

# ADVANCED BAYESIAN LEARNING APPROXIMATE METHODS SPRING 2014

Mattias Villani

**Division of Statistics  
Department of Computer and Information Science  
Linköping University**

# TOPIC OVERVIEW

- ▶ Variational Bayes (VB)
- ▶ Approximate Bayesian Computations (ABC)

# VARIATIONAL BAYES

- ▶ Let  $\theta = (\theta_1, \dots, \theta_M)$ . Approximate the posterior  $p(\theta|y)$  with a (simpler) distribution  $q(\theta)$ .

- ▶ **Nonparametric/Factorization/Mean field approximation**

$$q(\theta) = \prod_{i=1}^M q_i(\theta_i)$$

- ▶ **Parametric**, where  $q_\lambda(\theta)$  is a parametric family with parameters  $\lambda$ .
- ▶ Find the  $q(\theta)$  that **minimizes the Kullback-Leibler distance**:

$$KL(p, q) = \int p(\theta|y) \ln \frac{p(\theta|y)}{q(\theta)} d\theta = E_p \left[ \ln \frac{p(\theta|y)}{q(\theta)} \right].$$

- ▶ Computing the expectation wrt  $p(\theta|y)$  is often hard.
- ▶ **Reverse KL** problem is often simpler (but somewhat unnatural):

$$KL(q, p) = \int q(\theta) \ln \frac{q(\theta)}{p(\theta|y)} d\theta = E_q \left[ \ln \frac{q(\theta)}{p(\theta|y)} \right].$$

## VB GIVES A LOWER BOUND ON $p(\mathbf{y})$

- ▶ Using that  $p(\theta|\mathbf{y}) = p(\mathbf{y}, \theta) / p(\mathbf{y})$  we have

$$\begin{aligned} KL(q, p) &= \int q(\theta) \ln \frac{q(\theta)}{p(\theta|\mathbf{y})} d\theta \\ &= \int q(\theta) \ln \left( \frac{q(\theta)}{p(\mathbf{y}, \theta)} \right) d\theta + \int q(\theta) \ln p(\mathbf{y}) d\theta \\ &= \int q(\theta) \ln \frac{q(\theta)}{p(\mathbf{y}, \theta)} d\theta + \ln p(\mathbf{y}) \end{aligned}$$

- ▶ Since  $KL(q, p) \geq 0$ , we have the following **lower bound** for  $\ln p(\mathbf{y})$

$$\ln p(\mathbf{y}) \geq - \int q(\theta) \ln \frac{q(\theta)}{p(\mathbf{y}, \theta)} d\theta = \int q(\theta) \ln \frac{p(\mathbf{y}, \theta)}{q(\theta)} d\theta \stackrel{\text{def}}{=} \ln \underline{p}(\mathbf{y}; q),$$

where  $p(\mathbf{y}, \theta) = p(\mathbf{y}|\theta)p(\theta)$  is the unnormalized posterior.

- ▶ Minimizing  $KL(q, p)$  is the same as maximizing  $\ln \underline{p}(\mathbf{y}; q)$ .

# MEAN FIELD APPROXIMATION

- ▶ Factorization

$$q(\theta) = \prod_{i=1}^M q_i(\theta_i)$$

- ▶ **No functional forms are assumed** for the  $q_i(\theta)$ . Nonparametric.
- ▶ **Optimal densities** can be shown to satisfy:

$$q_i(\theta) \propto \exp(E_{-\theta_i} \ln p(\mathbf{y}, \theta))$$

where  $E_{-\theta_i}(\cdot)$  is the expectation with respect to  $\prod_{i \neq j} q_j(\theta_j)$ .

- ▶ Alternative formulation that **connects to Gibbs sampling**

$$q_i(\theta_i) \propto \exp(E_{-\theta_i} \ln p(\theta_i | \text{rest}))$$

where  $p(\theta_i | \text{rest})$  is the full conditional posterior of  $\theta_i$ .

- ▶ **Structured mean field approximation.** Group subset of parameters in tractable blocks.

# MEAN FIELD APPROXIMATION - ALGORITHM

- ▶ Initialize:  $q_2^*(\theta_2), \dots, q_M^*(\theta_M)$
- ▶ Repeat until increase in  $\ln \underline{p}(\mathbf{y}; q)$  is negligible:
  - ▶  $q_1^*(\theta_1) \leftarrow \frac{\exp[E_{-\theta_1} \ln p(\mathbf{y}, \theta)]}{\int \exp[E_{-\theta_1} \ln p(\mathbf{y}, \theta)] d\theta_1}$
  - ▶  $\vdots$
  - ▶  $q_M^*(\theta_M) \leftarrow \frac{\exp[E_{-\theta_M} \ln p(\mathbf{y}, \theta)]}{\int \exp[E_{-\theta_M} \ln p(\mathbf{y}, \theta)] d\theta_M}$
- ▶ Note: we make no assumptions about parametric form of the  $q_i(\theta)$ , but the optimal  $q_i(\theta)$  often turn out to be parametric (normal, gamma etc).
- ▶ The updates above then boil down to just updating of hyperparameters in the optimal densities.

