

LECTURE 12: DISTRIBUTIONAL APPROXIMATIONS

- MCMC is expensive, specially for hierarchical models, so a number of approximations have been developed
- Expectation-Maximization
- Variational Inference

Expectation-Maximization (EM) Algorithm

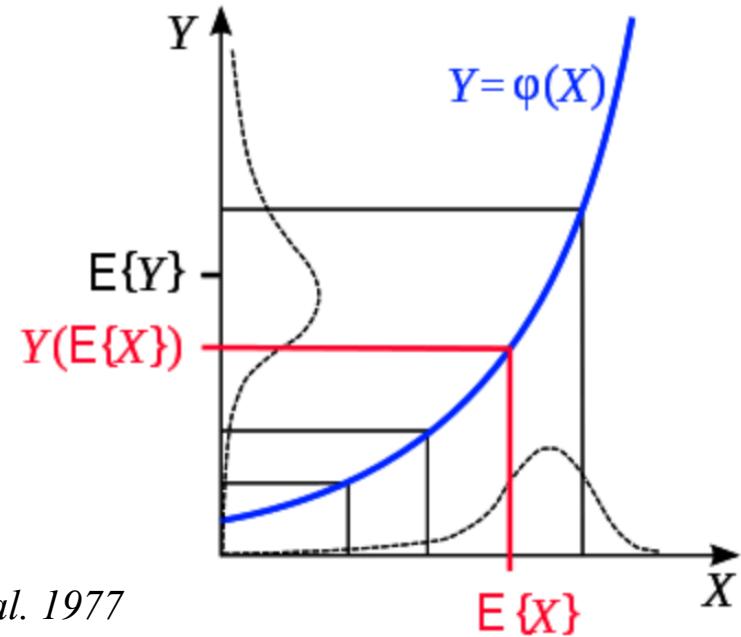
- We have **data X , parameters θ and latent variables Z** (which often are of the same size as X).
- In hierarchical models we know how to write conditionals $p(X|Z, \theta)$ and $p(Z|\theta)$ but it is hard to integrate out Z to write directly $p(X|\theta)$, and thus posterior $p(\theta|X)$ (we will assume flat prior), i.e. it is hard to compute

$$p(X|\theta) = \int p(X,Z|\theta)dZ = \int p(X|Z,\theta)p(Z|\theta)dZ$$

- **Jensen inequality** for convex Y :

$$\varphi(E[X]) \leq E[\varphi(X)] = E[Y]$$

- Opposite for concave (log)

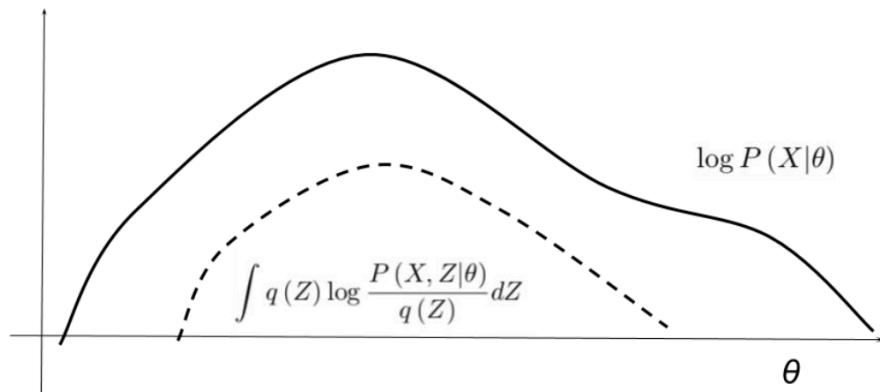


Credit: Dempster et al. 1977

Jensen Inequality applied to $\log P$

- For any $q(Z)$ we have

$$\log \int P(X, Z|\theta) dZ = \log \int P(X, Z|\theta) \frac{q(Z)}{q(Z)} dZ \geq \int q(Z) \log \frac{P(X, Z|\theta)}{q(Z)} dZ$$

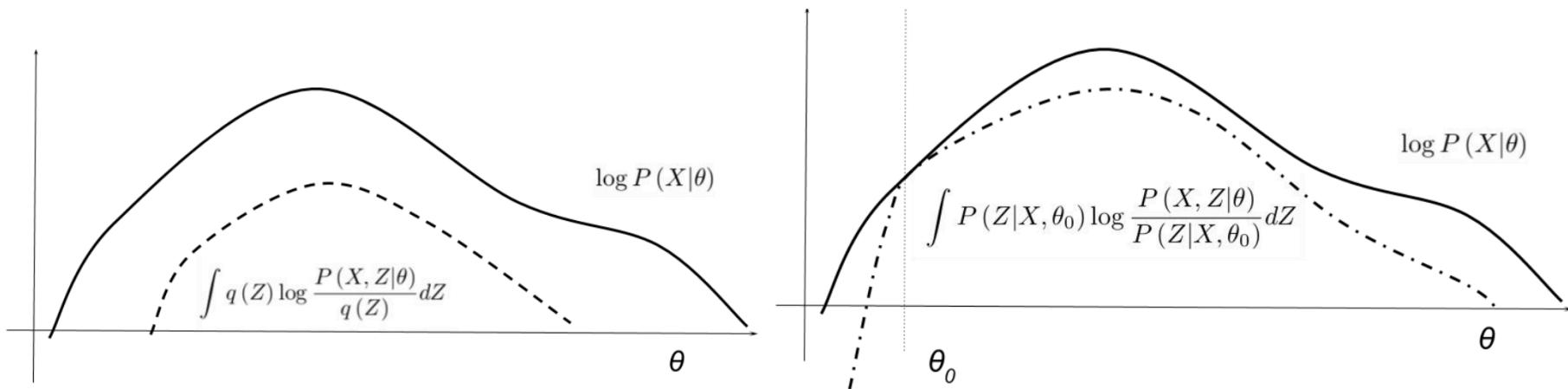


Credit: Slide from R. Giordano

Jensen Equality

- This can be equality if $q(Z) = p(Z|X, \theta_0)$, but only at $\theta = \theta_0$

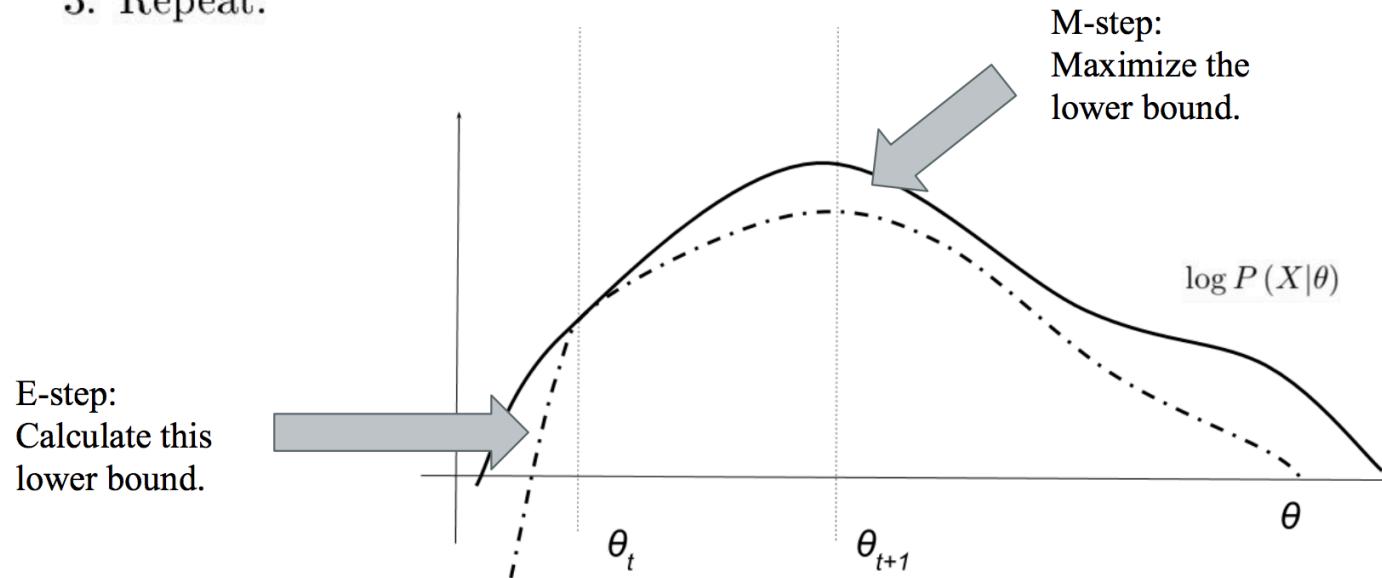
$$\begin{aligned}\int P(Z|X, \theta_0) \log \frac{P(X, Z|\theta_0)}{P(Z|X, \theta_0)} dZ &= \int P(Z|X, \theta_0) \log \frac{P(X, Z|\theta_0) P(X|\theta_0)}{P(X, Z|\theta_0)} dZ \\ &= \log P(X|\theta_0) \int P(Z|X, \theta_0) dZ = \log P(X|\theta_0)\end{aligned}$$



