Approximate methods

Lecturer: Riccardo Corradin

Being approximate: why

- **MCMC methods could be expensive to compute**, especially for large sample sizes n.
- The computational cost increases mainly by two components: **evaluating the (log)likelihood function**, usually when we need its normalization constant and is hard to compute, and **proposing a new candidate**.
- Moreover, many mcmc algorithms require a rough estimate of some key posterior quantities, such as the posterior variance. Recall, e.g., the MALA.
- These issues motivate the development of deterministic approximations of the posterior distribution or the definition of approximate sampling schemes, **that avoid the usual bottlenecks**.
- 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, while approximate sampling schemes can be used in challenging and intractable scenarios.

Laplace Approximation

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 $\hat{\boldsymbol{\theta}}_{MAP}$, yielding

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

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

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

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

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) \approx N_p(\hat{\boldsymbol{\theta}}_{MAP}, \hat{M}^{-1}).$$

Comments on the Laplace approximation

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

- Suppose the data $X_1, \ldots, X_n$ are iid from a "true" model $P_{\boldsymbol{\theta}_0}$.

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

$$||\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) - N_p(\hat{\boldsymbol{\theta}}_{MAP}, \hat{M}^{-1})||_{TV} \xrightarrow{P_{\boldsymbol{\theta}_0}} 0, \qquad n \to \infty,$$

  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 $P_{\boldsymbol{\theta}_0}$.

- Here we are also assuming that $\hat{\boldsymbol{\theta}}_{MAP}$ and $n\hat{M}^{-1}$ 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.

Comments on the Laplace approximation

- The Laplace approximation is an old and simple method with 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 $\Rightarrow$ a reparametrization should be considered.
- Finally, it is unclear how to handle discrete parameter spaces.

More general approximations

- Let $\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) \in \mathbb{P}_{\boldsymbol{\theta}}$ be a posterior distribution (intractable), and let $q(\boldsymbol{\theta}) \in \mathcal{Q}$ be a density, where $\mathbb{Q}_{\boldsymbol{\theta}} \subseteq \mathbb{P}_{\boldsymbol{\theta}}$ denotes the space **tractable densities** with support Θ.

- In a general fashion, 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 distribution.

- Among other, one choice of $\mathcal{D}$ commonly used is the Kullback-Leibler divergence $\mathcal{D}(\cdot, \cdot) = \mathrm{KL}(\cdot \mid\mid \cdot)$, resulting in well-known approximation approaches.

- 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, this approach is relevant whenever the posterior $\pi(\boldsymbol{\theta} \mid \boldsymbol{X})$ is not included in the subspace of tractable density $\mathbb{Q}$.

More general approximations

- 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 Q be the space of multivariate Gaussians, **finding the optimal approximate 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 $KL(q(\boldsymbol{\theta}) \,||\, \pi(\boldsymbol{\theta} \mid \boldsymbol{X}))$ divergence, leading to the variational Bayes method.
- The $KL(\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) \,||\, q(\boldsymbol{\theta}))$ divergence, leading to the expectation propagation method.

The evidence lower bound (ELBO)

- In the first place, we recall that

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

- Hence, by multiplying and dividing the fractional term by $\pi(\boldsymbol{X})$, we obtain

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

  where the first term in the last expression is called **evidence lower bound** $ELBO(q(\boldsymbol{\theta}))$, and the last term does not depend on $\boldsymbol{\theta}$.

- Hence, if we want the best approximating density function $q(\boldsymbol{\theta}) \in \mathbb{Q}_{\boldsymbol{\theta}}$, we have

$$\hat{q}(\boldsymbol{\theta}) = \underset{q \in \mathbb{Q}}{\arg\min}\, KL(q(\boldsymbol{\theta}) \,||\, \pi(\boldsymbol{\theta} \mid \boldsymbol{X})) = \underset{q \in \mathbb{Q}}{\arg\max}\, ELBO(q(\boldsymbol{\theta})),$$

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

The evidence lower bound (ELBO)

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

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

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

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

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

  will be **small in absolute terms**.

- Moreover, quantifying the value of $KL(\hat{q}(\boldsymbol{\theta}) \,||\, \pi(\boldsymbol{\theta} \mid \boldsymbol{X})) = \log \pi(\boldsymbol{X}) - ELBOq(\boldsymbol{\theta})$ 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.

The evidence lower bound (ELBO)

- 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 approximation, we can derive the so-called coordinate ascent variational inference (CAVI) algorithm.

- The optimization of the ELBO can be written as optimization of

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

- 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

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

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

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

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

Properties and convergence

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

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

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

- In practice, the above expectation is often straightforward to compute, and some **known kernel** can usually 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.

