



UiO : **Matematisk institutt**

Det matematisk-naturvitenskapelige fakultet

STK-4051/9051 Computational Statistics Spring 2020 **Variational Inference**

Instructor: Odd Kolbjørnsen, oddkol@math.uio.no

Slides partly made by Geir Olve Storvik



Exam spring 2020

- *The exam is run as a 7-day individual home exam with all aids allowed. The tasks will be similar to a mandatory task. Collaboration will not be considered cheating, but you must have formulated and written the answer that is submitted, and it should reflect your understanding of the syllabus.). The exam is published in Inspera at the original examination date (9th of June) and must be submitted as one PDF file in Inspera within the same time 7 days later. The faculty works on solutions for students who may not have access to a computer/network. Please contact administration if you have problems.*
- *STK4051. The grade will be pass/fail, and the limit for passing will be 40% (like E at the regular exam)*
- *STK9051. Ordinary rules for pass/fail will be applied. There will be an additional problem for you to solve.*

Susceptible-Infected-Recovered (SIR)

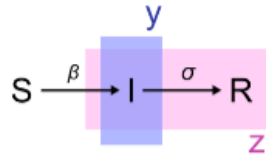
From : Jose lourenço, Robert Paton, Mahan Ghafari, Sunetra Gupta (2020)
 Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the SARS-CoV-2 epidemic
 DOI: [10.1101/2020.03.24.20042291](https://doi.org/10.1101/2020.03.24.20042291)

equation 1: $\frac{dy}{dt} = \beta y (1 - z) - \sigma y$

equation 2: $\frac{dz}{dt} = \beta y (1 - z)$

equation 3: $\Lambda_t = N \rho \theta z_{t-\psi}$

equation 4: $R_0 = \beta / \sigma$



[9]: DOI: [10.7554/ELIFE.29820](https://doi.org/10.7554/ELIFE.29820) eLife 2017;6:e29820

Markov chain monte carlo fitting approach

[Request a detailed protocol](#)

For the fitting process, the MCMC algorithm by Lourenço et al. is here altered to a Bayesian approach by formalising a likelihood and parameter priors (Lourenço and Recker, 2014). For this, the proposal distributions (q) of each parameter were kept as Gaussian (symmetric), effectively retaining a random walk Metropolis kernel. We define our acceptance probability α of a parameter set Θ , given model ODE output y as:

$$\alpha = \min\left\{1, \frac{\pi(y|\Theta^*)p(\Theta^*)q(\Theta^\circ|\Theta^*)}{\pi(y|\Theta^\circ)p(\Theta^\circ)q(\Theta^*|\Theta^\circ)}\right\} \quad (18)$$

where Θ^* and Θ° are the proposed and current (accepted) parameter sets (respectively); $\pi(y|\Theta^*)$ and $\pi(y|\Theta^\circ)$ are the likelihoods of the ODE output representing the epidemic data given each parameter set; $p(\Theta^\circ)$ and $p(\Theta^*)$ are the prior-related probabilities given each parameter set. We fit the Zika virus cumulative case counts per week, for which no age-related or geographical data is

Model output on cumulative death counts (Λ) is fitted to the reported time series of deaths (see Data) using a Bayesian MCMC approach previously implemented in other modelling studies [7–10]. Model variables are summarized in Table 1.

Variable / Parameter		Assumptions / Priors	Support
proportion infectious	y	equation 1	---
proportion of population no longer susceptible	z	equation 2	---
cumulative deaths	Λ	equation 3	---
time (day) of introduction	τ	Uniform distribution ($-\infty, +\infty$)	---
basic reproduction number	R_0	Gaussian distributions: G1(M=2.25, SD=0.025), G2(M=2.75, SD=0.025)	[11–13]
infectious period (days)	$1/\sigma$	Gaussian distribution G(M=4.5, SD=1)	[11,14–16]
transmission coefficient	β	$\beta = \sigma R_0$	---
time (days) between infection and death	ψ	Gaussian distribution G(M=17, SD=2)	[14]
probability of dying with severe disease	θ	Gaussian distribution G(M=0.14, SD=0.007)	[1,2,11,17]
proportion of population at risk of severe disease	ρ	Gamma distribution G1(S=5, R=5/0.01), G2(S=5, R=5/0.001)	---
population size	N	UK 66.87M, Italy 60M	---

Table 1 - Model variables and parameters. M=mean. SD=standard deviation. S=scale. R=rate.

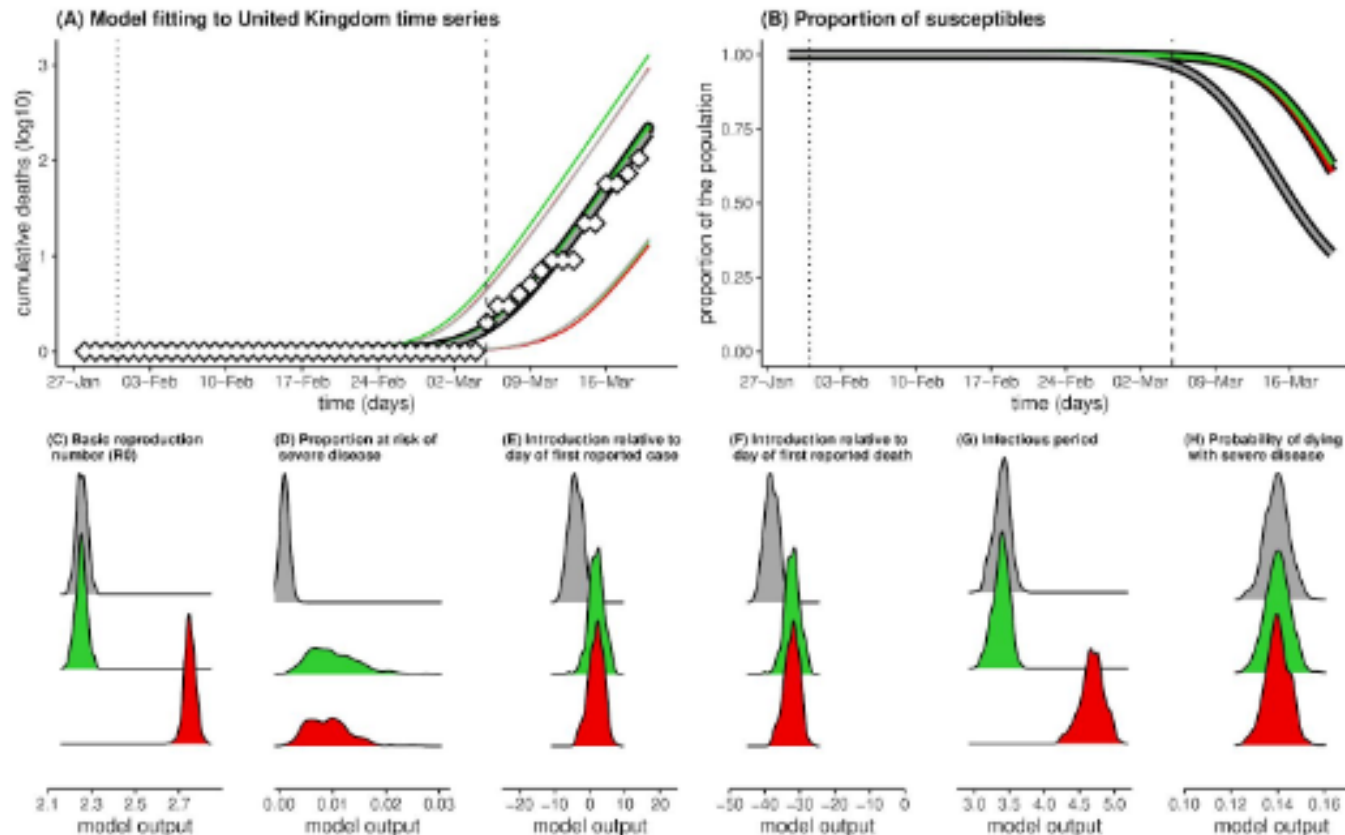
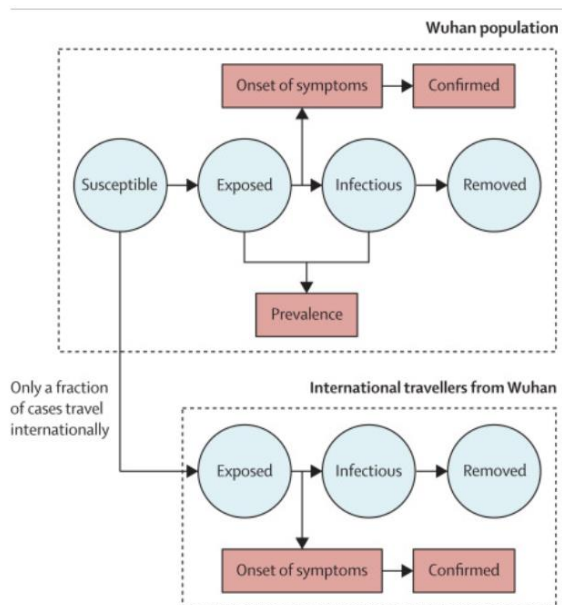