- Suppose we want to determine MLE/MAP of $p(X|\theta)$ or $p(\theta | X)$ over q : this suggests a strategy is to maximize over θ given previous solution

EM Algorithm

1. E-step: Starting at θ_t , calculate the expectation $E(\theta) = \int P(Z|X, \theta_t) \log P(X, Z|\theta) dZ$
2. M-step: Optimize $\theta_{t+1} = \operatorname{argsup} E(\theta)$
3. Repeat.

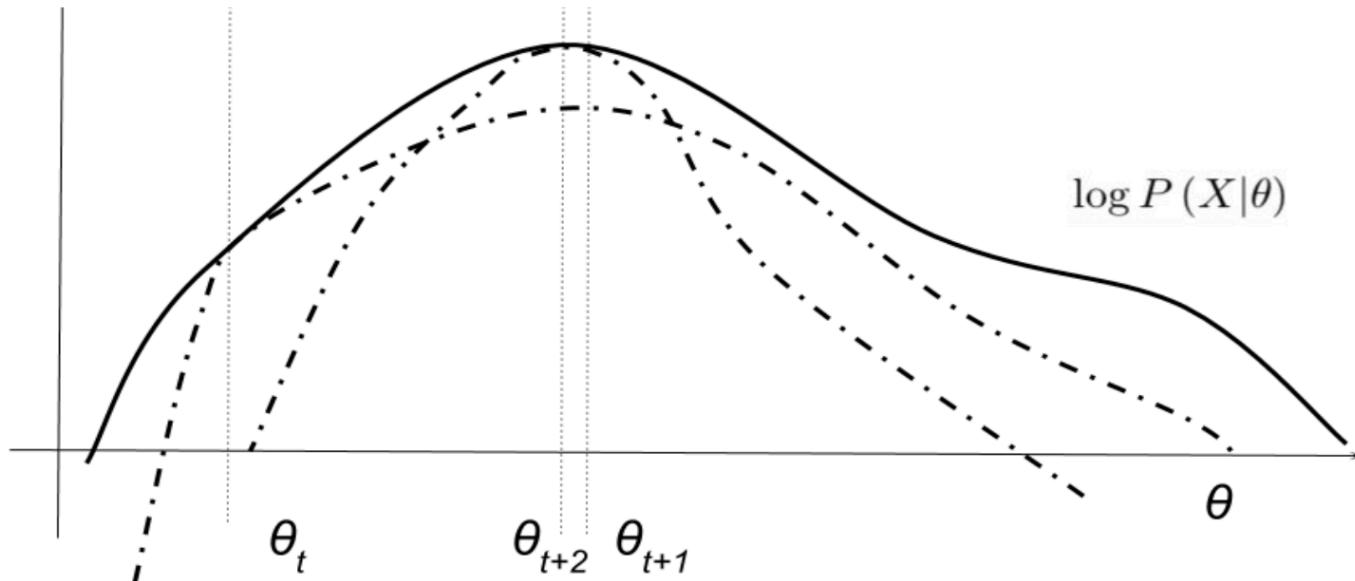


Generalized EM: if M is unsolvable then instead of maximization over θ make any move in the direction of increasing the value (similar to NL optimizations)

Guaranteed to work

This is guaranteed to increase the marginal likelihood $\log P(\theta|X)$ since

$$\begin{aligned}\sup_{\theta} \int P(Z|X, \theta_t) \log P(X, Z|\theta) dZ &= \sup_{\theta} \int P(Z|X, \theta_t) \log \frac{P(X, Z|\theta)}{P(Z|X, \theta_t)} dZ \\ &\geq \int P(Z|X, \theta_t) \log \frac{P(X, Z|\theta_t)}{P(Z|X, \theta_t)} dZ = \log P(\theta_t|X)\end{aligned}$$



Often rapid convergence if good starting point

Note however that it solves an optimization problem: finds the nearest local maximum

Why is it useful?

- Two reasons: performs marginalization over latent parameters and avoids evaluating the normalizations

$$\sup_{\theta} \int P(Z|X, \theta_t) \log P(X, Z|\theta) dZ$$

Conditionals are simpler than marginals



Log probabilities are simpler than probabilities (no Z in the normalizing constant!)