# MEAN FIELD APPROXIMATION - NORMAL MODEL

- ▶ **Model:**  $X_i|\mu, \sigma^2 \stackrel{iid}{\sim} N(\mu, \sigma^2)$ .
- ▶ **Prior:**  $\mu \sim N(\mu_\mu, \sigma_\mu^2)$  **independent** of  $\sigma^2 \sim IG(A, B)$ .
- ▶ Note: this is NOT the conjugate prior.
- ▶ **Variational approximation:**  $q(\mu, \sigma^2) = q_\mu(\mu) \cdot q_{\sigma^2}(\sigma^2)$ .
- ▶ Optimal densities

$$q_\mu^*(\mu) \propto \exp \left[ E_{q(\sigma^2)} \ln p(\mu|\sigma^2, \mathbf{x}) \right]$$

$$q_{\sigma^2}^*(\sigma^2) \propto \exp \left[ E_{q(\mu)} \ln p(\sigma^2|\mu, \mathbf{x}) \right]$$

- ▶ Full conditional posteriors

$$\mu|\sigma^2, \mathbf{x} \sim N \left( \frac{n\bar{x}/\sigma^2 + \mu_\mu/\sigma_\mu^2}{n/\sigma^2 + 1/\sigma_\mu^2}, \frac{1}{n/\sigma^2 + 1/\sigma_\mu^2} \right)$$

$$\sigma^2|\mu, \mathbf{x} \sim IG \left( A + \frac{n}{2}, B + \frac{1}{2}(\mathbf{x} - \mu)'(\mathbf{x} - \mu) \right)$$

## NORMAL MODEL EXAMPLE - UPDATING $q_{\sigma^2}^*(\sigma^2)$

- Full conditional posterior of  $\sigma^2$

$$\sigma^2 | \mu, \mathbf{x} \sim IG \left( A + \frac{n}{2}, B + \frac{1}{2}(\mathbf{x} - \mu)'(\mathbf{x} - \mu) \right)$$

- So,  $E_{q(\mu)} \ln p(\sigma^2 | \mu, \mathbf{x})$  is proportional to

$$E_{q(\mu)} \left[ - \left( A + \frac{n}{2} + 1 \right) \ln \sigma^2 - \left( B + \frac{1}{2}(\mathbf{x} - \mu)'(\mathbf{x} - \mu) \right) / \sigma^2 \right]$$

and therefore

$$q_{\sigma^2}^*(\sigma^2) \propto (\sigma^2)^{-(A+n/2+1)} \exp \left( - \left( B + \frac{1}{2} E_{q(\mu)}(\mathbf{x} - \mu)'(\mathbf{x} - \mu) / \sigma^2 \right) \right)$$

which shows that  $q_{\sigma^2}^*(\sigma^2)$  is

$$IG \left( A + \frac{n}{2}, B + \frac{1}{2} E_{q(\mu)}(\mathbf{x} - \mu)'(\mathbf{x} - \mu) \right)$$



## NORMAL MODEL EXAMPLE - UPDATING $q_{\sigma^2}^*(\sigma^2)$

- So  $q_{\sigma^2}^*(\sigma^2)$  is

$$IG \left( A + \frac{n}{2}, B + \frac{1}{2} E_{q(\mu)}(\mathbf{x} - \mu)'(\mathbf{x} - \mu) \right)$$

and

$$E_{q(\mu)}(\mathbf{x} - \mu)'(\mathbf{x} - \mu) =$$

$$E_{q(\mu)} \left[ \left( \mathbf{x} - E_{q(\mu)}(\mu) \right) + \left( E_{q(\mu)}(\mu) - \mu \right) \mathbf{1}_n \right]' \left[ \left( \mathbf{x} - E_{q(\mu)}(\mu) \right) + \left( E_{q(\mu)}(\mu) - \mu \right) \mathbf{1}_n \right]$$

$$= E_{q(\mu)} \left[ \left( \mathbf{x} - E_{q(\mu)}(\mu) \right)' \left( \mathbf{x} - E_{q(\mu)}(\mu) \right) + \left( E_{q(\mu)}(\mu) - \mu \right)^2 n \right]$$

$$= \left( \mathbf{x} - E_{q(\mu)}(\mu) \right)' \left( \mathbf{x} - E_{q(\mu)}(\mu) \right) + n \cdot \text{Var}_{q(\mu)}(\mu)$$

because  $E_{q(\mu)} \left( E_{q(\mu)}(\mu) - \mu \right) = 0$ .