Figure 1. Results for the United Kingdom for three scenarios: $R_0 = 2.25$ and $\rho = 0.001$ (grey), $R_0 = 2.25$ and $\rho = 0.01$ (green), and $R_0 = 2.75$ and $\rho = 0.01$ (red). MCMC ran for 1 million steps. Results presented are the posteriors (model output) using 1000 samples after a burnout of 50% (A) Model fits showing reported (diamonds) and model (lines) cumulative death counts. Deaths are \log_{10} transformed for visualisation. (B) Mean proportion of the population still susceptible to infection ($1-z$, see Model). (A-B) Vertical lines mark the date of the first confirmed case (dotted) and date of first confirmed death (dashed). (C) Posteriors for R_0 , (D) proportion of population at risk of severe disease (ρ), (E) Time of introduction relative to the date of the first reported case, (F) Time of introduction relative to the date of first reported death, (G) infectious period, (H) probability of dying with severe disease.

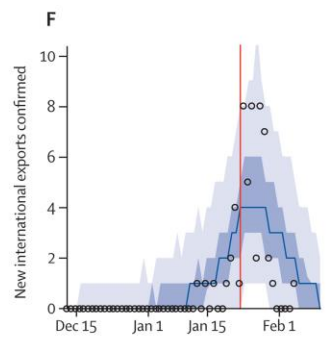
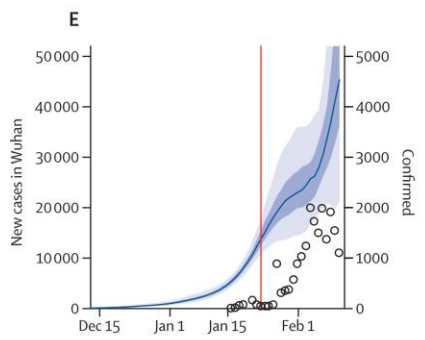
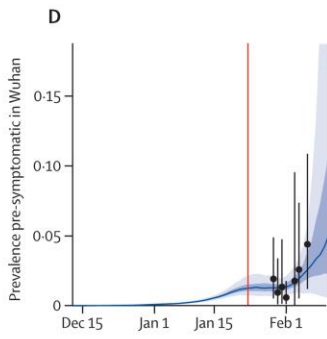
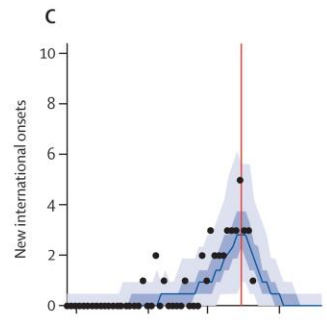
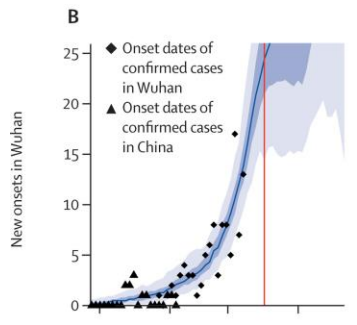
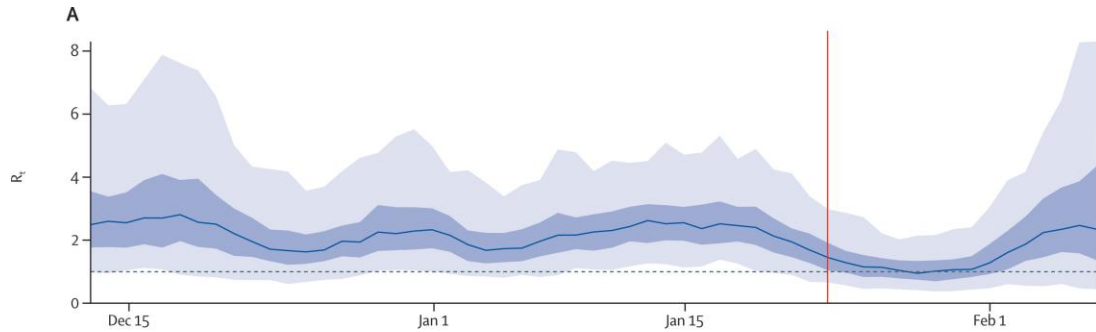
Early dynamics of transmission and control of COVID-19: a mathematical modelling study

The Lancet Infectious Diseases, (2020)

Adam J. Kucharski, Timothy W. Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebastian Funk, Rosalind M. Eggo



We modelled transmission as a geometric random walk process, and we used sequential Monte Carlo simulation to infer the transmission rate over time, as well as the resulting number of cases and the time-varying basic reproduction number (R_t), defined here as the mean number of secondary cases generated by a typical infectious individual on each day in a full susceptible population. The model had three unknown parameters, which we estimated: magnitude of temporal variability in transmission, proportion of cases that would eventually be detectable, and relative probability of reporting a confirmed case within Wuhan compared with an internationally exported case that originated in Wuhan. We assumed the outbreak started with a single infectious case on Nov 22, 2019, and the entire population was initially susceptible.



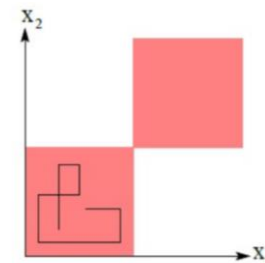
The red line marks travel restrictions starting on Jan 23, 2020. For parts (A) to (F) blue lines represent median, light blue shading represents 50% confidence intervals of the model estimate, and dark blue shading represents 95% confidence intervals of the model estimate. In all panels, datasets that were fitted to are shown as solid points; non-fitted data are shown as empty circles. (A) Estimated R_t over time. The dashed line represents an R_t of 1. (B) Onset dates of confirmed cases in Wuhan and China. (C) Reported cases by date of onset (black points) and estimated internationally exported cases from Wuhan by date of onset (blue line). (D) Estimated prevalence of infections that did not have detectable symptoms (blue line), and proportion of passengers on evacuation flights that tested positive for severe acute respiratory syndrome coronavirus 2 (black points; error bars show 95% binomial CIs). (E) New confirmed cases by date in Wuhan (circles, right hand axis) and estimated new symptomatic cases (blue line, left hand axis). (F) International exportation events by date of confirmation of case, and expected number of exports in the fitted model. (G) Estimated number of internationally exported cases from Wuhan confirmed up to Feb 10, 2020 and observed number in 20 countries with the highest connectivity to China. R_t =daily reproduction number.