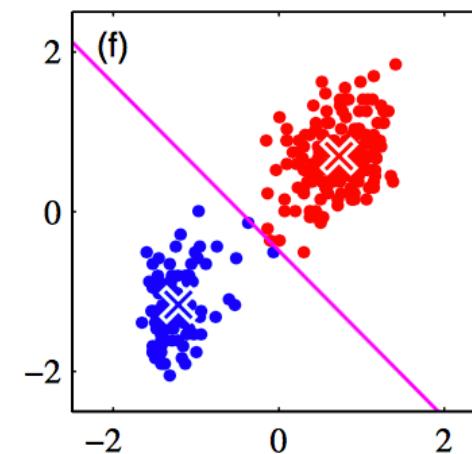
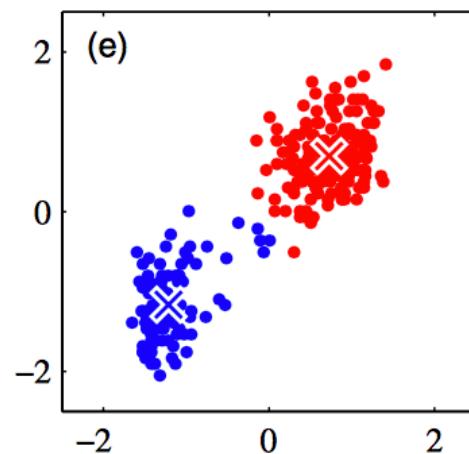
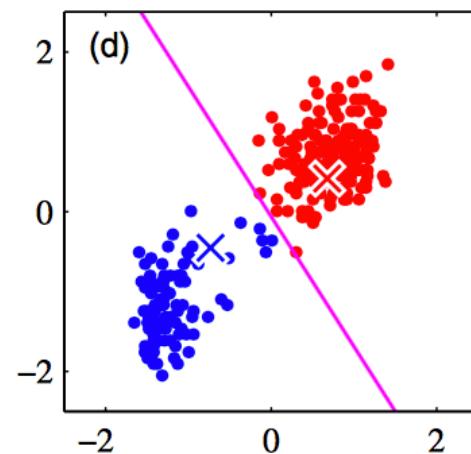
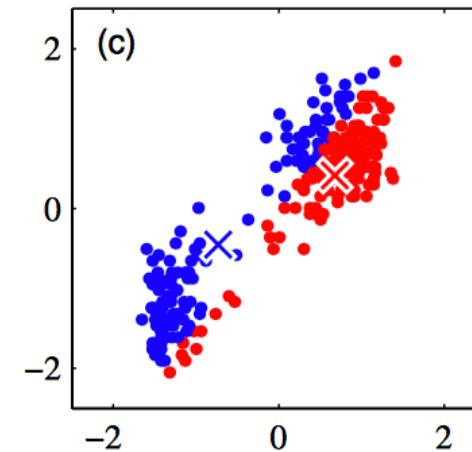
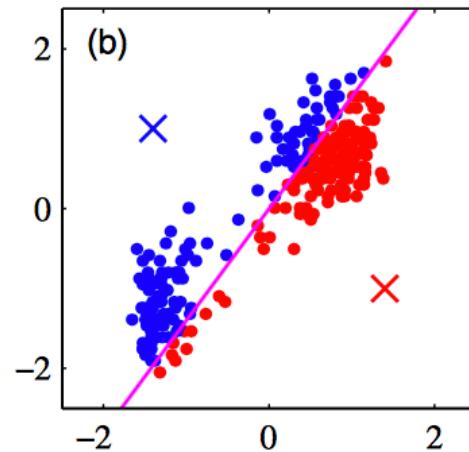
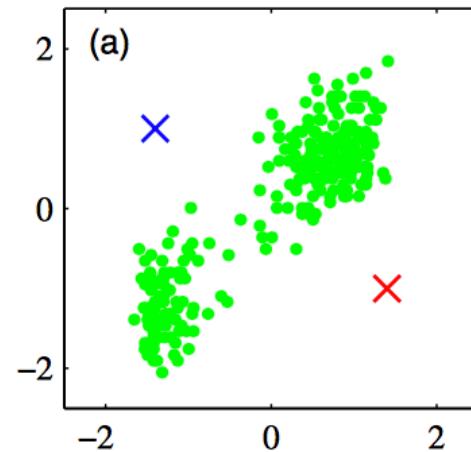
- However, it only gives MLE/MAP
- Extension called supplemented EM evaluates curvature matrix at MLE/MAP (see Gelman et al)

Cluster Classification: K-means

- Before looking at EM let's look at a non-probabilistic approach called K-means clustering
- We have N observations x_n and each x_n is in D-dimensions
- We want to partition it into K clusters
- Let's assume they are given simply by K means μ_k representing cluster centers
- We can define loss or objective function $J = \sum_n \sum_k r_{nk} (x_n - \mu_n)^2$ where $r_{nk} = 1$ for one k and $r_{nj} = 0$ for $j \neq k$, so that each data point is assigned to a single cluster k .
- Optimizing J for r_{nk} gives us $r_{nk} = 1$ for whichever k minimizes the distance $(x_n - \mu_k)^2$, set $r_{nj} = 0$ for $j \neq k$. This is the expectation part in EM language.
- Optimizing J for μ_k at fixed r_{nk} we take a derivative of J wrt μ_k which gives $\mu_k = \sum_n r_{nk} x_n / (\sum_n r_{nk})$. This is M part. Repeat.

Example (Bishop Chap. 9)

- Random starting μ_k (crosses). Magenta line is the cluster divider



Gaussian Mixture with Latent Variables

- We have seen GM before: $p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- Now we also introduce a latent variable z_{nk} playing the role of r_{nk} , i.e. for each n one is 1 and the other $K-1$ are 0. The marginal distribution is $p(z_k=1) = \pi_k$, where $\sum_k \pi_k = 1$ and $0 \leq \pi_k \leq 1$. Conditional of x given $z_{k=1}$ is a gaussian

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$
$$p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k} \quad p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^K \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

More variables make it easier

- We have defined latent variables z we want to marginalize over. Advantage is that we can work with $p(x,z)$ rather than $p(x)$. Lesson: adding many parameters sometimes makes the problem easier.
- We also need responsibility $\gamma(z_k) = p(z_k=1|x)$, using Bayes
- Here π_k is prior for $p(z_k=1)$, $\gamma(z_k)$ is posterior given x

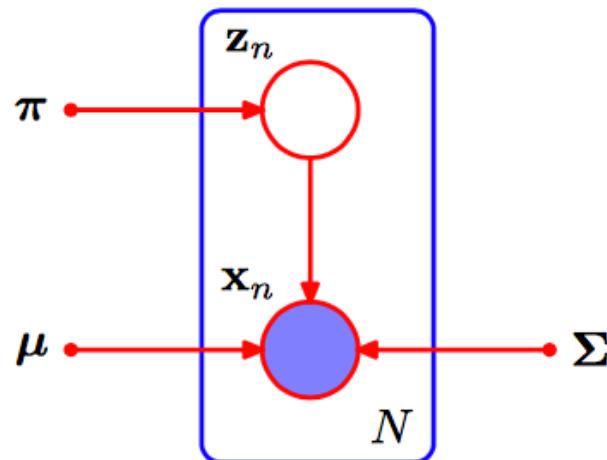
$$\begin{aligned}\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) &= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^K p(z_j = 1)p(\mathbf{x}|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.\end{aligned}$$

Mixture Models

- We want to solve

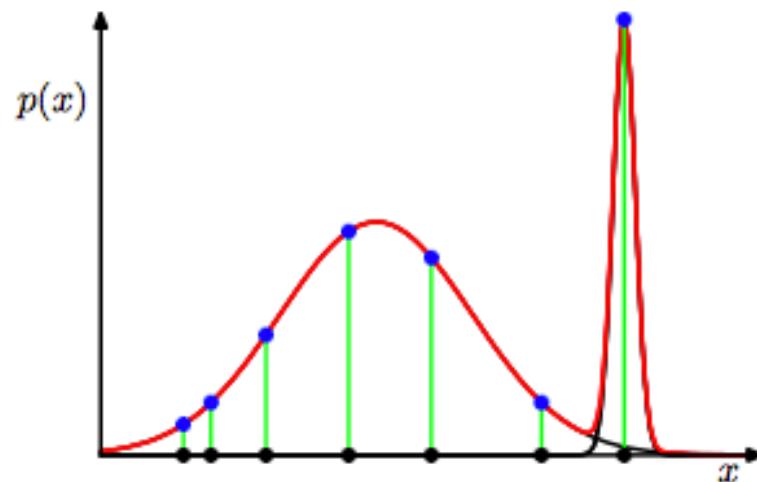
$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

- This could have been solved with optimization.
- Instead we solve it with latent variables z
- Graphical model



Beware of pitfalls of GM models

- Collapse onto a point: 2nd Gaussian can simply decide to fit a single point with infinitely small error



- Identifiability: there are $K!$ equivalent solutions since we can swap their identities. No big deal, EM will give us one of them.