- 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 algorithm stops whenever the ELBO sequence has converged.

Underestimation of the variability

- The combination of mean-field assumption + VB approach typically leads to a sensible **underestimation** of the variability.
- In the 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 second place, this is a consequence of the chosen divergence. Indeed, the quantity

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

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

The cavi for a Gaussian example

- Let us assume the observations $(x_1, \ldots, x_n)$ are draws from

$$X_i \mid \mu, \tau \sim N(\mu, \tau^{-1}), \qquad i = 1, \ldots, n,$$

with independent priors $\mu \sim N(m_0, s_0^2)$ and $\tau \sim Ga(a_0, b_0)$.

- 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

$$q(\mu) \stackrel{d}{=} N(m_n, s_n^2), \qquad m_n = s_n^2 \left(\frac{m_0}{s_0^2} + \mathrm{E}_q[\tau] \sum_{i=1}^{n} x_i \right), \qquad s_n^2 = \left(n\mathrm{E}_q[\tau] + \frac{1}{s_0^2} \right)^{-1}.$$

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

$$q(\tau) \stackrel{d}{=} Ga(a_n, b_n), \qquad a_n = a_0 + \frac{n}{2}, \qquad b_n = b_0 + \frac{1}{2} \sum_{i=1}^{n} \mathrm{E}_q\left[(x_i - \mu)^2 \right].$$

The Pima dataset

- The logistic regression case has often been presented as an example in which mean-field variational Bayes can not be applied. See, for example, Section 10.5 of Bishop (2006).

- A priori, we set $\boldsymbol{\beta} \sim N_p(\boldsymbol{b}_0, B_0)$.
- Recently, a mean-field approximation for an augmented version of the model has been proposed (Durante, D. and Rigon, T., 2019).

- We consider a sequence of augmented variables $Z_1, \ldots, Z_n \sim PG(1, 0)$, where $PG(\alpha, \gamma)$ denotes the Pólya-gamma random variable, i.e.

$$Z \stackrel{d}{=} \frac{1}{2\pi^2} \sum_{j \geq 1} \frac{G_j}{(j - 1/2)^2 + \gamma^2/(4\pi^2)},$$

  with $G_j \sim Ga(\alpha, 1)$, $\alpha > 0$ and $\gamma \in \mathbb{R}$.

- The density of a Pólya-gamma random variable is expressed in terms of an infinite summation, but it can be easily simulated.

The Pima dataset

- The algorithm iterates between two simple steps.
- Update $q(\boldsymbol{\beta})$. The locally optimal variational distribution for $q(\boldsymbol{\beta})$ is

$$\begin{aligned} q(\boldsymbol{\beta}) &\propto \exp\left[\mathrm{E}_q\{\log \pi(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\beta}) + \log \pi(\boldsymbol{\beta})\}\right] \\ &\propto \pi(\boldsymbol{\beta}) \left\{ \sum_{i=1}^{n} (y_i - 1/2) \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta} - \frac{1}{2} \mathrm{E}_q[z_i] (\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta})^2 \right\} \end{aligned}$$

  Re-arranging the above equation, we obtain that $q(\boldsymbol{\beta}) \stackrel{d}{=} N_p(\boldsymbol{\mu}, \Sigma)$ with

$$\boldsymbol{\mu} = \Sigma(\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{y} - 1/2) + B_0^{-1}\boldsymbol{b}_0), \qquad \Sigma = (\boldsymbol{X}^{\mathsf{T}} \mathrm{E}_q[Z] \boldsymbol{X} + B^{-1})^{-1},$$

  where $Z = \mathrm{diag}(z_1, \ldots, z_n)$ and its expectation is taken w.r.t. $q(\boldsymbol{z})$.
- Hence, the optimal variational distribution for $\boldsymbol{\beta}$ is Gaussian. This is an implication of the mean-field structure and not an assumption.

The Pima dataset

- The second step involves the variational distribution $q(\boldsymbol{z})$.
- Update $q(\boldsymbol{z})$. The locally optimal variational distribution for $q(\boldsymbol{z})$ is

$$\begin{aligned} q(\boldsymbol{z}) &\propto \exp \mathrm{E}_q[\log p(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\beta})] \\ &\propto \prod_{i=1}^{n} p(z_i \mid 1, 0) \exp\left(-\frac{z_i}{2} \mathrm{E}_q[\eta_i^2]\right) \end{aligned}$$

  Re-arranging the above equation, we have

$$q(\boldsymbol{z}) = \prod_{i=1}^{n} PG(1, \mathrm{E}_q[\eta_i^2]).$$

- Hence, the optimal variational distribution for $\boldsymbol{z}$ are independent Pólya-gamma distributions. As before, this is an implication and not an assumption.