Not fitted.
 Quality issue?

Last time

- Recap

- Irreducible / aperiodic/ recurrent
- Limiting distribution = stationary distribution
- Reparameterization
- The Gelman-Rubin diagnostic (Monte Carlo variance)



- Advanced topics

- Adaptive MCMC
- Multiple-Try M-H
- Slice sampler
- Simulated tempering
- Reversible Jump MCMC (model selection)
- Langevin
- Hamiltonian

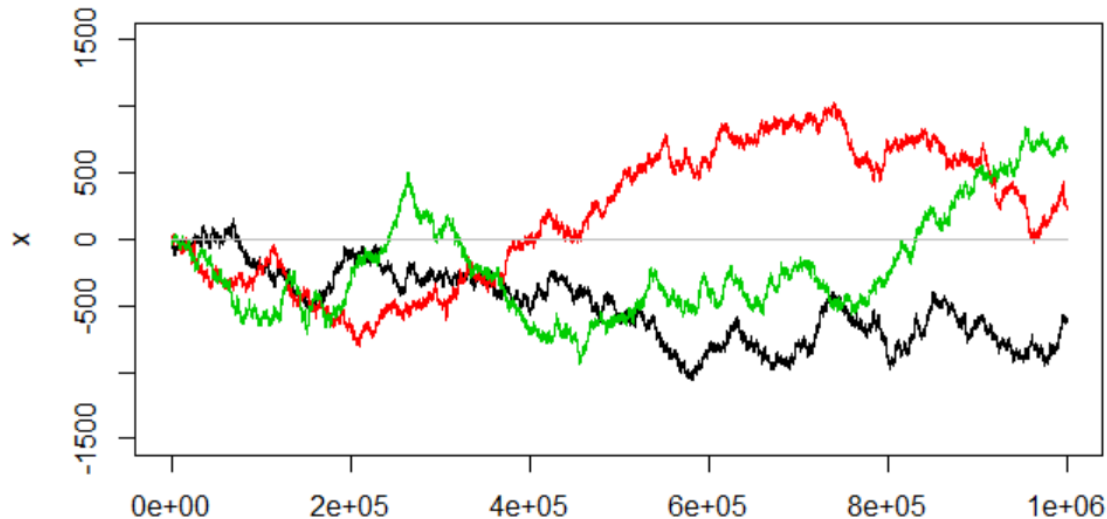
Question

- In the book Chapter 7 it appears that it is sufficient that the Markov chain is irreducible and aperiodic what happened to recurrent?
- Answer: Read also section 1.7 in book, here the term recurrent appears. The distinction is related to state space that are infinite.
- Example: random walk in R^d , $d \geq 3$

Example: improper prior

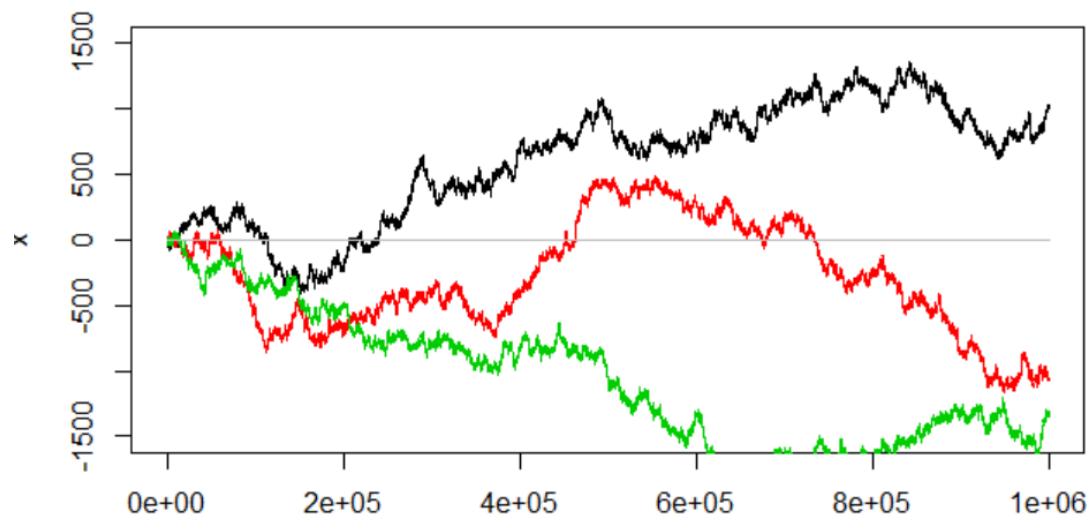
- $f(\mathbf{x}) \propto 1$, $\mathbf{x} = (x_1, x_2, x_3) \in R^3$
- $p(\mathbf{x}^* | \mathbf{x}) = \phi(x_1^*; x_1, 1) \cdot \phi(x_2^*; x_2, 1) \cdot \phi(x_3^*; x_3, 1)$
- Irreducible? (possible to reach any point with a finite number of steps)
 - Yes, there is a positive probability for any set of non-zero measure in one step.
- Aperiodic?
 - Yes, any non zero set can be reached at any time
- Detailed balance?
 - Yes we have $p(\mathbf{x}^* | \mathbf{x})f(\mathbf{x}) = p(\mathbf{x} | \mathbf{x}^*)f(\mathbf{x}^*)$
- So what could go wrong??
 - The chain is not recurrent

Example random walk in R^3



If you get sample paths like these, you might have a recurrence issue

Perhaps your target distribution is not a proper distribution [not easy to tell upfront]



If you safe guard yourself against zero density regions by setting a minimum density value. [you get into trouble]

Bayesian inference

- Assume observations \mathbf{x} , variables of interest \mathbf{z}
- \mathbf{z} may include parameters and/or latent variables
- Core: Posterior distribution

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})}$$

- Inference:

$$E[h(\mathbf{z})|\mathbf{x}] = \int_{\mathbf{z}} h(\mathbf{z})p(\mathbf{z}|\mathbf{x})d\mathbf{z}$$

- $p(\mathbf{z}|\mathbf{x})$ can be very complex and high dimensional
- No “black box” algorithm to solve “all problems”
 - Need diagnostics / burn in / effective number of samples
- Standard" methods such as MCMC do not scale well with big data

Variational Inference

- Main source: Blei et al. (2017) Variational inference: A review for statisticians. JASA

- Idea:

- **Approximate** $p(\mathbf{z}|\mathbf{x})$ by a simpler $q^*(\mathbf{z})$
- Perform inference by

$$E[h(\mathbf{z})|\mathbf{x}] \approx \int_{\mathbf{z}} h(\mathbf{z})q^*(\mathbf{z})d\mathbf{z}$$

← This integral is «easy»

- Question: How to find $q^*(\mathbf{z})$?

- Approach: Define $q^*(\mathbf{z})$ as

$$q^*(\mathbf{z}) = \arg \min_{q(\mathbf{z}) \in \mathcal{Q}} \mathcal{D}(q(\mathbf{z}), p(\mathbf{z}|\mathbf{x}))$$

← This optimization must be solved

where

- \mathcal{Q} is some class of distributions; **variational family**
- $\mathcal{D}(\cdot, \cdot)$ is some distance measure between distributions
- Note: q^* will depend on \mathbf{x} although not explicitly shown in the notation!