EM Solution

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n$$

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

$$N_k = \sum_{n=1}^N \gamma(z_{nk}).$$

- Take derivative wrt μ_k : $0 = - \sum_{n=1}^N \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_k (\mathbf{x}_n - \boldsymbol{\mu}_k)$
- Derivative wrt Σ_k : $\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k) (\mathbf{x}_n - \boldsymbol{\mu}_k)^T$
- Derivative wrt π_k subject to Lagrange multiplier due to $\sum_k \pi_k = 1$ constraint

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left(\sum_{k=1}^K \pi_k - 1 \right)$$

- Gives $0 = \sum_{n=1}^N \frac{\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda$. So $\lambda = -N$ and $\pi_k = \frac{N_k}{N}$

Summarizing EM for Gaussian Mixtures

- Iterative, needs more iterations than K-means
 - Note that K-means is EM in the limit of variance Σ constant and going to 0
1. Initialize the means μ_k , covariances Σ_k and mixing coefficients π_k , and evaluate the initial value of the log likelihood.
 2. **E step.** Evaluate the responsibilities using the current parameter values

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \quad (9.23)$$

Summarizing EM for Gaussian Mixtures

3. **M step.** Re-estimate the parameters using the current responsibilities

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n \quad (9.24)$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}) (\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})^T \quad (9.25)$$

$$\pi_k^{\text{new}} = \frac{N_k}{N} \quad (9.26)$$

where

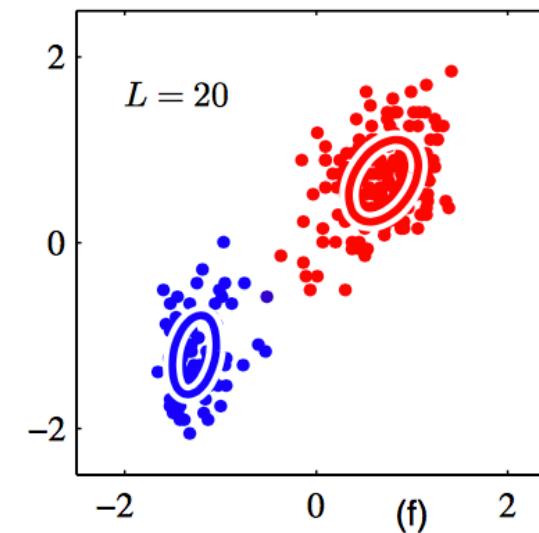
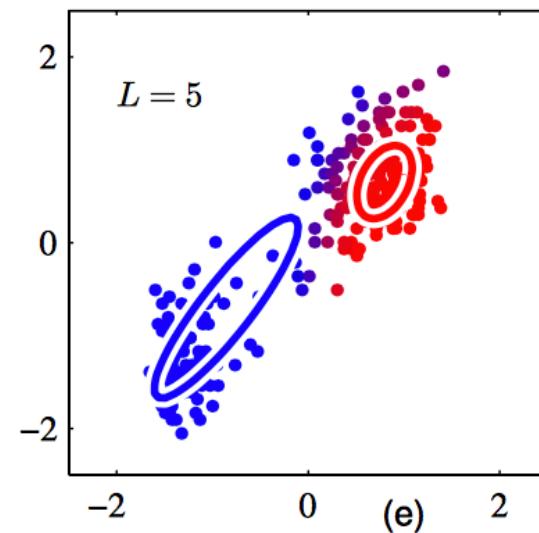
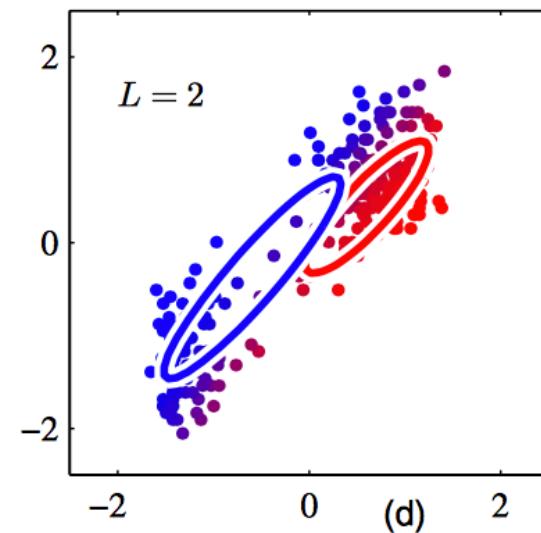
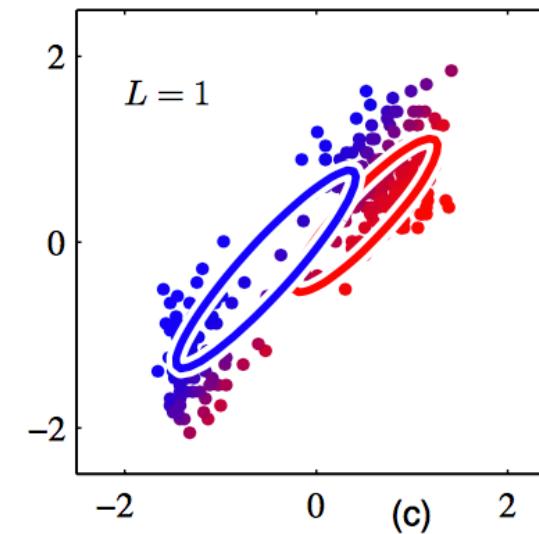
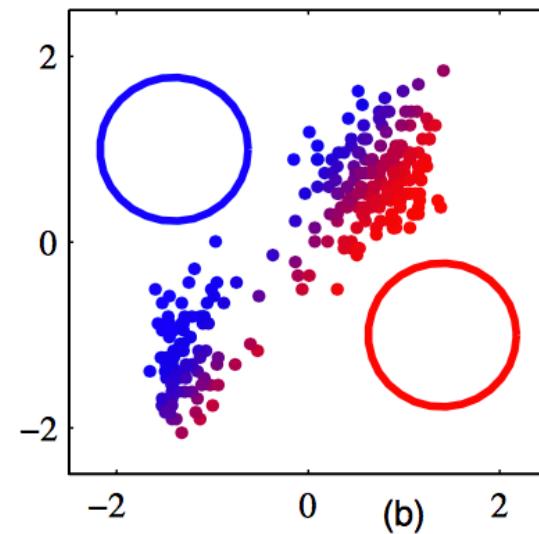
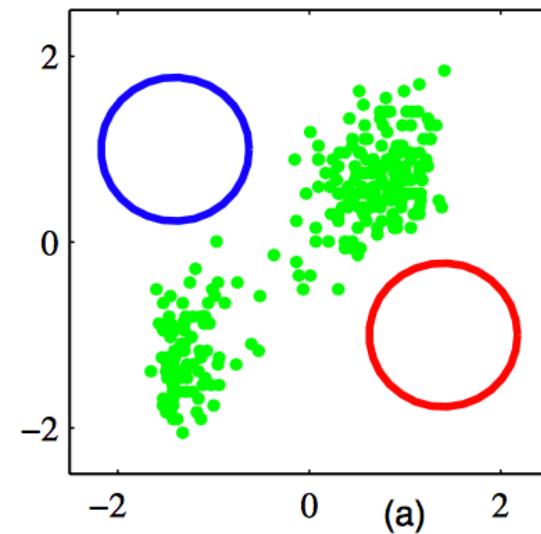
$$N_k = \sum_{n=1}^N \gamma(z_{nk}). \quad (9.27)$$

4. Evaluate the log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \quad (9.28)$$

and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.