CAVI algorithm with Pima dataset

- We consider again the Pima dataset, using the CAVI algorithm.
- We use a $N_p(\mathbf{0}, 100\mathrm{I}_p)$ as prior distribution.

```
# prior parameters
b <- rep(0, 8)
B <- diag(100, 8)

# Running the CAVI
fit_CAVI <- logit_CAVI(y, X, B, b)

est_coef <- fit_CAVI$mu
# X       Xnpreg   Xglu    Xbp  Xskin   Xbmi   Xped   Xage
# -0.990  0.406  1.097 -0.095  0.072  0.569  0.452  0.284
```

Approximate Bayesian computation

Even more approximate

- In the previous case, we introduced **approximate methods** to avoid the intractability of $\pi(\boldsymbol{\theta} \mid \boldsymbol{X})$, with $X_i \in \mathbb{X}$ and $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$, but we assumed that we can **evaluate/deal** with $\pi(\boldsymbol{\theta}, \boldsymbol{X})$.

- Recall that $\pi(\boldsymbol{\theta}, \boldsymbol{X}) = \mathrm{L}(\boldsymbol{X} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$, where the **likelihood term** is given by

$$\mathrm{L}(\boldsymbol{X} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta}).$$

- In many scenarios, even evaluating the **likelihood is unfeasible**, since we have

$$f(x_i \mid \boldsymbol{\theta}) = \frac{1}{Z_{\boldsymbol{\theta}}} g(x_i \mid \boldsymbol{\theta}),$$

  where $Z_{\boldsymbol{\theta}} < +\infty$ is an intractable normalizing constant, which depends on the value of $\boldsymbol{\theta}$.

- At the same time, even if we cannot evaluate the density function $f(x_i \mid \boldsymbol{\theta})$ of X_i, in many scenarios we can easily sample X_i.

Some examples

- **g-and-k** distribution, extends the Gaussian distribution with added skewness and heavier/lighter tails, defined by the quantile function

$$F_{gk}^{-1}(u) = a + b\Big[1 + c\tanh(gu/2)\Big]\Phi^{-1}(u)\Big[1 + (\Phi^{-1}(u))^2\Big]^k,$$

with $\Phi^{-1}(\cdot)$ being the quantile function of a standard Gaussian distribution, $a \in \mathbb{R}$, $b > 0$, $g \geq 0$, $k \geq 0$ are location, scale, shape (affecting the skewness) and shape (affecting the kurtosis) parameters respectively.

- **Hidden Potts model**, let $i \in \{1, \ldots, n\}$ denotes the notes of a lattice (e.g. pixels in an image), with $y_i \in \{1, \ldots, k\}$ denotes the node's ith node's state. The Potts model is a Markov random field of the form

$$\mathrm{P}(y_i \mid y_{i\sim\ell}) = \frac{\exp\left(\theta \sum_{i\sim\ell} 1_{[y_i = y_\ell]}\right)}{\sum_{j=1}^{k} \exp\left(\theta \sum_{i\sim\ell} 1_{[y_i = y_\ell]}\right)}$$

where $i \sim \ell$ are neighboring nodes of i, and $\theta \geq 0$.

Key ideas of ABC

- Suppose we observe $X_1, \ldots, X_n$ with

$$X_i \mid \boldsymbol{\theta} \overset{iid}{\sim} f(x_i \mid \boldsymbol{\theta}), \qquad \text{and} \qquad \boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}).$$

- The **intuition of ABC** methods is that, once we have a parameter value $\boldsymbol{\theta}$, we generate a **set of synthetic data** $S_1, \ldots, S_n$.

- If the synthetic data $S_1, \ldots, S_n$ are close enough to the observed one $X_1, \ldots, X_n$, then the data generating process of the synthetic data is similar to the one of the observed data.

- But the two data generating process share the same structure and are indexed by $\boldsymbol{\theta}$. Hence, if the two set of data are close enough, $\boldsymbol{\theta}$ is a reasonable parameter also for $X_1, \ldots, X_n$.

- Such approach can be combined with **different sampling strategies** and refined to obtain efficient sampler for posterior arising from intractable likelihood.

- This motivates also labeling these approaches as "likelihood free methods",

Rejection sampler, an old friend

- We recall one of the basic Monte Carlo sampler, the rejection sampler.
- Suppose we want to sample from $\pi(\boldsymbol{\theta} \mid X)$ using an auxiliary distribution $h(\boldsymbol{\theta})$. With rejection sampling we can produce a sample as follows.