Variational inference

- Replace a «hard integral»

$$E[h(\mathbf{z})|\mathbf{x}] = \int_{\mathbf{z}} h(\mathbf{z})p(\mathbf{z}|\mathbf{x})d\mathbf{z}$$

with an «easy integral»

$$E[h(\mathbf{z})|\mathbf{x}] \approx \int_{\mathbf{z}} h(\mathbf{z})q^*(\mathbf{z})d\mathbf{z}$$

and optimization

$$q^*(\mathbf{z}) = \arg \min_{q(\mathbf{z}) \in \mathcal{Q}} \mathcal{D}(q(\mathbf{z}), p(\mathbf{z}|\mathbf{x}))$$

- Approximation without «general bounds» on error
 - Often: preserve the mean but not the variance
- We need:
 - \mathcal{D} a distance measure between distributions (KL-divergence)
 - \mathcal{Q} class of distributions (often: independent & Gaussian)

Kullback-Leibler divergence

$$\begin{aligned} \mathcal{D}(q(\mathbf{z}), p(\mathbf{z}|\mathbf{x})) \\ &= \text{KL}(q(\mathbf{z}) || p(\mathbf{z}|\mathbf{x})) = \int_{\mathbf{z}} \log \left(\frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})} \right) q(\mathbf{z}) d\mathbf{z} \\ &= E^q(\log q(\mathbf{Z})) - E^q(\log p(\mathbf{Z}|\mathbf{x})) \end{aligned}$$

If $p(\mathbf{z}|\mathbf{x})$ is zero and $q(\mathbf{z})$ is positive ☹️
 If $p(\mathbf{z}|\mathbf{x})$ is positive and $q(\mathbf{z})$ is zero 😊

- Properties:

- $\text{KL}(q(\mathbf{z}) || p(\mathbf{z}|\mathbf{x})) \geq 0$
- $\text{KL}(q(\mathbf{z}) || p(\mathbf{z}|\mathbf{x})) = 0 \Leftrightarrow q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x})$
- $\text{KL}(q(\mathbf{z}) || p(\mathbf{z}|\mathbf{x})) \neq \text{KL}(p(\mathbf{z}|\mathbf{x}) || q(\mathbf{z}))$ (not symmetric)
- Not a metric

Asymmetry of Kullback Liebler divergence

$$p(\mathbf{z}) \sim N(\mu_p, \Sigma_p), \quad \text{where } \mathbf{z} \in \mathbb{R}^d$$

$$q(\mathbf{z}) \sim N(\mu_q, \Sigma_q), \quad \text{where } \Sigma_q \text{ is diagonal}$$

Find μ_q , and diagonal Σ_q which minimizes:

* Problem 1 (Variational inference- VI)

$$\min_q \text{KL}(q(\mathbf{z}) || p(\mathbf{z})) = \min_q E^q(\log q(\mathbf{Z})) - E^q(\log p(\mathbf{Z}))$$

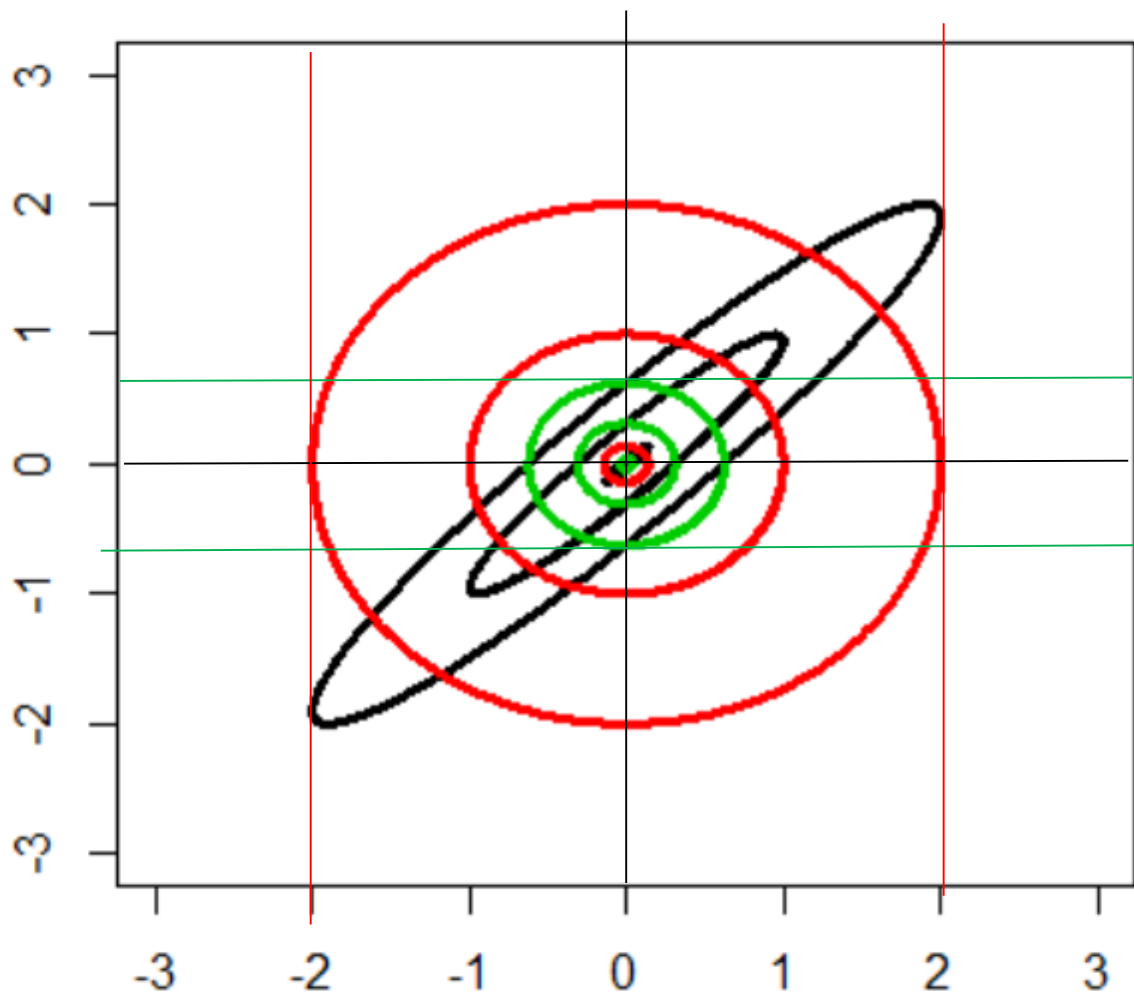
$$\Rightarrow E^q(z_k) = E^p(z_k), \quad \text{Var}^q(z_k) = \text{Var}^p(z_k | z_{-k})$$

* Problem 2 (Expectation propagation- EP)

$$\min_q \text{KL}(p(\mathbf{z}) || q(\mathbf{z})) = \min_q E^p(\log p(\mathbf{Z})) - E^p(\log q(\mathbf{Z}))$$

$$\Rightarrow E^q(z_k) = E^p(z_k), \quad \text{Var}^q(z_k) = \text{Var}^p(z_k)$$

VI vs EP



— GT
— EP
— VI

Ground truth:
Bi-normal
Correlation 0.95

Contours at 1 and 2 standard deviations

Computations: Variational inference

$$p(\mathbf{z}) \sim N(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p), \quad \text{where } \mathbf{z} \in \mathbb{R}^d$$

$$q(\mathbf{z}) \sim N(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q), \quad \text{where } \boldsymbol{\Sigma}_q \text{ is diagonal}$$

$$\text{KL}(q(\mathbf{z}) || p(\mathbf{z})) = E^q(\log q(\mathbf{Z})) - E^q(\log p(\mathbf{Z}))$$

$$\log(q(\mathbf{z})) = -\frac{1}{2} \log|\boldsymbol{\Sigma}_q| - \frac{d}{2} \log(2\pi) - \frac{1}{2} (\mathbf{z} - \boldsymbol{\mu}_q)^T \boldsymbol{\Sigma}_q^{-1} (\mathbf{z} - \boldsymbol{\mu}_q)$$

