^2_\mu)$ and $\tau \sim \mathrm{Ga}(a_\tau, b_\tau)$.

- Assuming a mean-field approximation $q(\mu, \tau) = q(\mu)q(\tau)$, the CAVI algorithm iterates between the following steps simple steps.

- **Update $q(\mu)$**. The locally optimal variational distribution for $q(\mu)$ is

$$q(\mu) = \mathrm{N}(\mu \mid \mu_n, \sigma^2_n), \quad \mu_n = \sigma^2_n \left(\frac{\mu_\mu}{\sigma^2_\mu} + \mathbb{E}_q(\tau) \sum_{i=1}^n x_i \right), \quad \sigma^2_n = \left(n\, \mathbb{E}_q(\tau) + \frac{1}{\sigma^2_\mu} \right)^{-1}.$$

- **Update $q(\tau)$**. The locally optimal variational distribution for $q(\tau)$ is

$$q(\tau) = \mathrm{Ga}\left(\tau \mid a_n, b_n\right), \quad a_n = a_\tau + n/2, \quad b_n = b_\tau + \frac{1}{2} \sum_{i=1}^n \mathbb{E}_q\{(x_i - \mu)^2\}.$$

Underestimation of the variability

- As previously mentioned, the combination of mean-field assumption $+$ VB approach typically leads to a sensible **underestimation of the variability**.

- In first place, this is a consequence of the **insufficient flexibility** of the mean-field class of approximating densities.

- Indeed, if the densities in $\mathbb{Q}$ were arbitrarily close to the posterior, this phenomenon would be negligible in practice.

- In second place, this is a consequence of the chosen divergence. Indeed, the quantity

$$\mathrm{KL}\{q(\boldsymbol{\theta}) \,||\, \pi(\boldsymbol{\theta} \mid \boldsymbol{X})\} = -\int_{\Theta} q(\boldsymbol{\theta}) \log \frac{\pi(\boldsymbol{\theta} \mid \boldsymbol{X})}{q(\boldsymbol{\theta})}, \mathrm{d}\boldsymbol{\theta}$$

  favors the choice of densities $q(\boldsymbol{\theta})$ which are included in the support of $\pi(\boldsymbol{\theta} \mid \boldsymbol{X})$.

- Indeed, there is a large positive contribution to the above KL for those values of $\boldsymbol{\theta}$ such that $\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) \approx 0$, unless $q(\boldsymbol{\theta}) \approx 0$ as well.

Expectation propagation (EP)

- The **Expectation Propagation** algorithm (EP) has been proposed by Minka (2001).
- The EP approach aims at minimizing the divergence $\text{KL}\{\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) \mid\mid q(\boldsymbol{\theta})\}$, which is the reversed situation compared to the VB.
- At least in principle, the EP is expected to **overestimate** the variability of the posterior, but this is not a big concern in practice.
- Indeed, the EP does not rely on the mean-field approximation for $\mathbb{Q}$. In contrast, the class $\mathbb{Q}$ will be some **parametric exponential family of distributions**.
- The EP is essentially a **heuristic method** for minimizing $\text{KL}\{\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) \mid\mid q(\boldsymbol{\theta})\}$, as there are little theoretical guarantees that this is indeed occurring.
- On the other hand, in specific contexts the EP approach outperforms other approaches.

EP and exponential families

- Let us assume $\mathbb{Q}$ is an **exponential family of distributions**, with natural parameters $\boldsymbol{\eta} \in \mathbb{R}^p$, so that
$$q(\boldsymbol{\theta} \mid \boldsymbol{\eta}) = h(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\theta}^\intercal \boldsymbol{\eta} - K(\boldsymbol{\eta})\right\}.$$

- Then, it can be shown that the minimum of the KL divergence is such that
$$\min_{q \in \mathbb{Q}} \mathrm{KL}\{\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) \mid\mid q(\boldsymbol{\theta} \mid \boldsymbol{\eta})\} = \min_{\boldsymbol{\eta} \in \mathbb{R}^p} \mathrm{KL}\{\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) \mid\mid q(\boldsymbol{\theta} \mid \boldsymbol{\eta})\},$$
where the optimal set of parameters $\hat{\boldsymbol{\eta}}$ minimizing the divergence is such that
$$\mathbb{E}_q(\boldsymbol{\theta}) = \mathbb{E}(\boldsymbol{\theta} \mid \boldsymbol{X}).$$

- In words, the optimal parameter $\hat{\boldsymbol{\eta}}$ is the one **matching** the true posterior mean of the **natural parameter** $\mathbb{E}(\boldsymbol{\theta} \mid \boldsymbol{X})$, with the mean $\mathbb{E}_q(\boldsymbol{\theta})$ under the variational distribution.