At the rth sampling step

i) We generate $\boldsymbol{\theta}^{(r)} \overset{iid}{\sim} h(\boldsymbol{\theta})$ (independent of the previous state).

ii) Accept $\boldsymbol{\theta}^{(r)}$ with probability

$$\frac{\pi(\boldsymbol{\theta}^{(r)} \mid \boldsymbol{X})}{Kh(\boldsymbol{\theta}^{(r)})}, \qquad \text{with} \qquad K \geq \max_{\boldsymbol{\theta}} \frac{\pi(\boldsymbol{\theta}^{(r)} \mid \boldsymbol{X})}{h(\boldsymbol{\theta}^{(r)})},$$

otherwise go back to i).

- Here we still have to evaluate the likelihood function, we can extend this algorithm accommodating for synthetic data.
- The support of $h(\cdot)$ should cover the support of $\pi(\boldsymbol{\theta}^{(r)} \mid \boldsymbol{X})$.

ABC rejection sampler

- A first intuitive ABC sampler can be obtained as follows.

At the rth sampling step

i) We generate $\boldsymbol{\theta}^{(r)} \overset{iid}{\sim} h(\boldsymbol{\theta})$ (independent of the previous state).

ii) We generate $\boldsymbol{S} \overset{iid}{\sim} f(s \mid \boldsymbol{\theta}^{(r)})$.

ii) Accept $\boldsymbol{\theta}^{(r)}$ with probability

$$\frac{\pi(\boldsymbol{\theta}^{(r)})}{Kh(\boldsymbol{\theta}^{(r)})}, \qquad \text{with} \qquad K \geq \max_{\boldsymbol{\theta}} \frac{\pi(\boldsymbol{\theta}^{(r)})}{h(\boldsymbol{\theta}^{(r)})},$$

otherwise go back to i).

- We do not need anymore to evaluate the data density function, but just to use it as sampling mechanism.
- However, having $\boldsymbol{S}, \boldsymbol{X} \in \mathbb{X}^n \subseteq \mathbb{R}^d$, then we have $P(\boldsymbol{S} = \boldsymbol{X}) = 0$.
- Further, when the number of observation is large, using the whole sample information is slowing the sampler possibly without practical benefits.

Relaxing the matching condition

- In general, we **do not assume** an identity function to compare the observe data and the synthetic one.

- In fact, we require that the observed data are **close enough** to the synthetic one, for example requiring that

$$||\boldsymbol{X} - \boldsymbol{S}|| \geq \epsilon,$$

for some threshold $\epsilon \geq 0$ and some distance measure $||\cdot||$.

- When $\epsilon = 0$, we go back on the previous rejection sampler and we produce a sample from $\pi(\boldsymbol{\theta} \mid \boldsymbol{X})$. However, usually we consider $\epsilon > 0$, producing a sample from an approximation of such a distribution.

- The previous can be further generalized by applying a kernel function to the previous distance, hence moving from

$$1_{[||\boldsymbol{X}-\boldsymbol{S}||\leq\epsilon]} \qquad \text{to} \qquad \mathrm{K}_\lambda(||\boldsymbol{X} - \boldsymbol{S}||) = \frac{1}{\lambda}\mathrm{K}\left(\frac{||\boldsymbol{X} - \boldsymbol{S}||}{\lambda}\right).$$

Common choices are the triangular kernel $\mathrm{K}(u) = (1 - |u|)1_{[|u|\leq 1]}$ and the Gaussian kernel $\mathrm{K}(u) = \phi(u)$, with ϕ being the density function of a standard Gaussian distribution.

Simplify the comparison

- Secondly, when the number of observations increases, the comparison can be based on a **reduced information** instead of the whole sample.
- Further, is highly unlikely that $\boldsymbol{S} \approx \boldsymbol{X}$ can be generated from $f(s \mid \boldsymbol{\theta})$ for any choice of $\boldsymbol{\theta}$.
- This results in the need of a **large scale parameter** λ in the kernel function, in order to achieve decent acceptance rates with the rejection algorithm.
- A common practice is to consider summary statistics of both observed and synthetic data, and then use those statistics in the comparison

$$\mathrm{K}_\lambda(||\boldsymbol{\phi}(X) - \boldsymbol{\phi}(S)||)$$

  where $\phi(\cdot) : \mathbb{X}^n \to \mathbb{R}^q$ is a function producing a vector of summaries of $\boldsymbol{X}$ or $\boldsymbol{S}$.
- This approach actually produce a sample from $\pi_{ABC}(\boldsymbol{\theta} \mid \phi(\boldsymbol{X}))$. However, if the vector of summary statistics is sufficient for the model parameters, then the approximating distribution corresponds to $\pi_{ABC}(\boldsymbol{\theta} \mid \boldsymbol{X})$.