$E^q(\log q(\mathbf{Z}))$:

$$E^q \left((\mathbf{z} - \boldsymbol{\mu}_q)^T \boldsymbol{\Sigma}_q^{-1} (\mathbf{z} - \boldsymbol{\mu}_q) \right) = \text{Tr} \left(\boldsymbol{\Sigma}_q^{-1} E^q \left((\mathbf{z} - \boldsymbol{\mu}_q)(\mathbf{z} - \boldsymbol{\mu}_q)^T \right) \right)$$

$$= \text{Tr}(\boldsymbol{\Sigma}_q^{-1} \boldsymbol{\Sigma}_q) = \text{Tr}(\mathbf{I}) = d$$

$$E^q(\log q(\mathbf{Z})) = -\frac{1}{2} \log|\boldsymbol{\Sigma}_q| - \frac{d}{2} \log(2\pi) - \frac{d}{2}$$

$E^q(\log p(\mathbf{Z})):$

$$\log(p(\mathbf{z})) = -\frac{1}{2}\log|\Sigma_p| - \frac{d}{2}\log(2\pi) - \frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}_p)^T \boldsymbol{\Sigma}_p^{-1}(\mathbf{z} - \boldsymbol{\mu}_p)$$

$$E^q \left((\mathbf{z} - \boldsymbol{\mu}_p)^T \boldsymbol{\Sigma}_p^{-1}(\mathbf{z} - \boldsymbol{\mu}_p) \right) = \text{Tr} \left(\boldsymbol{\Sigma}_p^{-1} E^q \left((\mathbf{z} - \boldsymbol{\mu}_p)(\mathbf{z} - \boldsymbol{\mu}_p)^T \right) \right)$$

$$= \text{Tr} \left(\boldsymbol{\Sigma}_p^{-1} E^q \left((\mathbf{z} - \boldsymbol{\mu}_q + \boldsymbol{\mu}_q - \boldsymbol{\mu}_p)(\mathbf{z} - \boldsymbol{\mu}_q + \boldsymbol{\mu}_q - \boldsymbol{\mu}_p)^T \right) \right)$$

$$= \text{Tr} \left(\boldsymbol{\Sigma}_p^{-1} E^q \left(((\mathbf{z} - \boldsymbol{\mu}_q) + (\boldsymbol{\mu}_q - \boldsymbol{\mu}_p))((\mathbf{z} - \boldsymbol{\mu}_q) + (\boldsymbol{\mu}_q - \boldsymbol{\mu}_p))^T \right) \right)$$

$$= \text{Tr} \left(\boldsymbol{\Sigma}_p^{-1} E^q \left(\underbrace{(\mathbf{z} - \boldsymbol{\mu}_q)(\mathbf{z} - \boldsymbol{\mu}_q)^T}_{E()=\Sigma_q} + \underbrace{(\mathbf{z} - \boldsymbol{\mu}_q)(\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)^T}_{E()=0} + \underbrace{(\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)(\mathbf{z} - \boldsymbol{\mu}_q)^T}_{E()=0} + (\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)(\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)^T \right) \right)$$

$$= \text{Tr}(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q) + (\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)^T \boldsymbol{\Sigma}_p^{-1}(\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)$$

$$E^q(\log p(\mathbf{Z})) = -\frac{1}{2}\log|\Sigma_p| - \frac{d}{2}\log(2\pi) - \frac{1}{2}\text{Tr}(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q) - \frac{1}{2}(\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)^T \boldsymbol{\Sigma}_p^{-1}(\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)$$

$$p(\mathbf{z}) \sim N(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p), \quad \text{where } \mathbf{z} \in \mathbb{R}^d$$

$$q(\mathbf{z}) \sim N(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q), \quad \text{where } \boldsymbol{\Sigma}_q \text{ is diagonal}$$

$$\text{KL}(q(\mathbf{z}) || p(\mathbf{z})) =$$

$$-\frac{1}{2} \log |\boldsymbol{\Sigma}_q| - \frac{d}{2} + \frac{1}{2} \log |\boldsymbol{\Sigma}_p| + \frac{1}{2} \text{Tr}(\boldsymbol{\Sigma}_p^{-1} \boldsymbol{\Sigma}_q) + \frac{1}{2} (\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)^T \boldsymbol{\Sigma}_p^{-1} (\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)$$

$$\frac{\partial \text{KL}(q(\mathbf{z}) || p(\mathbf{z}))}{\partial \boldsymbol{\mu}_q}: \boldsymbol{\Sigma}_p^{-1} (\boldsymbol{\mu}_q - \boldsymbol{\mu}_p) = 0 \Rightarrow \boldsymbol{\mu}_q = \boldsymbol{\mu}_p$$

$$\frac{\partial \text{KL}(q(\mathbf{z}) || p(\mathbf{z}))}{\partial \boldsymbol{\Sigma}_q}: -\frac{1}{2} (\boldsymbol{\Sigma}_q^{-1} - \boldsymbol{\Sigma}_p^{-1}) = 0, \text{ (but only for diagonal elements)}$$

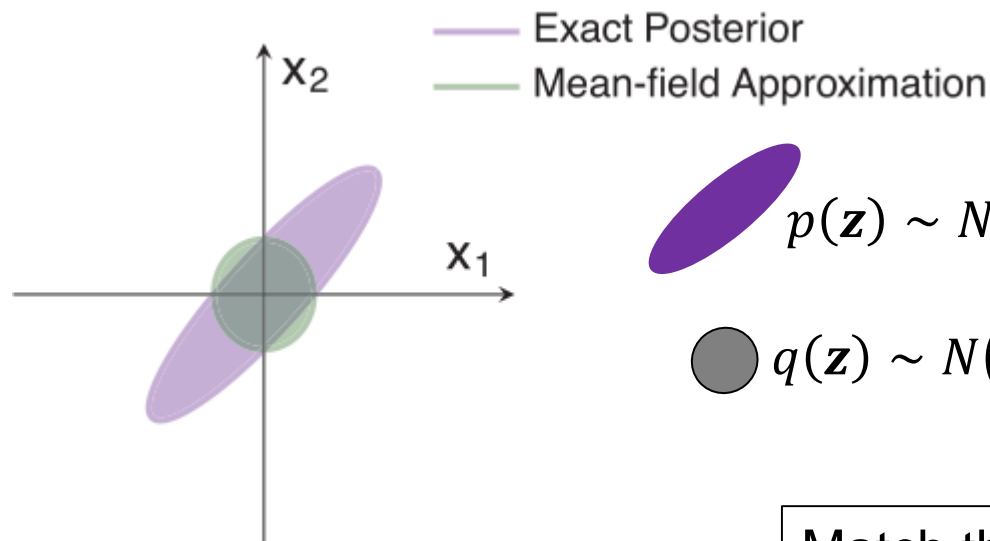
since $\boldsymbol{\Sigma}_q$ is diagonal


Solution:


$$E^q(z_k) = E^p(z_k)$$

$$\text{Var}^q(z_k) = \text{Var}^p(z_k | z_{-k})$$

Independent Gaussian approximation to a general Gaussian




 $p(\mathbf{z}) \sim N(\mu_p, \Sigma_p), \quad \text{where } \mathbf{z} \in \mathbb{R}^d$


 $q(\mathbf{z}) \sim N(\mu_q, \Sigma_q), \quad \text{where } \Sigma_q \text{ is diagonal}$

$$E^q(z_k) = E^p(z_k)$$

$$\text{Var}^q(z_k) = \text{Var}^p(z_k | z_{-k})$$

Match the mean ☺

Underestimate the variance ☹

This is due to the direction we define the KL divergence:

- Avoid «low probability» regions
- Little penalty for not including «high probability» regions

Is this what we want?