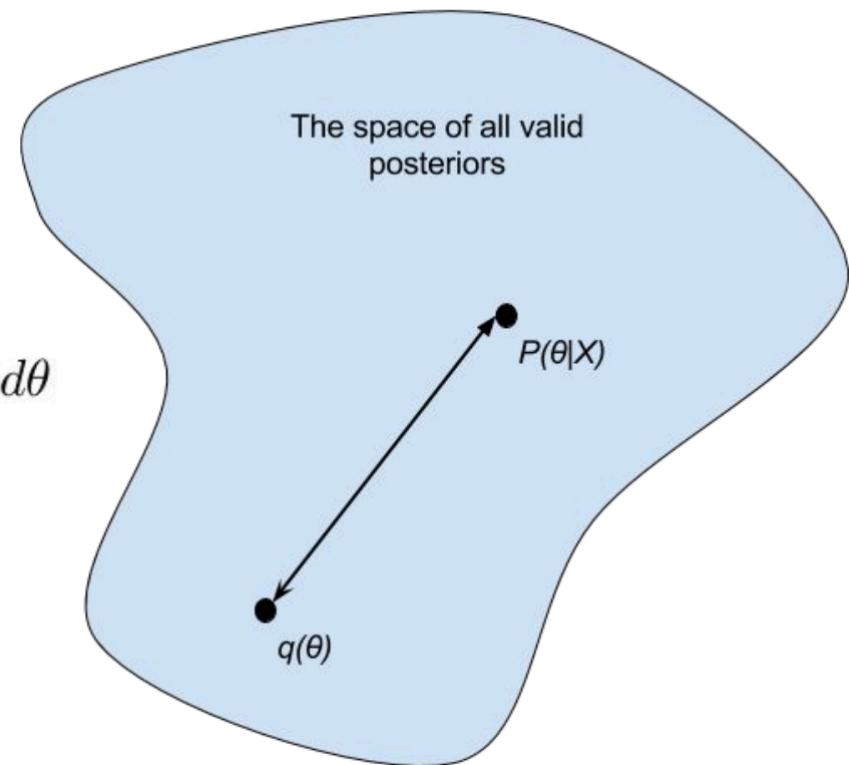
Example (same as before)



Variational Inference/ Bayes

- We want to approximate the posterior $P(\theta|X)$ using simple distributions $q(\theta)$ that are analytically tractable
- We do this by minimizing KL divergence

$$KL(q(\theta) || P(\theta|X)) = \int q(\theta) \log \frac{P(\theta|X)}{q(\theta)} d\theta$$



Why is this useful?

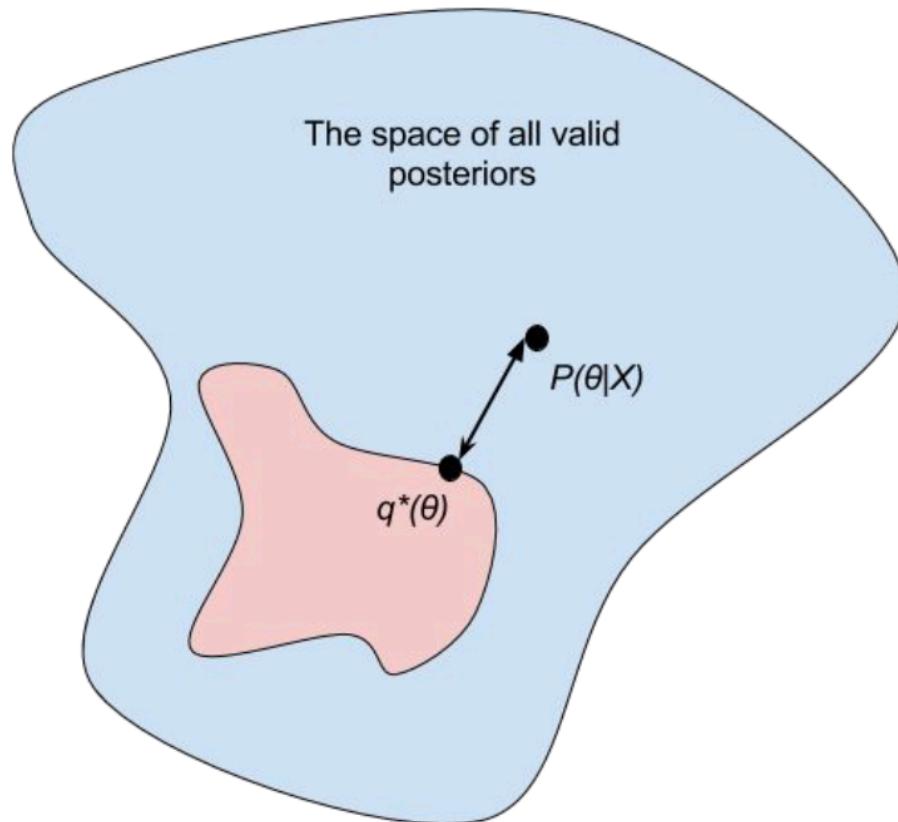
- We do not know the normalizing integral constant of $P(\theta|X)$ but we know it for $q(\theta)$

$$\begin{aligned} P(\theta|X) &= \operatorname{argmin}_q KL(q(\theta) || P(\theta|X)) \\ &= \operatorname{argmin}_q \int q(\theta) \log \frac{q(\theta)}{P(\theta|X)} d\theta \\ &= \operatorname{argmin}_q \left\{ \int q(\theta) \log q(\theta) d\theta - \int q(\theta) \log P(\theta, X) P(\theta) d\theta - P(X) \right\} \\ &= \operatorname{argmax}_q \left\{ - \underbrace{\int q(\theta) \log q(\theta) d\theta}_{\text{Entropy of approximation}} + \underbrace{\int q(\theta) \log P(\theta, X) P(\theta) d\theta}_{\text{Data fit (without the normalizing constant!)}} \right\} \end{aligned}$$

We limit $q(\theta)$ to tractable distributions

- Entropies are hard to compute except for tractable distributions
- We find $q^*(\theta)$ that minimizes KL distance in this space
- Mean field approach:

$$\mathcal{Q} = \left\{ q(\theta) = \prod_k q(\theta_k) \right\}$$



Bivariate Gaussian Example

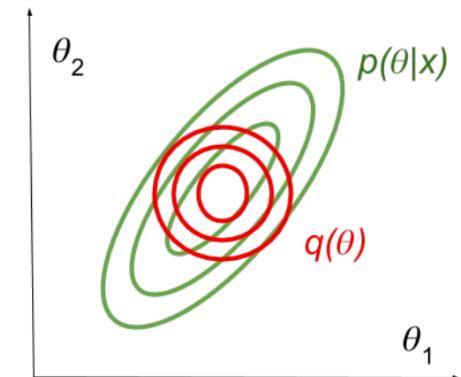
- MFVB does a good job at finding the mean
- MFVB does not describe correlations and tends to underestimate the variance

$$\mathcal{Q} = \{q(\theta) = \mathcal{N}(\theta_1; \mu_1, \sigma_1^2) \mathcal{N}(\theta_2; \mu_2, \sigma_2^2)\}$$

$$\eta_1 = (\mu_1, \sigma_1^2)$$

$$\eta_2 = (\mu_2, \sigma_2^2)$$

$$\eta = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$$

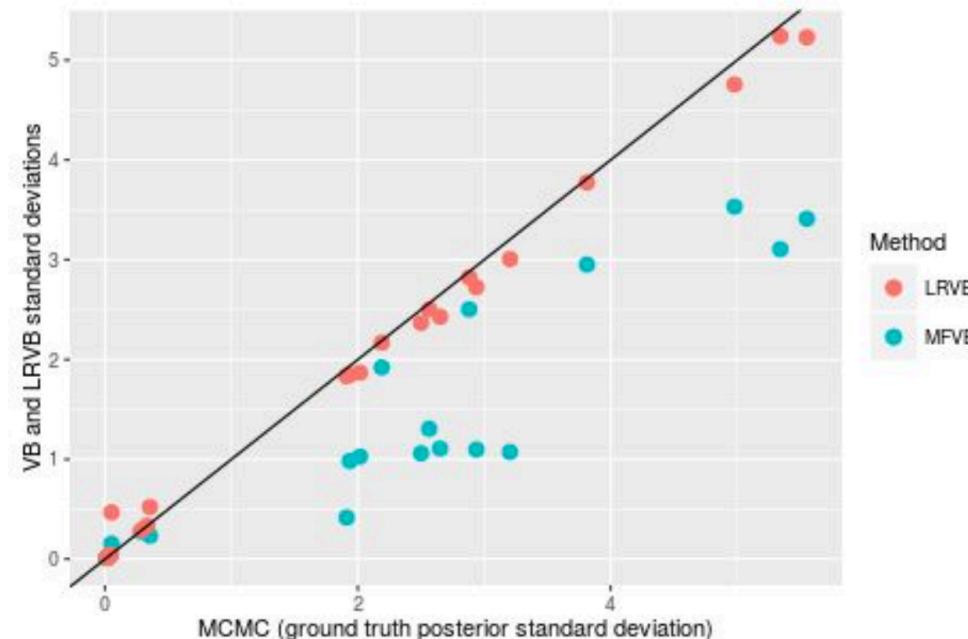


VB and EM

- EM can be viewed as a special case of VB where $q(\theta, Z) = \delta(\theta - \theta_0)q(Z)$
- E step: update $q(Z)$ keeping θ_0 fixed
- M step: update θ_0 at fixed Z

Why use (or not) VB?