ABC rejection sampler (better version)

At the rth sampling step

i) We generate $\boldsymbol{\theta}^{(r)} \overset{iid}{\sim} h(\boldsymbol{\theta})$ (independent of the previous state).

ii) We generate $\boldsymbol{S} \overset{iid}{\sim} f(s \mid \boldsymbol{\theta}^{(r)})$.

ii) Compute the synthetic data summary statistics $\psi(\boldsymbol{S})$.

iv) Accept $\boldsymbol{\theta}^{(r)}$ with probability

$$\frac{K_\lambda(||\phi(X) - \phi(S)||)\pi(\boldsymbol{\theta}^{(r)})}{Ch(\boldsymbol{\theta}^{(r)})}, \qquad \text{with} \qquad C \geq K_\lambda(0) \max_{\boldsymbol{\theta}} \frac{\pi(\boldsymbol{\theta}^{(r)})}{h(\boldsymbol{\theta}^{(r)})},$$

otherwise go back to i).

- The observed data summary statistics can be pre-computed once, before the sampling, saving computational time.
- The degree of approximation we are introducing is tuned by specific choices of λ, K and $||\cdot||$.

The target distribution

- In practice, we are sampling from the joint distribution

$$\pi(\boldsymbol{\theta}, \phi(\boldsymbol{S}) \mid \phi(\boldsymbol{X})) \propto \mathrm{K}_\lambda(||\boldsymbol{\phi}(X) - \boldsymbol{\phi}(S)||)\mathrm{L}(\phi(\boldsymbol{S}) \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}),$$

  and the posterior ABC distribution si defined as

$$\pi_{ABC}(\boldsymbol{\theta} \mid \phi(\boldsymbol{X})) = \int_{\mathbb{X}^n} \pi(\boldsymbol{\theta}, \phi(\boldsymbol{S}) \mid \phi(\boldsymbol{X}))\mathrm{d}\boldsymbol{S}.$$

- We can see that, as the scale parameter of the kernel decreases, we have

$$\begin{aligned}\lim_{\lambda \to 0} \pi_{ABC}(\boldsymbol{\theta} \mid \phi(\boldsymbol{X})) &\propto \int_{\mathbb{X}^n} \lim_{\lambda \to 0} \mathrm{K}_\lambda(||\boldsymbol{\phi}(X) - \boldsymbol{\phi}(S)||)\mathrm{L}(\phi(\boldsymbol{S}) \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{S} \\ &= \int_{\mathbb{X}^n} \delta_{\phi(\boldsymbol{X})}(\phi(\boldsymbol{S}))\mathrm{L}(\phi(\boldsymbol{S}) \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{S} = \pi(\boldsymbol{\theta} \mid \phi(\boldsymbol{X}))\end{aligned}$$

- If $\phi(\cdot)$ is sufficient or it is simply the whole sample, it is apparent that the previous corresponds to $\pi(\boldsymbol{\theta} \mid \boldsymbol{X})$.

How much approximate

- We can also quantify the accuracy of an ABC sampler. For simplicity of illustration, let us consider the a sampler without summary statistics, and an univariate case with a single data point, $x, s \in \mathbb{X} = \mathbb{R}$, and $||\cdot|| = |\cdot|$.

- We can see that the error committed in approximating the likelihood function (i.e. ignoring the prior distribution) corresponds to

$$\mathrm{L}_{ABC}(x \mid \boldsymbol{\theta}) = \int K_\lambda(|x - s|)\mathrm{L}(s \mid \boldsymbol{\theta})\mathrm{d}s = \int K(u)\mathrm{L}(x - u\lambda \mid \boldsymbol{\theta})\mathrm{d}u.$$

  Using a Taylor expansion, we have

$$\mathrm{L}_{ABC}(x \mid \boldsymbol{\theta}) = \int \mathrm{K}(u) \left[\mathrm{L}(x \mid \boldsymbol{\theta}) - u\lambda \frac{\mathrm{d}}{\mathrm{d}x}\mathrm{L}(x \mid \boldsymbol{\theta}) + \frac{u^2\lambda^2}{2}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\mathrm{L}(x \mid \boldsymbol{\theta}) - \cdots\right] \mathrm{d}u$$

  which can be truncated, with some simplification, in

$$\mathrm{L}_{ABC}(x \mid \boldsymbol{\theta}) \simeq \mathrm{L}(x \mid \boldsymbol{\theta}) + \frac{\lambda^2}{2}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\mathrm{L}(x \mid \boldsymbol{\theta}) \int u^2 K(u)\mathrm{d}u$$

- Hence, the bias we are committing is given by

$$b_\lambda(x \mid \boldsymbol{\theta}) = \frac{\lambda^2}{2}\sigma_{\mathrm{K}}^2 \frac{\mathrm{d}^2}{\mathrm{d}x^2}\mathrm{L}(x \mid \boldsymbol{\theta}),$$

  where we can see that is quadratic in λ and depends on the kernel choice through $\sigma_K^2 = \int u^2 \mathrm{K}(u)\mathrm{d}u$.