Variational inference

- General set up

$$E[h(\mathbf{z})|\mathbf{x}] \approx \int_{\mathbf{z}} h(\mathbf{z})q^*(\mathbf{z})d\mathbf{z}$$

$$q^*(\mathbf{z}) = \arg \min_{q(\mathbf{z}) \in \mathcal{Q}} \mathcal{D}(q(\mathbf{z}), p(\mathbf{z}|\mathbf{x}))$$

- If we can make \mathcal{Q} sufficiently general we can make as good approximations as we like
- When people talk about VI, they often mean:
 - \mathcal{Q} – Mean field approximation, (independence structure)
 - They maximize the ELBO which is equivalent to
$$\mathcal{D}(q, p) = KL(q||p)$$
- Much «common knowledge» of VI is related to this particular choice

The evidence lower bound (ELBO)

- Rewriting of the KL divergence:

$$\begin{aligned}
 KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) &= E^q[\log q(\mathbf{z})] - E^q[\log p(\mathbf{z}|\mathbf{x})] \\
 &= E^q[\log q(\mathbf{z})] - E^q[\log p(\mathbf{z}, \mathbf{x})] + E^q[\log p(\mathbf{x})] \\
 &= E^q[\log q(\mathbf{z})] - E^q[\log p(\mathbf{z}, \mathbf{x})] + \log p(\mathbf{x})
 \end{aligned}$$

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})}$$

- Minimizing KL is equivalent maximizing

$$ELBO(q) = E^q[\log p(\mathbf{z}, \mathbf{x})] - E^q[\log q(\mathbf{z})]$$

Log-evidence, not dependent on \mathbf{z}

$p(\mathbf{z}, \mathbf{x})$ implied

- ELBO = Evidence lower bound:

$$ELBO(q) = \log p(\mathbf{x}) - KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) \leq \log p(\mathbf{x})$$

The evidence $p(\mathbf{x})$ is the normalizing factor of Bayesian statistics
 Since KL is always positive. If KL is zero we have equality

$$ELBO(q) \leq \log p(\mathbf{x}) \quad (\text{hence the name})$$

ELBO practical use

- Minimizing KL is equivalent maximizing

$$\text{ELBO}(q) = E^q[\log p(\mathbf{z}, \mathbf{x})] - E^q[\log q(\mathbf{z})]$$

- Practical formulation:

$$p(\mathbf{z}, \mathbf{x}) = p(\mathbf{x}|\mathbf{z}) \cdot p(\mathbf{z})$$

$$\begin{aligned} \text{ELBO}(q) &= E^q[\log p(\mathbf{x}|\mathbf{z})] + E^q[\log p(\mathbf{z})] - E^q[\log q(\mathbf{z})] \\ &= E^q[\log p(\mathbf{x}|\mathbf{z})] - \text{KL}(q(\mathbf{z})||p(\mathbf{z})) \end{aligned}$$

Is maximum when
 $q(\mathbf{z}) \propto p(\mathbf{x}|\mathbf{z})$

Maximum when: (including sign)
 $q(\mathbf{z}) = p(\mathbf{z})$

$$\text{KL}(q(\mathbf{z})||c \cdot p(\mathbf{x}|\mathbf{z})) \geq 0$$

\Updownarrow

$$E^q(\log q(\mathbf{Z})) \geq E^q(\log(c \cdot p(\mathbf{x}|\mathbf{Z})))$$

Trade off between likelihood and prior

Common in Bayesian statistics

So what ?

- So far we have established that maximizing the ELBO gives a trade off between prior and posterior
- But we already knew that ELBO is maximized with the posterior:

$$\text{ELBO}(q) = \log p(\mathbf{x}) - KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) \leq \log p(\mathbf{x})$$

- So what have we achieved?
 - A variational form for the approximation
- We now restrict the class of distributions to \mathcal{Q} so that integrals are easy to compute thus $p(\mathbf{z}|\mathbf{x})$ is not in \mathcal{Q}

$$q^*(\mathbf{z}) = \arg \min_{q(\mathbf{z}) \in \mathcal{Q}} \mathcal{D}(q(\mathbf{z}), p(\mathbf{z}|\mathbf{x}))$$

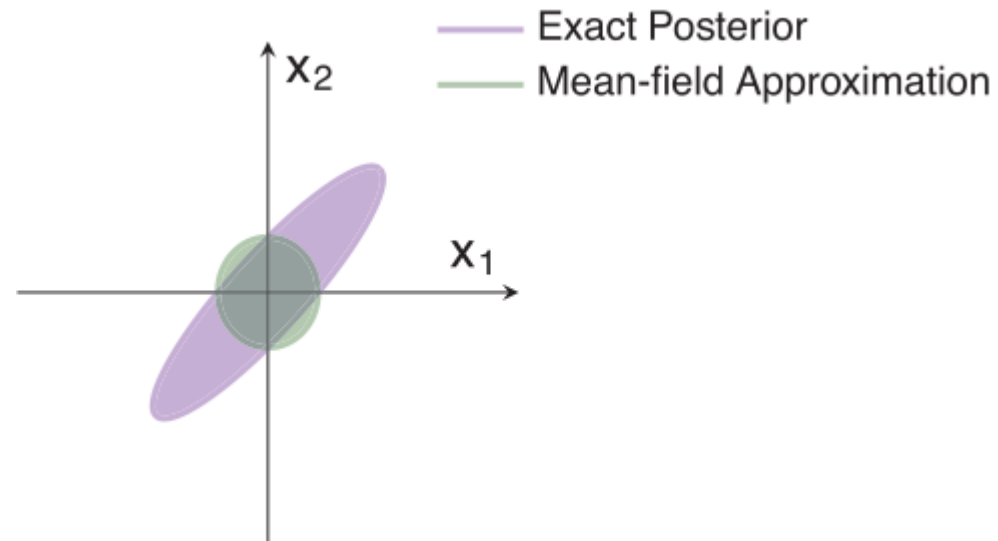
Mean field variational family

- Common choice: **Mean-field variational family**

$$q(\mathbf{z}) = \prod_{j=1}^m q_j(z_j)$$

that is an **independence** structure

- Also need to specify $q_j(z_j)$.
 - For continuous variables, Gaussians is possible
 - For categorical variables, any discrete distribution.
- Typically, point estimates quite accurate, uncertainty measures too low



Example mixture Gaussian

- Model

$$\mu_k \sim \mathcal{N}(0, \sigma^2), \quad k = 1, \dots, K$$

$$c_i \sim \text{Categorical}(1/K, \dots, 1/K), \quad i = 1, \dots, n$$

$$x_i | c_i, \boldsymbol{\mu} \sim \mathcal{N}(\mu_{c_i}, 1)$$

Of interest

$$p(\boldsymbol{\mu}, \mathbf{c} | \mathbf{x}) \propto p(\boldsymbol{\mu}, \mathbf{c}, \mathbf{x}) = p(\boldsymbol{\mu}) \prod_{i=1}^n p(c_i) p(x_i | c_i, \boldsymbol{\mu})$$

- Variational approximation:

$$q(\boldsymbol{\mu}, \mathbf{c}) = \prod_{k=1}^K q(\mu_k; m_k, s_k^2) \prod_{i=1}^n q(c_i; \boldsymbol{\phi}_i)$$

where $\boldsymbol{\phi}_i = (\phi_{i,1}, \dots, \phi_{i,K})$, $\sum_{k=1}^K \phi_{i,k} = 1$ and

$$q(\mu_k; m_k, s_k^2) = \mathcal{N}(\mu_k; m_k, s_k^2)$$

$$q(c_i; \boldsymbol{\phi}_i) = \phi_{i,c_i}$$

- Need to specify $\{(m_k, s_k^2), k = 1, \dots, K\}$ and $\{\phi_{i,k}, i = 1, \dots, n, k = 1, \dots, K\}$.