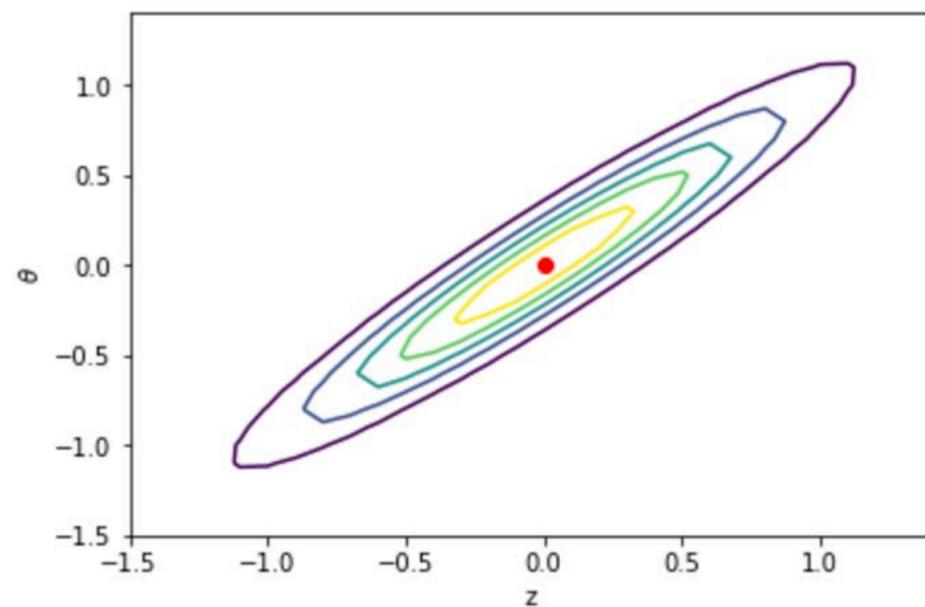
- Very fast compared to MCMC
- Typically gives good means
- Mean field often fails on variance
- Recent developments (ADVI, LRVB) improve on MFVB variance, but still no full posteriors