Interpretation

- There are many way to describe and interpret ABC samplers and their impact on statistical analysis.

- Commonly (e.g. Blum, 2010), in the ABC setting we are interested in the joint sample $(\boldsymbol{\theta}^{(r)}, \boldsymbol{S}^{(r)})$, and we produce an empirical conditional density of $\pi(\boldsymbol{\theta} \mid \boldsymbol{X})$ by weighting the $\boldsymbol{\theta}^{(r)}$s by $||\boldsymbol{S}^{(r)} - \boldsymbol{X}||$

- Fearnhead and Prangle (2012) noted that the ABC approximation of the posterior is a continuous mixture

$$\pi_{ABC}(\boldsymbol{\theta} \mid \boldsymbol{X}) \propto \int w(\boldsymbol{S})\pi(\boldsymbol{\theta} \mid \boldsymbol{S})\mathrm{d}\boldsymbol{S},$$

  where $w(\boldsymbol{S}) \propto \mathrm{K}_\lambda(||\boldsymbol{S} - \boldsymbol{X}||)\mathrm{m}(\boldsymbol{X})$, with $m(\boldsymbol{X})$ being the marginal distribution of $\boldsymbol{X}$.

- Wilkinson (2013) pointed out that ABC methods can be considered as exact if $||\boldsymbol{S} - \boldsymbol{X}||$ represents an error term (observational error or model misspecification), and K_λ is the error distribution.

ABC rejection sampler with Pima dataset

- We consider again the Pima dataset, but using a ABC rejection sampler.
- We use a $N_p(\mathbf{0}, 100\mathrm{I}_p)$ as prior distribution, $N_p(\boldsymbol{\mu}, S)$ as proposal, and as kernel, we consider $K_\lambda(||\boldsymbol{S} - \boldsymbol{X}||) = \left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{I}_{[S_i = X_i]}\right)^{\lambda}$.

```
# not informed
S <- diag(100, 8)
mu <- rep(0, 8)
lambda <- 3

# Running the ABCr (R = 3000)
ABCrejectionNI <- ABCrejection(R, lambda, y, X, S, mu)

colMeans(ABCrejectionNI$param)
# X       Xnpreg   Xglu    Xbp  Xskin   Xbmi   Xped   Xage
# -2.975  1.978  3.760  1.060  1.640  2.047  1.793  2.402

ABCrejectionNI$acc
#  0.152
```

ABC rejection sampler with Pima dataset

- Same but using the Laplace approximation as proposal.

```
# not informed
S <- vcov(fit_logit)
mu <- coefficients(fit_logit)
lambda <- 3

# Running the ABCr (R = 3000)
ABCrejectionNI <- ABCrejection(R, lambda, y, X, S, mu)

colMeans(ABCrejectionNI$param)
# X       Xnpreg   Xglu    Xbp  Xskin   Xbmi   Xped   Xage
# -0.991  0.405  1.097 -0.098  0.068  0.571  0.449  0.284

ABCrejectionNI$acc
#  0.995
```

Improving ABC methods

Motivation

- We can improve our sampler by combining the previous strategy with other sampling strategy, for example the one presented in the previous sessions.
- In general, we target the joint distribution of parameter vector and summary statistics

$$\pi(\boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{S}) \mid \boldsymbol{\phi}(\boldsymbol{X})) \propto \mathrm{K}_{\lambda}(||\boldsymbol{\phi}(\boldsymbol{S}) - \boldsymbol{\phi}(\boldsymbol{X})||)\mathrm{L}(\boldsymbol{S} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

- Samples of $\boldsymbol{\theta}$ can be obtained by sampling jointly $(\boldsymbol{\theta}, \boldsymbol{S})$ and marginalizing by discarding $\boldsymbol{S}$.
- Ideally, we will see ...

ABC importance sampling

- We discussed that importance sampling is a procedure that, in the spirit of rejection sampler, use an instrumental distribution to propose possible values of the quantity of interest. However, rather than calculate acceptance probabilities, produce a weighted sample.

- Once we draw a candidate $\boldsymbol{\theta}^{(r)} \sim q(\boldsymbol{\theta}^{(r)})$, we also compute a weight associated to this value, $w(\boldsymbol{\theta}^{(r)}) = \pi(\boldsymbol{\theta}^{(r)} \mid \boldsymbol{X})/q(\boldsymbol{\theta}^{(r)})$.