- Important:  $E_{q(\mu)}(\mu)$  and  $\text{Var}_{q(\mu)}(\mu)$  is the mean and variance of the  $q_\mu$  distribution.

## NORMAL MODEL EXAMPLE - UPDATING $q_{\mu}^*(\mu)$

- Easier to go back to the form  $q_{\mu}(\mu) \propto \exp \left( E_{q(\sigma^2)} \ln p(\mathbf{y}, \mu, \sigma^2) \right)$

$$\begin{aligned} \ln p(\mathbf{y}, \mu, \sigma^2) &= \ln p(\mathbf{y}|\mu, \sigma^2) + \ln p(\mu) + \ln p(\sigma^2) \\ &= -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 - \frac{1}{2\sigma_{\mu}^2} (\mu - \mu_{\mu})^2 + \text{const} \end{aligned}$$

$$q_{\mu}(\mu) \propto \exp \left( -E_{q(\sigma^2)} \left( \frac{1}{\sigma^2} \right) \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 - \frac{1}{2\sigma_{\mu}^2} (\mu - \mu_{\mu})^2 \right)$$

- Completing the square shows

$$q_{\mu}(\mu) = N \left( E_{q(\mu)}(\mu), \text{Var}_{q(\mu)}(\mu) \right)$$

where

$$E_{\mu}(\mu) = \frac{n\bar{x} E_{q(\sigma^2)} \left( \frac{1}{\sigma^2} \right) + \mu_{\mu} / \sigma_{\mu}^2}{n E_{q(\sigma^2)} \left( \frac{1}{\sigma^2} \right) + 1 / \sigma_{\mu}^2}$$

$$\text{Var}_{\mu}(\mu) = \frac{1}{n E_{q(\sigma^2)} \left( \frac{1}{\sigma^2} \right) + 1 / \sigma_{\mu}^2}$$

# NORMAL MODEL EXAMPLE - SUMMARY

- Variational density for  $\sigma^2$

$$q_{\sigma^2}(\sigma^2) = IG(A_q, B_q)$$

where  $A_q = A + n/2$  and  $B_q = B + \frac{1}{2} \left( \|\mathbf{x} - \mu_q \cdot \mathbf{1}_n\|^2 + n \cdot \sigma_q^2 \right)$

- Variational density for  $\mu$

$$q_{\mu}(\mu) = N(\mu_q, \sigma_q^2)$$

where

$$\sigma_q^2 = \frac{1}{n \frac{A_q}{B_q} + 1/\sigma_{\mu}^2}$$

$$\mu_q = \left( n\bar{x} \frac{A_q}{B_q} + \mu_{\mu}/\sigma_{\mu}^2 \right) \sigma_q^2$$

## NORMAL MODEL EXAMPLE - ALGORITHM

- ▶ Set  $A_q = A + n/2$ .
- ▶ Initialize  $E_\mu(\mu) = \bar{x}$  and  $Var_\mu(\mu) = s^2/n$ .
- ▶ Repeat

- ▶
$$B_q \leftarrow B + \frac{1}{2} \left( \|\mathbf{x} - \mu_q \cdot \mathbf{1}_n\|^2 + n \cdot \sigma_q^2 \right)$$

- ▶
$$\sigma_q^2 \leftarrow \frac{1}{n \frac{A_q}{B_q} + 1/\sigma_\mu^2}$$

- ▶
$$\mu_q \leftarrow \left( n\bar{x} \frac{A_q}{B_q} + \mu_\mu / \sigma_\mu^2 \right) \sigma_q^2$$

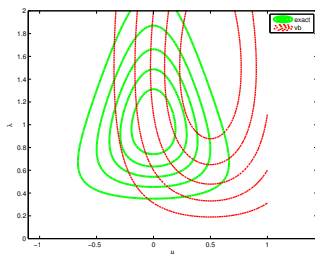
- ▶ Until the change in

$$\ln \underline{p}(\mathbf{x}; q) = \frac{1}{2} - \frac{n}{2} \log(2\pi) + \frac{1}{2} \ln \left( \sigma_q^2 / \sigma_\mu^2 \right) - \frac{(\mu_q - \mu_\mu)^2 + \sigma_q^2}{2\sigma_\mu^2}$$

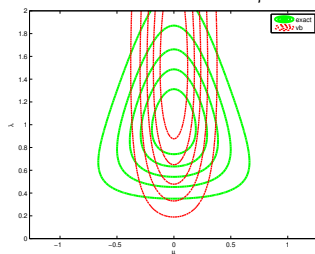
is negligible.