Example: MAP/Laplace vs. MFVB

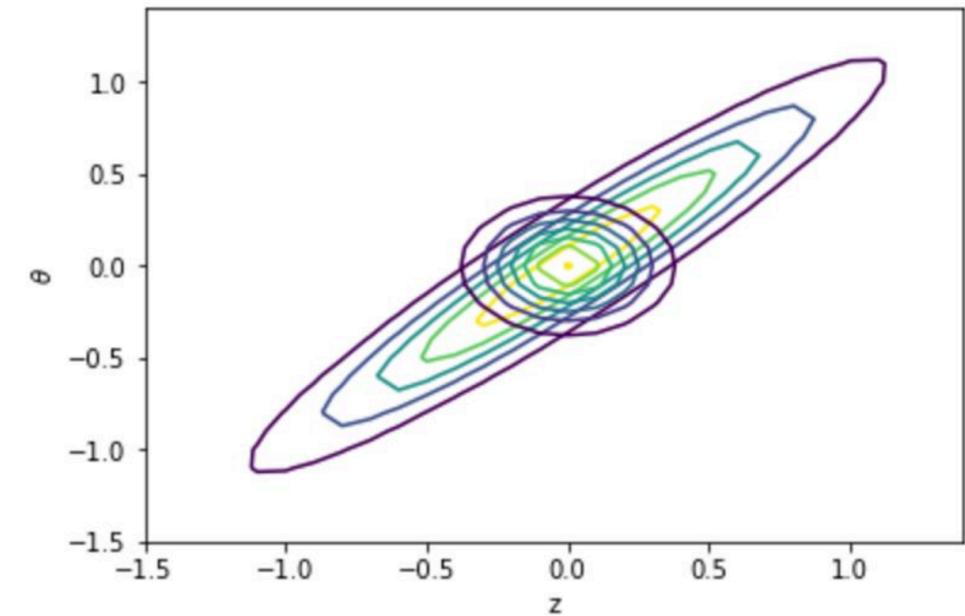
- On a multivariate Gaussian

MAP+Laplace

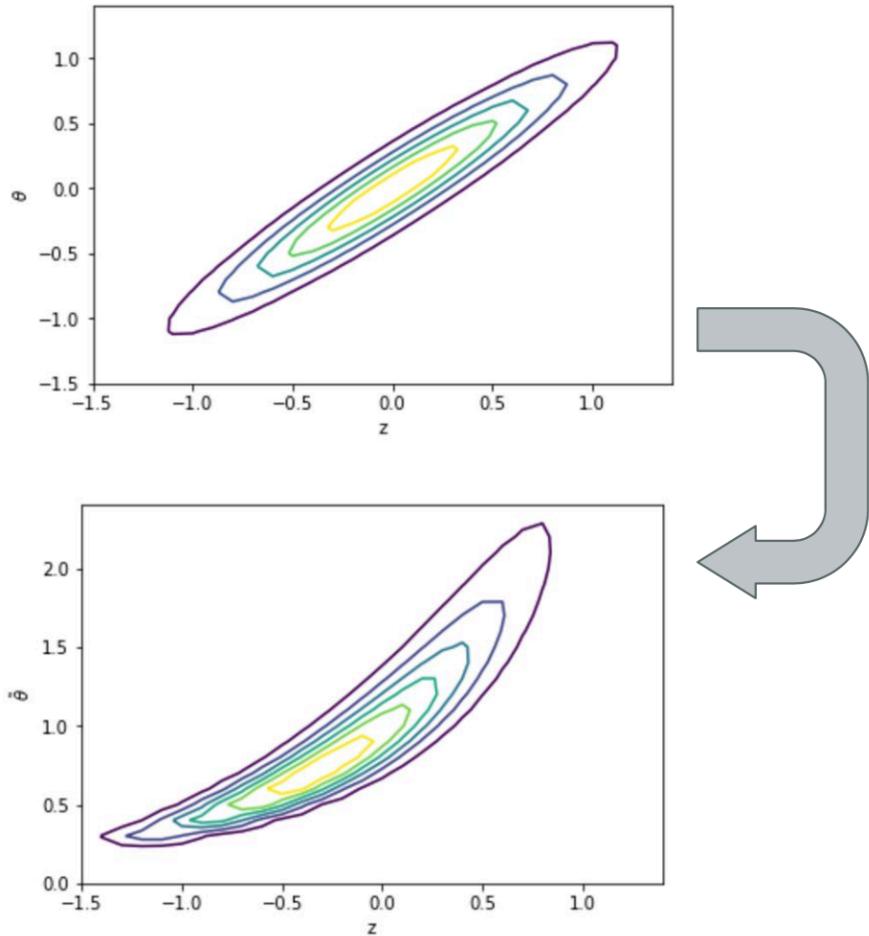


beats

MFVB



Example: bad banana

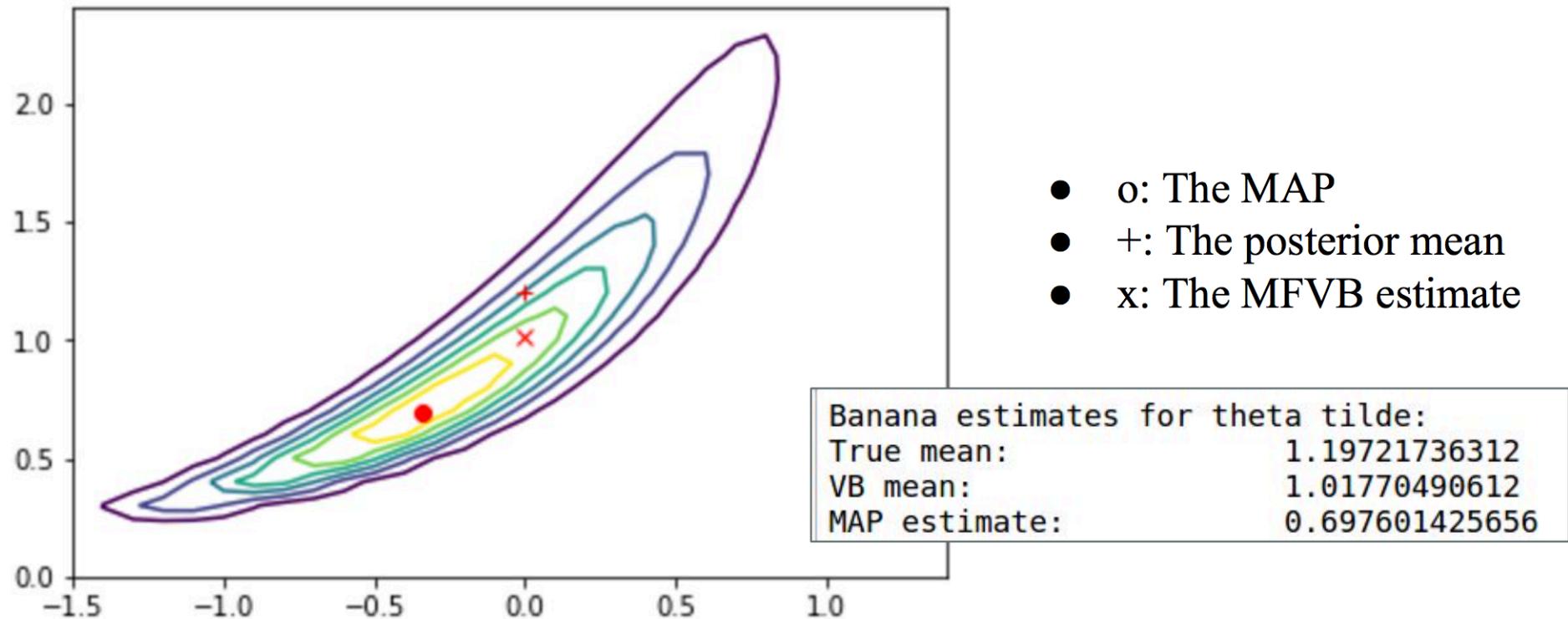


Suppose we instead had modeled

$$\begin{aligned}\tilde{\theta} &= \exp(\theta) \\ P_{\tilde{\theta}, z}(\tilde{\theta}, z) &= P_{\theta, z}(\log \tilde{\theta}, z) \frac{d\theta}{d\tilde{\theta}} \\ &= P_{\theta, z}(\log \tilde{\theta}, z) \exp(-\theta)\end{aligned}$$

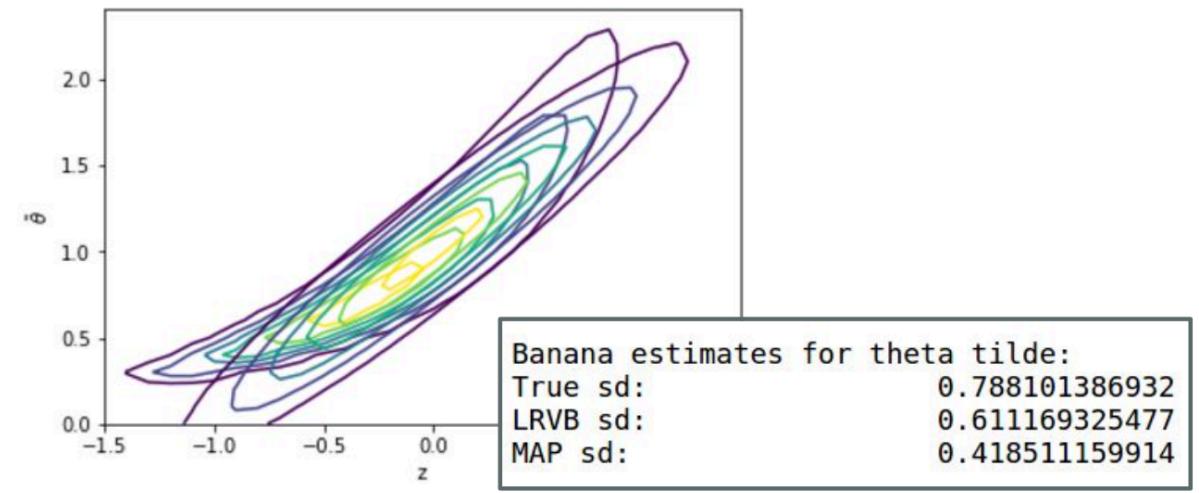
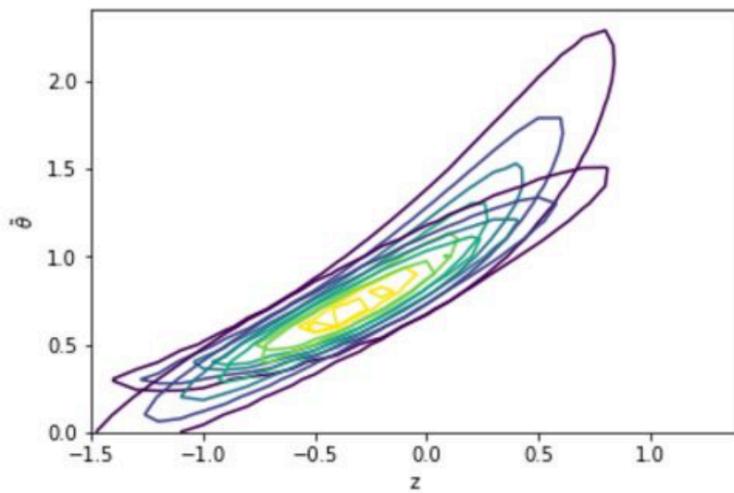
In $(\tilde{\theta}, Z)$ space the problem is not as easy.