Optimization - CAVI

- $\text{ELBO}(q) = E^q (\log p(\mathbf{Z}, \mathbf{x})) - E^q (\log q(\mathbf{z}))$

$$q(\mathbf{z}) = \prod_{j=1}^m q_j(z_j)$$

- Optimize one $q_j(z_j)$ at the time keep the others fixed

$$\min_{q_j} E^q (\log p(\mathbf{Z}, \mathbf{x})) - E^q (\log q(\mathbf{z}))$$

$$\min_{q_j} E_{z_j}^q \left(\underbrace{E_{z_{-j}}^q (\log p(\mathbf{Z}, \mathbf{x}))}_{= \log q_j(z_j)} \right) - E_{z_j}^q (\log q_j(z_j)) + \text{const}$$

$$q_j(z_j) \propto \exp \left\{ E_{z_{-j}}^q (\log p(\mathbf{Z}, \mathbf{x})) \right\}$$

Optimization - CAVI

- CAVI = Coordinate ascent variational inference
- Given all other variables, the optimal $q_j^*(z_j)$ is

$$q_j^*(z_j) \propto \exp\{E_{\mathbf{z}_{-j}}^q[\log p(z_j | \mathbf{z}_{-j}, \mathbf{x})]\}$$

$$\propto \exp\{E_{\mathbf{z}_{-j}}^q[\log p(z_j, \mathbf{z}_{-j}, \mathbf{x})]\}$$

- Algorithm

Input: A model $p(\mathbf{x}, \mathbf{z})$, a data set \mathbf{x}

Output: A variational density $q(\mathbf{x}) = \prod_{j=1}^m q_j(z_j)$

Initialize: Variational factors $q_j(z_j)$

while *the ELBO has not converged* **do**

for $j \in \{1, \dots, m\}$ **do**

 Set $q_j(z_j) \propto \exp\{E_{\mathbf{z}_{-j}}^q[\log p(z_j, \mathbf{z}_{-j}, \mathbf{x})]\}$

end

 Compute $\text{ELBO}(q) = E^q[\log p(\mathbf{z}, \mathbf{x})] - E^q[\log q(\mathbf{z})]$

end

- A type of "message passing" algorithm: Enables automated software for a large class of models
- Note: Similar structure as the **Gibbs sampler!**

$$p(\boldsymbol{\mu}, \mathbf{c}|\mathbf{x}) \propto p(\boldsymbol{\mu}, \mathbf{c}, \mathbf{x}) = p(\boldsymbol{\mu}) \prod_{i=1}^n p(c_i) p(x_i|c_i, \boldsymbol{\mu})$$

Example - updating for c_i :

$$q(\boldsymbol{\mu}, \mathbf{c}) = \prod_{k=1}^K q(\mu_k; m_k, s_k^2) \prod_{i=1}^n q(c_i; \phi_i)$$

$$q^*(c_i; \phi_i) \propto \exp\{E_{\mathbf{c}_{-i}, \boldsymbol{\mu}}[\log p(c_i, \mathbf{c}_{-i}, \boldsymbol{\mu}, \mathbf{x})]\}$$

$$\propto \exp\{E_{\mathbf{c}_{-i}, \boldsymbol{\mu}}[\log(p(c_i)) + \log(p(\mathbf{c}_{-i})) + \log p(\boldsymbol{\mu}) + \sum_{j=1}^n \log p(x_j|c_j, \boldsymbol{\mu})]\}$$

$$\propto \exp\{E_{\mathbf{c}_{-i}, \boldsymbol{\mu}}[\log(p(c_i)) + \log p(x_i|c_i, \boldsymbol{\mu})]\}$$

$$\propto \exp\{E_{\mathbf{c}_{-i}, \boldsymbol{\mu}}[\frac{1}{K} + \log p(x_i|\mu_{c_i})]\}$$

$$\propto \exp\{E_{\boldsymbol{\mu}}[[-\frac{1}{2}(x_i - \mu_{c_i})^2]]\}$$

$$\propto \exp\{E_{\boldsymbol{\mu}}[-\frac{1}{2}(x_i^2 - 2x_i\mu_{c_i} + \mu_{c_i}^2)]\}$$

$$\propto \exp\{[x_i m_{c_i} - \frac{1}{2} s_{c_i}^2 - \frac{1}{2} m_{c_i}^2]\}$$

$$p(\boldsymbol{\mu}, \mathbf{c} | \mathbf{x}) \propto p(\boldsymbol{\mu}, \mathbf{c}, \mathbf{x}) = p(\boldsymbol{\mu}) \prod_{i=1}^n p(c_i) p(x_i | c_i, \boldsymbol{\mu})$$

Example, updating for μ_k

$$q(\boldsymbol{\mu}, \mathbf{c}) = \prod_{k=1}^K q(\mu_k; m_k, s_k^2) \prod_{i=1}^n q(c_i; \phi_i)$$

$$q^*(\mu_k; m_k, s_k^2) \propto \exp\{E_{\mathbf{c}, \boldsymbol{\mu}_{-k}}[\log p(\mathbf{c}, \boldsymbol{\mu}, \mathbf{x})]\}$$

$$\propto \exp\{E_{\mathbf{c}, \boldsymbol{\mu}_{-k}}[\log p(\mu_k) + \sum_{i=1}^n \log p(x_i | c_i, \boldsymbol{\mu})]\}$$

$$\propto \exp\{E_{\mathbf{c}, \boldsymbol{\mu}_{-k}}[-\frac{1}{2\sigma^2}\mu_k^2 + \sum_{i=1}^n I(c_i = k) \log p(x_i | \mu_k)]\}$$

$$\propto \exp\{[-\frac{1}{2\sigma^2}\mu_k^2 + \sum_{i=1}^n \phi_{ik}[-\frac{1}{2}(x_i - \mu_k)^2]]\} \quad \longleftarrow \text{Independence in } q$$

$$\propto \exp\{[-\frac{1}{2\sigma^2}\mu_k^2 - \frac{1}{2} \sum_{i=1}^n \phi_{ik} [x_i^2 - 2x_i\mu_k + \mu_k^2]]\}$$

$$\propto \exp\{-\frac{1}{2}[(\frac{1}{\sigma^2} + \sum_{i=1}^n \phi_{ik})\mu_k^2 - 2 \sum_{i=1}^n \phi_{ik} x_i \mu_k]\}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2} + \sum_{i=1}^n \phi_{ik}\right) \left[\mu_k - \frac{\sum_{i=1}^n \phi_{ik} x_i}{\frac{1}{\sigma^2} + \sum_{i=1}^n \phi_{ik}}\right]^2\right\}$$

$$\propto N\left(\mu_k; \frac{\sum_{i=1}^n \phi_{ik} x_i}{\frac{1}{\sigma^2} + \sum_{i=1}^n \phi_{ik}}, \left(\frac{1}{\sigma^2} + \sum_{i=1}^n \phi_{ik}\right)^{-1}\right)$$

$$p(\boldsymbol{\mu}, \mathbf{c} | \mathbf{x}) \propto p(\boldsymbol{\mu}, \mathbf{c}, \mathbf{x}) = p(\boldsymbol{\mu}) \prod_{i=1}^n p(c_i) p(x_i | c_i, \boldsymbol{\mu})$$