# NORMAL EXAMPLE FROM MURPHY ( $\lambda^{-1} = \sigma^2$ )

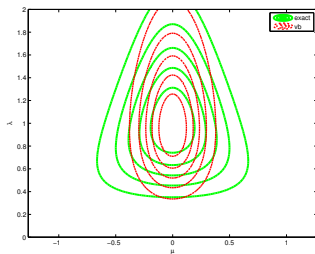
Initial values



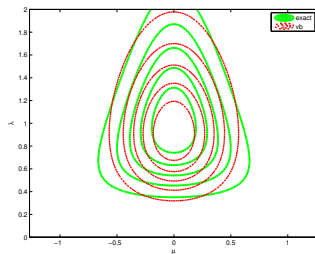
After updating  $q_\mu$



After updating  $q_{\sigma^2}$



At convergence



# PROBIT REGRESSION

- ▶ **Model:**

$$Y_i | x_i \stackrel{ind.}{\sim} \text{Bern} [\Phi(x_i' \beta)]$$

- ▶ **Prior:**  $\beta \sim N(\mu_\beta, \Sigma_\beta)$

- ▶ **Latent variable formulation** with  $\mathbf{a} = (a_1, \dots, a_n)'$

$$\mathbf{a} | \beta \sim N(X\beta, 1)$$

and

$$p(y_i | a_i) = I(a_i \geq 0)^{y_i} I(a_i < 0)^{1-y_i}$$

- ▶ **Factorized variational approximation**

$$q(\mathbf{a}, \beta) = q_{\mathbf{a}}(\mathbf{a}) q_{\beta}(\beta)$$

# PROBIT REGRESSION - UPDATING **A**

## ► Log joint distribution

$$\begin{aligned}\ln p(y, \beta, \mathbf{a}) &= \ln p(y|\beta, \mathbf{a}) + \ln p(\mathbf{a}|\beta) + \ln p(\beta) \\ &\propto \sum_{i=1}^n (y_i \ln I(a_i \geq 0) + (1 - y_i) \ln I(a_i < 0)) \\ &\quad - \frac{1}{2}(\mathbf{a} - \mathbf{X}\beta)'(\mathbf{a} - \mathbf{X}\beta) - \frac{1}{2}(\beta - \mu_\beta)' \Sigma_\beta (\beta - \mu_\beta)\end{aligned}$$

## ► Updating **a**

$$\begin{aligned}\ln q_{\mathbf{a}}(\mathbf{a}) &\propto E_{q(\beta)} \ln p(y, \beta, \mathbf{a}) \propto \sum_{i=1}^n (y_i \ln I(a_i \geq 0) + (1 - y_i) \ln I(a_i < 0)) \\ &\quad - \frac{1}{2} E_{q(\beta)} (\mathbf{a} - \mathbf{X}\beta)'(\mathbf{a} - \mathbf{X}\beta) + \text{const}\end{aligned}$$

Note that

$$\begin{aligned}E_{q(\beta)} (\mathbf{a} - \mathbf{X}\beta)'(\mathbf{a} - \mathbf{X}\beta) &= \mathbf{a}'\mathbf{a} - 2\mathbf{a}'\mathbf{X}E_{q(\beta)}(\beta) + \text{const} \\ &= (\mathbf{a} - \mathbf{X}E_{q(\beta)}(\beta))'(\mathbf{a} - \mathbf{X}E_{q(\beta)}(\beta)) + \text{const}\end{aligned}$$

## PROBIT REGRESSION - UPDATING **A**

► So

$$q_{\mathbf{a}}(\mathbf{a}) \propto \prod_{i=1}^n I(a_i \geq 0)^{y_i} I(a_i < 0)^{1-y_i} \\ \times \exp \left( -\frac{1}{2} (\mathbf{a} - \mathbf{X} \mu_{q(\beta)})' (\mathbf{a} - \mathbf{X} \mu_{q(\beta)}) \right)$$

► Normalizing gives

$$q_{\mathbf{a}}(\mathbf{a}) = \prod_{i=1}^n \left[ \frac{I(a_i \geq 0)}{\Phi \left( (\mathbf{X} \mu_{q(\beta)})_i \right)} \right]^{y_i} \left[ \frac{I(a_i < 0)}{1 - \Phi \left( (\mathbf{X} \mu_{q(\beta)})_i \right)} \right]^{1-y_i} \\ \times (2\pi)^{-n/2} \exp \left( -\frac{1}{2} \left\| \mathbf{a} - \mathbf{X} \mu_{q(\beta)} \right\|^2 \right)$$

which is a product of  $n$  truncated normals.