- From the previous, it is easy to see that

$$\mathrm{E}_q[w(\boldsymbol{\theta})g(\boldsymbol{\theta})] = \int g(\boldsymbol{\theta})w(\boldsymbol{\theta})q(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \int g(\boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{X})\mathrm{d}\boldsymbol{\theta} = \mathrm{E}_\pi[g(\boldsymbol{\theta})],$$

  which can be approximated via Monte Carlo methods.

- When the target distribution is not normalized, we simply used a normalized version of the weights

$$W^{(r)} = \frac{w(\boldsymbol{\theta}^{(r)})}{\sum_{r=1}^{R} w(\boldsymbol{\theta}^{(r)})}.$$

ABC importance sampling

- In the following slides, we do not consider the summary statistics case. However, everything here presented can be done also using summaries of the samples, both observed and synthetic.
- From an ABC perspective, importance sampling works similarly to rejection sampling.
- Our target distribution of the importance sampling is $\pi_{ABC}(\boldsymbol{\theta}, \boldsymbol{S} \mid \boldsymbol{X})$, and the proposal (importance) distribution is of the form $q(\boldsymbol{\theta}, \boldsymbol{S}) = \mathrm{L}(\boldsymbol{S} \mid \boldsymbol{\theta}) q(\boldsymbol{\theta})$, defined jointly on the parameter and sample spaces.
- As results, the unnormalized importance weights are

$$w(\boldsymbol{\theta}) \propto \frac{\pi_{ABC}(\boldsymbol{\theta}, \boldsymbol{S} \mid \boldsymbol{X})}{q(\boldsymbol{\theta}, \boldsymbol{S})} \propto \frac{\mathrm{K}_\lambda(||\boldsymbol{S} - \boldsymbol{X}||)\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})},$$

  which is free of the intractable likelihood term.
- The choice of the importance distribution $q(\boldsymbol{\theta})$ is crucial for the algorithm performances.

importance ABC in practice

At the rth sampling step

i) We generate $\boldsymbol{\theta}^{(r)} \overset{iid}{\sim} q(\boldsymbol{\theta})$ (independent of the previous state).

ii) We generate $S_i \overset{iid}{\sim} f(s \mid \boldsymbol{\theta}^{(r)})$, for $i = 1, \ldots, n$.

ii) Compute the importance weights

$$w(\boldsymbol{\theta}_r) = \frac{\mathrm{K}_\lambda(||\boldsymbol{S} - \boldsymbol{X}||)\pi(\boldsymbol{\theta}^{(r)})}{q(\boldsymbol{\theta}^{(r)})}.$$

- The result is a weighted sample, which can be possibly resampled.
- The kernel choice impacts on the resulting weights. Specifically, if the kernel has non-compact support, the importance weights are guaranteed to be always non-zero. However, this may results in high variability and low ESS of the sampler.

Importance ABC sampling with Pima dataset

- We consider again the Pima dataset, but using a ABC importance sampler.
- We use a $N_p(\mathbf{0}, 100\mathrm{I}_p)$ as importance distribution, $N_p(\boldsymbol{\mu}, S)$ as proposal, and as kernel, we consider $K_\lambda(||\boldsymbol{S} - \boldsymbol{X}||) = \left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{I}_{[S_i = X_i]}\right)^{\lambda}$.

```
# not informed
S <- diag(100, 8)
mu <- rep(0, 8)
lambda <- 3

# Running the ABCis (R = 3000)
ABCimportanceNI <- ABCimportance(R, lambda, y, X, S, mu)

colMeans(ABCrejectionNI$param)
# X       Xnpreg   Xglu    Xbp  Xskin   Xbmi   Xped   Xage
# -3.639  1.782  3.760  1.013  1.403  2.154  1.812  2.388

ABCrejectionNI$acc
#  ESS
```

Importance ABC sampling with Pima dataset

- Same but with Laplace approximation as importance distribution.

```
# not informed
S <- vcov(fit_logit)
mu <- coefficients(fit_logit)
lambda <- 3

# Running the ABCis (R = 3000)
ABCimportanceI <- ABCimportance(R, lambda, y, X, S, mu)

colMeans(ABCrejectionNI$param)
# X       Xnpreg   Xglu    Xbp  Xskin   Xbmi   Xped   Xage
# -1.064  0.348  1.089 -0.092  0.086  0.539  0.428  0.326

ABCrejectionNI$acc
#  14.263
```

ABC-MCMC

- As we saw earlier in the days, MCMC can be used to define sampling strategies. In particular, with some regularity assumptions, it is possible to produce suitable algorithm whose transition kernel produce a Markov Chain which is ergodic to a specific target distribution.