Example - calculating ELBO

$$q(\boldsymbol{\mu}, \mathbf{c}) = \prod_{k=1}^K q(\mu_k; m_k, s_k^2) \prod_{i=1}^n q(c_i; \phi_i)$$

$$\text{ELBO}(q) = E^q[\log p(\mathbf{z}, \mathbf{x})] - E^q[\log q(\mathbf{z})]$$

$$= E^q \left[\sum_{i=1}^n \log p(c_i) + \sum_{k=1}^K \log p(\mu_k) + \sum_{i=1}^n \log p(x_i | c_i, \boldsymbol{\mu}) \right] -$$

$$E^q \left[\sum_{i=1}^n \log q(c_i) + \sum_{k=1}^K \log q(\mu_k) \right]$$

$$= E^q \left[\sum_{i=1}^n \log\left(\frac{1}{K}\right) - \frac{1}{2\sigma^2} \sum_{k=1}^K \mu_k^2 - \frac{1}{2} \sum_{i=1}^n \sum_{k=1}^K I(c_i = k) (x_i - \mu_k)^2 \right] -$$

$$E^q \left[\sum_{i=1}^n \sum_{k=1}^K \phi_{ik} \log(\phi_{ik}) + \sum_{k=1}^K \left[-\frac{1}{2} \log(s_k^2) - \frac{1}{2s_k^2} (\mu_k - m_k)^2 \right] \right] + \text{const}$$

$$= E^q \left[-\frac{1}{2\sigma^2} \sum_{k=1}^K [s_k^2 + m_j^2] - \frac{1}{2} \sum_{i=1}^n \sum_{k=1}^K \phi_{ik} [x_i^2 - 2x_i \mu_k + \mu_k^2] - \right.$$

$$\left. E^q \left[\sum_{i=1}^n \sum_{k=1}^K \phi_{ik} \log(\phi_{ik}) + \sum_{k=1}^K \left[-\frac{1}{2} \log(s_k^2) - \frac{1}{2} \right] \right] + \text{const}$$

$$= -\frac{1}{2\sigma^2} \sum_{k=1}^K [s_k^2 + m_j^2] + \sum_{i=1}^n \sum_{k=1}^K \phi_{ik} [x_i m_k - \frac{1}{2} s_k^2 - \frac{1}{2} m_k^2] -$$

$$\sum_{i=1}^n \sum_{k=1}^K \phi_{ik} \log(\phi_{ik}) + \frac{1}{2} \sum_{k=1}^K \log(s_k^2) + \text{const}$$

Implementation ELBO- CAVI

```
phi = matrix(1/K,nrow=n,ncol=K)
m = mu
s2 = rep(1,K)
more = TRUE
Elbo = 0
```

VI_mix_dim1.R

```
while(more)
```

```
{
```

```
  for(i in 1:n)
```

```
  {
```

```
    phi[i,] = exp(m*x[i]-0.5*s2-0.5*m^2)
```

```
    phi[i,] = phi[i,]/sum(phi[i,])
```

```
  }
```

```
  for(k in 1:K)
```

```
  {
```

```
    m[k] = sum(phi[,k]*x)/(tau.mu+sum(phi[,k]))
```

```
    s2[k] = 1/(tau.mu+sum(phi[,k]))
```

```
  }
```

```
  elbo = -0.5*tau.mu*sum(s2+m^2)-sum(rowSums(phi*log(phi)))+0.5*sum(log(s2))
```

```
  for(k in 1:K)
```

```
    elbo = elbo + sum(phi[,k]*(x*m[k]-0.5*s2[k]-0.5*m[k]^2))
```

```
  more = abs(tail(Elbo,n=1)-elbo)>eps
```

```
  Elbo = c(Elbo,elbo)
```

```
}
```

$$\propto \exp\left\{[x_i m_{c_i} - \frac{1}{2} s_{c_i}^2 - \frac{1}{2} m_{c_i}^2]\right\}$$

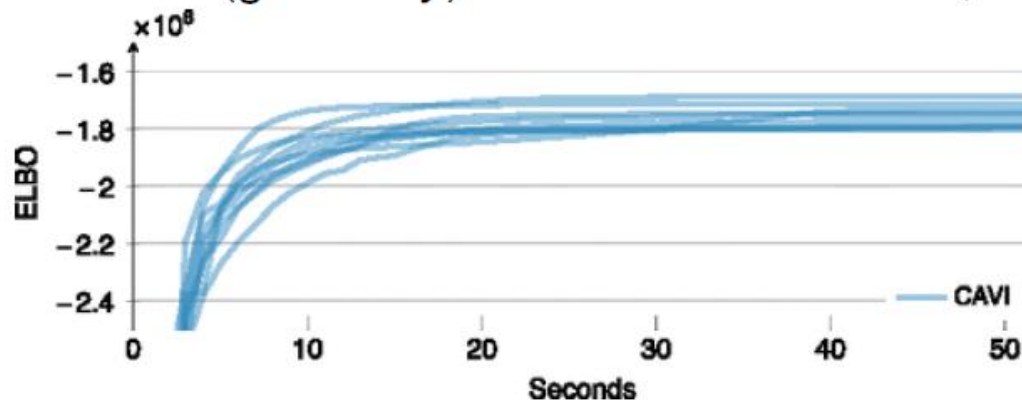
$$\propto N(\mu_k; \frac{\sum_{i=1}^n \phi_{ik} x_i}{\frac{1}{\sigma^2} + \sum_{i=1}^n \phi_{ik}}, (\frac{1}{\sigma^2} + \sum_{i=1}^n \phi_{ik})^{-1})$$

$$= -\frac{1}{2\sigma^2} \sum_{k=1}^K [s_k^2 + m_j^2] + \sum_{i=1}^n \sum_{k=1}^K \phi_{ik} [x_i m_k - \frac{1}{2} s_k^2 - \frac{1}{2} m_k^2] -$$

$$\sum_{i=1}^n \sum_{k=1}^K \phi_{ik} \log(\phi_{ik}) + \frac{1}{2} \sum_{k=1}^K \log(s_k^2)$$

Different initializations may lead CAVI to find different local optima of the ELBO

- ELBO is (generally) a nonconvex function, local optimum



Theory Results on the following type:

- Treat VI posterior means as point estimates
- Bayesian linear models: VI posterior means are consistent
- Poisson-mixed model with Gaussian VI
 - Consistency and asymptotic normality
- Network data with stochastic block models: Asymptotic normality (which might not be the case for ML estimates!)
- Gaussian variational approximations: Asymptotic covariance matrix estimator
- Mixture of Gaussians: CAVI converges to local optimum, VI estimator is consistent, VI posterior variance too small.

Extensions and open problems

- Other distance measures than KL
- Alternatives to mean-field
 - Add dependencies between variables: structured variational inference
 - Mixtures of variational densities
- Interface between VI and MCMC
- Statistical properties

VI compared to ML/MAP estimation

- VI retain Bayesian inference ideas but neglect correlations between parameters
- ML/MAP retain the model but neglect the uncertainty in the parameters,
- MCMC retain both.
- Expectation propagation is an other alternative (Minka 2001 & Hernández-Lobato et al., 2015)

McMC vs variational inference

- MCMC methods tend to be more computationally intensive than variational inference but they also provide guarantees of producing (asymptotically) exact samples from the target density
 - Suited for *small* datasets where precise inference is required
- Variational inference does not enjoy such guarantees - it can only find a density close to the target - but tends to be faster than MCMC.
 - Can be combined with stochastic gradient!
 - Suited for large datasets, many models to explore
- The relative accuracy of variational inference and MCMC is still unknown.
 - Variational inference generally underestimates variance
- Modern research
 - tackling Bayesian inference problems that involve massive data
 - improved optimization methods for minimizing KL
 - developing generic variational inference algorithms that are easy to apply to a wide class of models
 - increasing the accuracy of variational inference,

Reference

- D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.
- F. Guo, X. Wang, K. Fan, T. Broderick, and D. B. Dunson. Boosting variational inference. *arXiv preprint arXiv:1611.05559*, 2016.
- J. M. Hernández-Lobato, D. Hernández-Lobato, and A. Suárez. Expectation propagation in linear regression models with spike-and-slab priors. *Machine Learning*, 99(3):437–487, 2015.
- D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- A. Kucukelbir, R. Ranganath, A. Gelman, and D. Blei. Automatic variational inference in stan. In *Advances in neural information processing systems*, pages 568–576, 2015.
- T.P Minka, “Expectation Propagation for Approximate Bayesian Inference,” in *Uncertainty in Artificial Intelligence*, pp. 362–369. [860,873], 2001