# PROBIT REGRESSION - UPDATING $\beta$

- ▶ Updating  $\beta$

$$\begin{aligned}\ln q_{\beta}(\beta) &\propto E_{q(\mathbf{a})} \ln p(y, \beta, \mathbf{a}) \\ &\propto -\frac{1}{2} E_{q(\mathbf{a})} (\mathbf{a} - \mathbf{X}\beta)' (\mathbf{a} - \mathbf{X}\beta) - \frac{1}{2} (\beta - \mu_{\beta})' \Sigma_{\beta} (\beta - \mu_{\beta})\end{aligned}$$

- ▶ For any random vector  $\mathbf{y}$  with mean  $\mu$  and covariance matrix  $\Omega$

$$E(\mathbf{y} - \mathbf{m})'(\mathbf{y} - \mathbf{m}) = \text{trace}(\Omega) + (\mu - \mathbf{m})'(\mu - \mathbf{m})$$

- ▶ So

$$\begin{aligned}\ln q_{\beta}(\beta) &\propto -\frac{1}{2} \text{trace}(\Sigma_{\mathbf{a}}) - \frac{1}{2} (\mu_{q(\mathbf{a})} - \mathbf{X}\beta)' (\mu_{q(\mathbf{a})} - \mathbf{X}\beta) \\ &\quad - \frac{1}{2} (\beta - \mu_{\beta})' \Sigma_{\beta} (\beta - \mu_{\beta})\end{aligned}$$

- ▶ The variational approximation of  $\beta$  is like the posterior from regressing  $\mu_{q(\mathbf{a})}$  on  $\mathbf{X}$  with prior  $\beta \sim N(\mu_{\beta}, \Sigma_{\beta})$ .

# PROBIT REGRESSION - UPDATING $\beta$

- We therefore have

$$q_{\beta}(\beta) = N\left(\mu_{q(\beta)}, \left(\mathbf{X}'\mathbf{X} + \Sigma_{\beta}^{-1}\right)^{-1}\right)$$

and

$$\mu_{q(\beta)} = \left(\mathbf{X}'\mathbf{X} + \Sigma_{\beta}^{-1}\right)^{-1} \left(\mathbf{X}'\mu_{q(\mathbf{a})} + \Sigma_{\beta}^{-1}\mu_{\beta}\right)$$

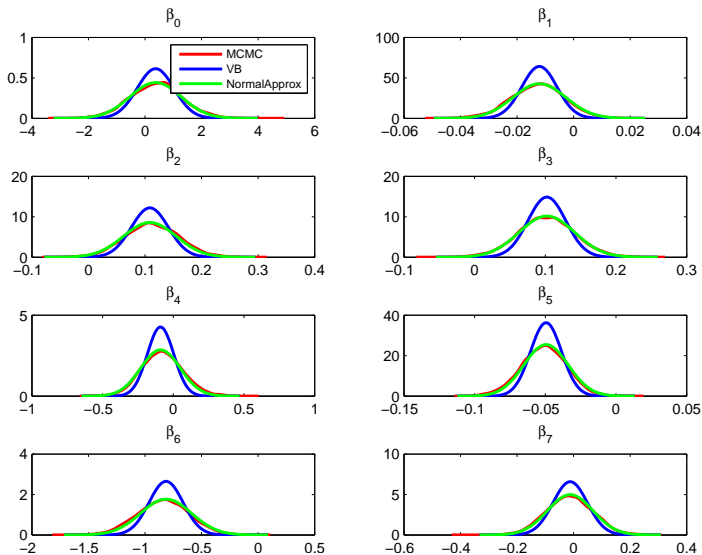
where

$$\mu_{q(\mathbf{a})} = X\mu_{q(\beta)} + \frac{\phi\left(X\mu_{q(\beta)}\right)}{\Phi\left(X\mu_{q(\beta)}\right)^{\mathbf{y}} \left[\Phi\left(X\mu_{q(\beta)}\right) - \mathbf{1}_n\right]^{\mathbf{1}_n - \mathbf{y}}}$$

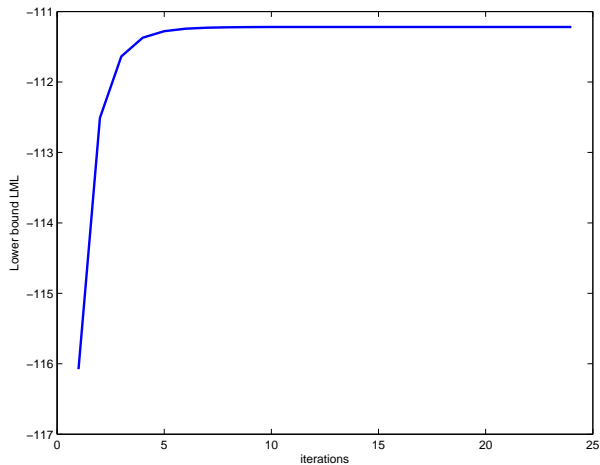
which follows from the expected value formula for a truncated distribution.

- The lower bound  $\ln \underline{p}(\mathbf{y}; q)$  is given in Ormerod and Wand (2010) where also the complete algorithm is given as Algorithm 4.