- In the multivariate Gaussian case, this implies that both the **mean** and the **variance** are matched.

The EP procedure

- The moment-matching procedure we just described is not directly applicable, because the posterior mean of the natural parameter $\boldsymbol{\theta}$ is unknown.

- The EP seeks for an heuristic procedure that iteratively minimize the KL using the principle of **moment-matching** local components.

- In first place, let us assume that the **joint likelihood factorizes** as follows

$$\pi(\boldsymbol{\theta}, \boldsymbol{X}) = \prod_{i=0}^{n} \pi_i(\boldsymbol{\theta}, \boldsymbol{X}),$$

  the first term corresponds to the prior, so that $\pi_0(\boldsymbol{\theta}, \boldsymbol{X}) = \pi(\boldsymbol{\theta})$.

- This is a common **modelling assumption**, which is satisfied for example if the data are conditionally independent (i.e. regression).

The EP procedure

- In second place, note that the **exponential family** assumption for $\mathbb{Q}$ guarantees that there exists a **decomposition** of the form

$$q(\boldsymbol{\theta} \mid \boldsymbol{\eta}) = \frac{1}{K}\prod_{i=0}^{n} q_i(\boldsymbol{\theta} \mid \boldsymbol{\eta}_i),$$

with $\boldsymbol{\eta} = \sum_{i=0}^{n} \boldsymbol{\eta}_i$ and K being the normalizing constant, and where the $q_i(\boldsymbol{\theta} \mid \boldsymbol{\eta}_i)$ is proportional to an exponential family of distributions.

- For example, if we consider a Gaussian kernel

$$q_i(\beta \mid \boldsymbol{r}_i, \boldsymbol{M}_i) = \exp\left\{-\frac{1}{2}\beta^\intercal \boldsymbol{M}_i\beta + \beta^\intercal \boldsymbol{r}_i\right\} \implies q(\boldsymbol{\theta} \mid \boldsymbol{\eta}) \propto \exp\left\{-\frac{1}{2}\beta^\intercal \boldsymbol{M}\beta + \beta^\intercal \boldsymbol{r}\right\},$$

with $\boldsymbol{r} = \sum_{i=0}^{n} \boldsymbol{r}_i$ and $\boldsymbol{M} = \sum_{i=0}^{n} \boldsymbol{M}_i$.

The EP procedure

- Recall that the goal is obtaining the value $\hat{\boldsymbol{\eta}}$ minimizing the following KL

$$\min_{\boldsymbol{\eta}\in\mathbb{R}^p} \mathrm{KL}\{\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) \mid\mid q(\boldsymbol{\theta} \mid \boldsymbol{\eta})\} = \min_{\boldsymbol{\eta}\in\mathbb{R}^p} \mathrm{KL}\left\{\frac{1}{\pi(\boldsymbol{X})}\prod_{i=0}^{n}\pi_i(\boldsymbol{\theta},\boldsymbol{X}) \mid\mid \frac{1}{K}\prod_{i=0}^{n} q_i(\boldsymbol{\theta} \mid \boldsymbol{\eta}_i)\right\}.$$

- Unfortunately, this is unfeasible so we proceed by iteratively updating each factor $q_j(\boldsymbol{\theta} \mid \boldsymbol{\eta}_i)$, for $j = 0, \dots, n$, keeping the other fixed.

- Hence, we iteratively update **only the jth factor** $q_j(\boldsymbol{\theta} \mid \boldsymbol{\eta}_i)$ so that

$$\min_{\boldsymbol{\eta}_j\in\mathbb{R}^p} \mathrm{KL}\left\{\frac{1}{K_j}\pi_j(\boldsymbol{\theta},\boldsymbol{X})\prod_{i\neq j} q_i(\boldsymbol{\theta} \mid \boldsymbol{\eta}_i) \mid\mid \frac{1}{K}q_j(\boldsymbol{\theta} \mid \boldsymbol{\eta}_i)\prod_{i\neq j} q_i(\boldsymbol{\theta} \mid \boldsymbol{\eta}_i)\right\},$$

where K_j is the normalizing constant.

- The minimizer $\hat{\boldsymbol{\eta}}_j$ of the above KL is indeed solved by **moment-matching**, possibly leveraging on a well-behaved numerical integration step.

The EP procedure

- The minimization of the previously considered local KL takes advantage of several recursive formulas, speeding up computations.
- There is no guarantee this algorithm is going to converge at all, especially if the target density is not log-concave.
- Moreover, the moment-matching step often involves **numerical integration**, which could be computationally delicate.
- Finally, the EP approach requires a particular likelihood structure and only works using exponential families.
- That said, when considering well-behaved posteriors (such as logistic regression), the EP strategy is very effective and often numerically stable.