- Among others, a broad class of algorithm is given by the Metropolis-Hastings algorithm, where given a current state $\boldsymbol{\theta}^{(r)}$, we propose a candidate from $q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r-1)})$ and then accept such a value with probability

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)}) = \min \left\{ 1, \frac{\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) q(\boldsymbol{\theta}^{(r-1)} \mid \boldsymbol{\theta})}{\pi(\boldsymbol{\theta}^{(r-1)} \mid \boldsymbol{X}) q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r-1)})} \right\}.$$

- it is only natural that MCMC-based ABC algorithms have been studied, as they combine both the tractability of ABC methods, and the feature of MCMC algorithms, such as Markovian dependence over time.

- ABC MCMC algorithms were originally studied by Marjoram et al. (2003), an later extended in many directions, by considering different MCMC strategies.

ABC-MCMC

- In an ABC setting, the target distribution is $\pi_{ABC}(\boldsymbol{\theta}, \boldsymbol{S} \mid \boldsymbol{X})$, whereas possibly we marginalize $\boldsymbol{S}$.

- Hence, our proposal distribution works on the product space $\Theta \times \mathbb{X}^n$. In practice, we want to have (Markovian) memories only for the parameter values. Hence, our proposal usually has form

$$q(\boldsymbol{\theta}, \boldsymbol{S} \mid \boldsymbol{\theta}^{(r-1)}, \boldsymbol{S}^{(r-1)}) = q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r-1)}) \mathrm{L}(\boldsymbol{S} \mid \boldsymbol{\theta}),$$

  so that the synthetic data are sampled independently from the past.

- The resulting acceptance rate equals

$$\begin{aligned} \alpha(\boldsymbol{\theta}, \boldsymbol{S} \mid \boldsymbol{\theta}^{(r-1)}, \boldsymbol{S}^{(r-1)}) &= \min \left\{ 1, \frac{\pi_{ABC}(\boldsymbol{\theta}, \boldsymbol{S} \mid \boldsymbol{X}) q(\boldsymbol{\theta}^{(r-1)}, \boldsymbol{S}^{(r-1)} \mid \boldsymbol{\theta}, \boldsymbol{S})}{\pi_{ABC}(\boldsymbol{\theta}^{(r-1)}, \boldsymbol{S}^{(r-1)} \mid \boldsymbol{X}) q(\boldsymbol{\theta}, \boldsymbol{S} \mid \boldsymbol{\theta}^{(r-1)}, \boldsymbol{S}^{(r-1)})} \right\} \\ &= \min \left\{ 1, \frac{\mathrm{K}_\lambda(||\boldsymbol{S} - \boldsymbol{X}||) \pi(\boldsymbol{\theta}) q(\boldsymbol{\theta}^{(r-1)} \mid \boldsymbol{\theta})}{\mathrm{K}_\lambda(||\boldsymbol{S}^{(r-1)} - \boldsymbol{X}||) \pi(\boldsymbol{\theta}^{(r-1)}) q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r-1)})} \right\}. \end{aligned}$$

- Note that, the previous acceptance rate does not involve the intractable likelihood term.

ABC-MCMC in practice

At the rth sampling step

i) We generate $\boldsymbol{\theta} \overset{iid}{\sim} q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r-1)})$.

ii) We generate $S_i \overset{iid}{\sim} f(s \mid \boldsymbol{\theta})$, for $i = 1, \ldots, n$.

ii) With probability

$$\alpha(\boldsymbol{\theta}, \boldsymbol{S} \mid \boldsymbol{\theta}^{(r-1)}, \boldsymbol{S}^{(r-1)}) = \min \left\{ 1, \frac{\mathrm{K}_\lambda(||\boldsymbol{S} - \boldsymbol{X}||)\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}^{(r-1)} \mid \boldsymbol{\theta})}{\mathrm{K}_\lambda(||\boldsymbol{S}^{(r-1)} - \boldsymbol{X}||)\pi(\boldsymbol{\theta}^{(r-1)})q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r-1)})} \right\}$$

set $(\boldsymbol{\theta}^{(r)}, \boldsymbol{S}^{(r)}) = (\boldsymbol{\theta}, \boldsymbol{S})$, otherwise $(\boldsymbol{\theta}^{(r)}, \boldsymbol{S}^{(r)}) = (\boldsymbol{\theta}^{(r-1)}, \boldsymbol{S}^{(r-1)})$.

- The algorithm satisfies the detailed balanced condition on $\pi_{ABC}(\boldsymbol{\theta}, \boldsymbol{S} \mid \boldsymbol{X})$.
- The algorithm can be sensible to the initial condition.
- Thus, the strategy can be combined with specific proposal (e.g. MALA).