# PROBIT EXAMPLE (N=200 OBSERVATIONS)



# PROBIT EXAMPLE (N=200 OBSERVATIONS)



# VARIATIONAL BAYES EM (VBEM)

- ▶ Latent variable models:
  - ▶ Model parameters,  $\theta = (\theta_1, \dots, \theta_p)$
  - ▶ Latent variables,  $\mathbf{z} = (z_1, \dots, z_n)$
- ▶ Examples:
  - ▶ Mixture models:  $z_i \in \{1, \dots, K\}$  is the mixture allocation for the  $i$ th observation.
  - ▶ Missing data:  $\mathbf{z}$  contains the missing values.
- ▶ EM:
  - ▶ Get rid of  $\mathbf{z}$  by computing the expected log-likelihood  $E_{\mathbf{z}|\theta} \ln L(\theta, \mathbf{z})$  (E-step)
  - ▶ Maximize  $E_{\mathbf{z}|\theta} \ln L(\theta, \mathbf{z})$  wrt to  $\theta$  (M-step)
- ▶ VBEM approximation of posterior

$$p(\theta, \mathbf{z} | \mathbf{y}) \approx q(\theta)q(\mathbf{z}) = q(\theta) \prod_{i=1}^n q(z_i)$$

- ▶ Improves on EM by modelling the uncertainty in  $\theta$ .

## INDIRECT USE OF VB

- ▶ VB can play a role as **initial values** for MCMC.
- ▶ **Variational MCMC**. Use VB to construct Metropolis-Hastings proposal. Need to combine it with Metropolis random walk moves since VB typically underestimates the posterior variance.
- ▶ VB is fast and accurate for computing **log predictive scores (LPS)**

$$\sum_{t=T+1}^{T^*} \ln p(y_t | y_{t-1}^H) = \sum_{t=T+1}^{T^*} \int \ln p(y_t | y_{t-1}^H, \theta) p(\theta | y_{t-1}^H) d\theta$$

- ▶ VB is **fast** in this setting since it can approximate each sequential posterior  $p(\theta | y_{t-1}^H)$  using the mode of  $\hat{p}(\theta | y_{t-2}^H)$  as excellent initial values.
- ▶ VB seems **accurate** for approximating LPS, at least when the prediction uncertainty is mainly dominated by the future error uncertainty and not by parameter uncertainty.

# APPROXIMATE BAYESIAN COMPUTATIONS (ABC)

- ▶ Suitable when the likelihood is very costly or even infeasible to compute, but simulation from the model is cheap.
- ▶ Examples:
  - ▶ Likelihood is given by an **intractable high-dimensional integral**

$$\ell(\theta|\mathbf{y}) = \int \ell^*(\theta|\mathbf{y}, \mathbf{u}) d\mathbf{u}.$$

- ▶ **Normalizing constant**  $Z_\theta$  is **costly** or intractable

$$\ell(\theta|\mathbf{y}) = \ell_1(\theta|\mathbf{y}) / Z_\theta$$

- ▶ Likelihood is unavailable because the **PDF does not exist in closed form**.  $\alpha$ -stable distributions.
- ▶ ABC is often very crude.
- ▶ Arbitrary (creative) choices needed when implementing it.
- ▶ Early days, likely to improve.

# LIKELIHOOD-FREE REJECTION SAMPLER 1

- ▶ Idea:  $\theta$ 's with large posterior should generate data  $\mathbf{z}$  that look like the actual data  $\mathbf{y}$ .
- ▶ Assume the data  $\mathbf{y}$  takes values in a finite or countable set  $\mathcal{D}$ .
- ▶ **for**  $i = 1$  to  $N$  do
  - ▶ **repeat**
    - ▶ Generate  $\theta'$  from the prior distribution  $\pi(\cdot)$
    - ▶ Generate  $\mathbf{z}$  from the data distribution  $f(\cdot|\theta')$
  - ▶ **until**  $\mathbf{z} = \mathbf{y}$
  - ▶ set  $\theta_i = \theta'$
- ▶ **end for**
- ▶ Algorithm 1 produces a sample  $\theta_1, \dots, \theta_N$  from the posterior  $\pi(\theta|\mathbf{y})$ :

$$f(\theta_i) \propto \sum_{\mathbf{z} \in \mathcal{D}} \pi(\theta_i) f(\mathbf{z}|\theta_i) \mathbb{I}_{\mathbf{y}}(\mathbf{z}) = \pi(\theta_i) f(\mathbf{y}|\theta_i) \propto \pi(\theta_i|\mathbf{y}).$$



## LIKELIHOOD-FREE REJECTION SAMPLER 2

- ▶ Extension to continuous sample spaces where  $Pr(\mathbf{z} = \mathbf{y}|\theta) = 0$ .
- ▶ Define summary statistics  $\eta(\mathbf{z})$  and a distance function  $\rho[\eta(\mathbf{z}), \eta(\mathbf{y})]$ .
- ▶ **for**  $i = 1$  to  $N$  do
  - ▶ **repeat**
    - ▶ Generate  $\theta'$  from the prior distribution  $\pi(\cdot)$
    - ▶ Generate  $\mathbf{z}$  from the data distribution  $f(\cdot|\theta')$
  - ▶ **until**  $\rho[\eta(\mathbf{z}), \eta(\mathbf{y})] \leq \varepsilon$
  - ▶ set  $\theta_i = \theta'$
- ▶ **end for**
- ▶ **Algorithmic choices:**
  - ▶  $\eta$  - a function on  $\mathcal{D}$  defining a summary statistic (close to sufficient)
  - ▶  $\rho > 0$ , a distance on  $\eta(\mathcal{D})$
  - ▶  $\varepsilon > 0$ , a tolerance level.

## LIKELIHOOD-FREE REJECTION SAMPLER 2, CONT.

- ▶ The algorithm samples from the joint distribution

$$\pi_{\varepsilon}(\theta, \mathbf{z}|\mathbf{y}) = \frac{\pi(\theta)f(\mathbf{z}|\theta)\mathbb{I}_{A_{\varepsilon,\mathbf{y}}}(\mathbf{z})}{\int_{A_{\varepsilon,\mathbf{y}} \times \Theta} \pi(\theta)f(\mathbf{z}|\theta)d\mathbf{z}d\theta}$$

where

$$A_{\varepsilon,\mathbf{y}} = \{\mathbf{z} \in \mathcal{D} | \rho[\eta(\mathbf{z}), \eta(\mathbf{y})] \leq \varepsilon\}$$

- ▶ The hope is that

$$\pi(\theta|\mathbf{y}) \approx \pi_{\varepsilon}(\theta|\mathbf{y}) = \int \pi_{\varepsilon}(\theta, \mathbf{z}|\mathbf{y})d\mathbf{z}$$

is a good approximation.

- ▶ “*The basic idea behind ABC is that using a representative (enough) summary statistic  $\eta$  coupled with a small (enough) tolerance  $\varepsilon$  should produce a good (enough) approximation to the posterior distribution*” (Marin et al, 2012).

# ABC - AN EXAMPLE

- ▶  $MA(q)$  model. Fairly complicated likelihood. Easy to simulate time series from  $MA(q)$ .
- ▶ Summary statistics:
  - ▶ Raw distance between time series:
$$\rho[(z_1, \dots, z_n), (y_1, \dots, y_n)] = \sqrt{\sum_{i=1}^n (y_i - z_i)^2}$$
  - ▶ Distance between estimated autocorrelation functions:
$$\sum_{j=1}^K (\tau_{y,j} - \tau_{z,j})^2.$$

# MCMC - ABC

- ▶ Likelihood-free rejection sampler 2 is inefficient since it proposes  $\theta$ 's from the prior  $\pi(\theta)$ , which is often far from the posterior (with informative data).
- ▶ Initialize  $(\theta^{(0)}, \mathbf{z}^{(0)})$
- ▶ **for**  $i = 1$  to  $N$  do
  - ▶ Propose  $\theta'$  from the Markov kernel  $q(\cdot|\theta^{(t-1)})$
  - ▶ Generate  $\mathbf{z}'$  from the data distribution  $f(\cdot|\theta')$
  - ▶ Generate  $u$  from  $\mathcal{U}_{[0,1]}$
  - ▶ **if**  $u \leq \frac{\pi(\theta')q(\theta^{(t-1)}|\theta')}{\pi(\theta^{(t-1)})q(\theta'|\theta^{(t-1)})}$  and  $\rho[\eta(\mathbf{z}'), \eta(\mathbf{y})] \leq \varepsilon$  **then**
    - ▶ set  $(\theta^{(t)}, \mathbf{z}^{(t)}) = (\theta', \mathbf{z}')$
  - ▶ **else**
    - ▶ set  $(\theta^{(t)}, \mathbf{z}^{(t)}) = (\theta^{(t-1)}, \mathbf{z}^{(t-1)})$
  - ▶ **end if**
- ▶ **end for**

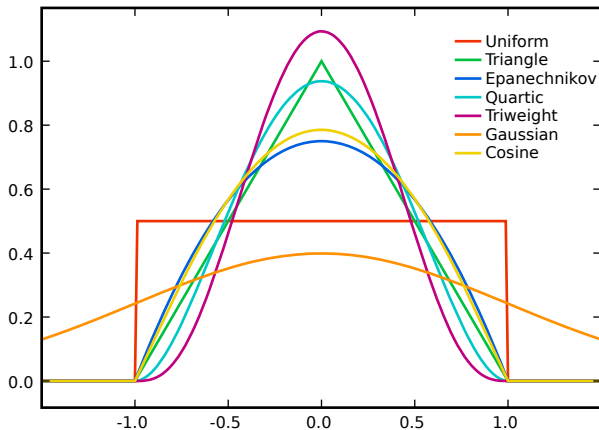
# NOISY ABC

- ▶ Replacing the crude  $\rho[\eta(\mathbf{z}), \eta(\mathbf{y})] \leq \varepsilon$  rejection rule with a smoother version

$$\pi_\varepsilon(\theta, \mathbf{z}|\mathbf{y}) = \frac{\pi(\theta)f(\mathbf{z}|\theta)K_\varepsilon(\mathbf{y} - \mathbf{z})}{\int \pi(\theta)f(\mathbf{z}|\theta)K_\varepsilon(\mathbf{y} - \mathbf{z})d\mathbf{z}d\theta}$$

- ▶  $K_\varepsilon(\cdot)$  is a kernel (think normal density) parametrized by the bandwidth  $\varepsilon > 0$  (think variance).
- ▶ The Bayes estimator from  $\pi_\varepsilon(\theta|\mathbf{y})$  is converging to the true value when  $n \rightarrow \infty$  and  $\varepsilon \rightarrow 0$ .
- ▶ See Wilkinson (2008) for details about the algorithm.

# SOME COMMON KERNELS



# CHOOSING THE ALGORITHMIC SETTINGS IN ABC

- ▶ Summary statistics  $\eta(\cdot)$  should be nearly sufficient. Which ones? Creativity ...
- ▶ Choice of  $\eta(\cdot)$  is crucial.
- ▶ Choice of  $\varepsilon$  is less important. Smaller  $\varepsilon$  gives better approximation at higher computational cost.
- ▶ Common choice of  $\varepsilon$ : small (0.1% or so) percentile of simulated distances  $\rho[\eta(\mathbf{z}'), \eta(\mathbf{y})]$ .

## POST-PROCESSING OF ABC OUTPUT

- ▶ Same algorithms, but allowing for larger  $\varepsilon$  by post-processing the ABC output.
- ▶ Keep **all**  $\theta$  draws regardless of how far  $\mathbf{z}$  is from the actual data  $\mathbf{y}$ , but shrink the  $\theta$  draws using

$$\theta_* = \theta - (\eta(\mathbf{z}) - \eta(\mathbf{y}))' \hat{\beta}$$

- ▶  $\hat{\beta}$  is obtained from a local kernel regression of  $\theta$  on  $\rho[\eta(\mathbf{z}), \eta(\mathbf{y})]$  with weights given by the kernel

$$K_\delta \{ \rho[\eta(\mathbf{z}), \eta(\mathbf{y})] \}.$$

- ▶ Kernel bandwidth  $\delta$  can for example be set equal to ABC tolerance  $\varepsilon$ .
- ▶ Alternative: heteroscedastic nonlinear regression.



# ABC FOR MODEL CHOICE

- ▶ You are entertaining a **set of  $M$  different** (competing, possibly non-nested) **models**.
- ▶ Let  $\mathcal{M}$  denote the unknown true model, and let  $\pi(\mathcal{M} = m)$  denote the **prior distribution** over the **model space**.
- ▶ The **Bayesian solution for model inference** is the posterior distribution:  $\pi(\mathcal{M} = m | \mathbf{y})$ .
- ▶ **ABC solution**: include  $\mathcal{M}$  in the set of parameters.
- ▶ Let  $\boldsymbol{\eta}(\mathbf{z}) = (\eta_1(\mathbf{z}), \dots, \eta_M(\mathbf{z}))$  be the concatenation of the **summary statistics** used for all models.

# ABC ALGORITHM FOR MODEL CHOICE

- ▶ **for**  $i = 1$  to  $N$  **do**
  - ▶ **repeat**
    - ▶ Generate  $m$  from the prior  $\pi(\mathcal{M} = m)$
    - ▶ Generate  $\theta_m$  from the prior  $\pi_m(\theta_m)$
    - ▶ Generate  $\mathbf{z}$  from the data distribution  $f_m(\cdot|\theta_m)$
  - ▶ **until**  $\rho[\eta(\mathbf{z}), \eta(\mathbf{y})] \leq \varepsilon$
  - ▶ Set  $m^{(i)} = m$  and  $\theta^{(i)} = \theta_m$
- ▶ **end for**
- ▶ **ABC estimate**

$$\pi(\mathcal{M} = m|\mathbf{y}) \approx \frac{1}{N} \sum_{i=1}^N \mathbb{I}_{m^{(i)}=m}$$

- ▶ **Example - comparing:**
  - ▶ AR-processes (homoscedastic variance)
  - ▶ GARCH models (volatility clustering)