Both MAP and MFVB get mean wrong



- MFVB is better than MAP on the mean

Covariances for MAP can also be wrong, but so are for MFVB and LRVB



Neyman-Scott “Paradox”

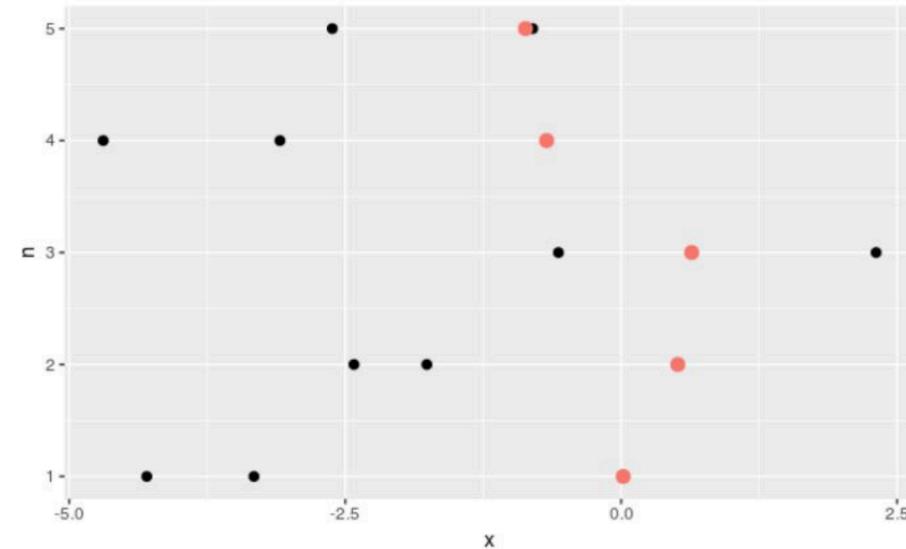
- Setup: we have N experiments, each measures $M = 2$ data X_{1n}, X_{2n} . Experiments are trying to determine the variance θ . However, there is an unknown mean offset for each experiment z_n .

For $n = 1, \dots, N$

$$X_{1n} \sim \mathcal{N}(z_n, \theta)$$

$$X_{2n} \sim \mathcal{N}(z_n, \theta)$$

We will investigate the “joint maximum likelihood estimator”.



Means are easy enough

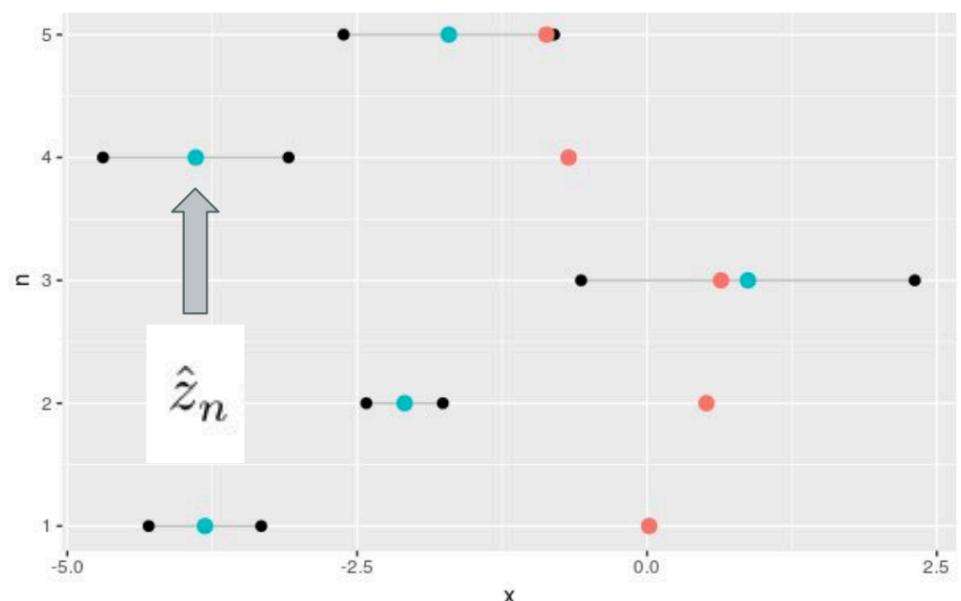
For $n = 1, \dots, N$

$$X_{1n} \sim \mathcal{N}(z_n, \theta)$$

$$X_{2n} \sim \mathcal{N}(z_n, \theta)$$

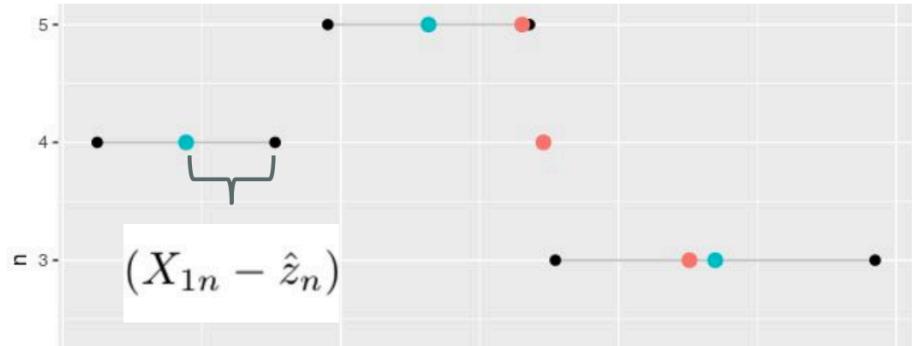
Irrespective of θ ,

$$\begin{aligned}\hat{z}_n &= \operatorname{argmax}_{z_n} P(X_{1n}, X_{2n} | z_n, \theta) \\ &= \frac{X_{1n} + X_{2n}}{2}\end{aligned}$$



How about variance θ ?

$$\hat{z}_n = \frac{X_{1n} + X_{2n}}{2}$$



$$\hat{\theta} = \operatorname{argmax}_{\theta} P(X_{1n}, X_{2n} | \hat{z}_n, \theta)$$

$$\begin{aligned} &= \frac{1}{2} \left(\frac{1}{N} \sum_n (X_{1n} - \hat{z}_n)^2 + \frac{1}{N} \sum_n (X_{2n} - \hat{z}_n)^2 \right) \\ &= \frac{1}{4N} \sum_n (X_{1n} - X_{2n})^2 \end{aligned}$$

- In intro labs/statistics courses we learn that the variance is computed from mean square distance of each point from mean divided by number of measurements M (here 2) if mean is known and divided by $M-1$ (here 1) if unknown. Where did it come from?

Even for large N , MLE is biased for low M by $M/(M-1)$ (=2 here)

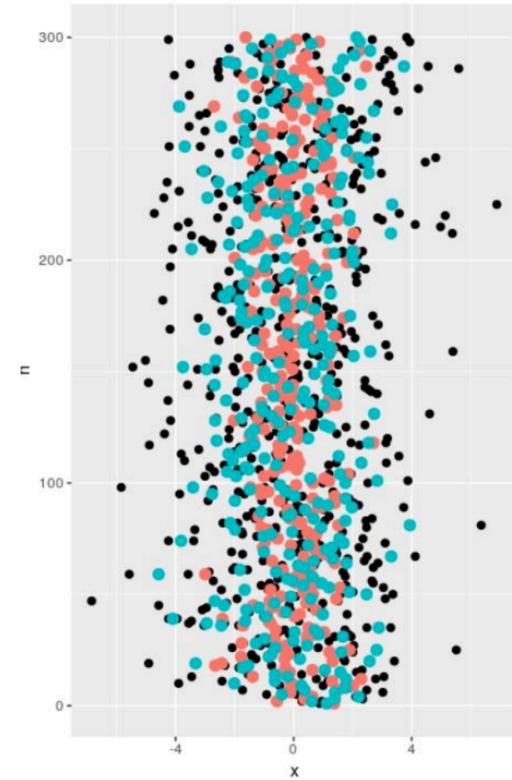
$$\hat{z}_n = \frac{X_{1n} + X_{2n}}{2} \quad \hat{\theta} = \frac{1}{4N} \sum_n (X_{1n} - X_{2n})^2$$

What does our estimate converge to as we get more data?

$$\mathbb{E} [(X_{1n} - X_{2n})^2] = \mathbb{E} [\mathbb{E} [(X_{1n} - X_{2n})^2 | z_n]] = 2\theta$$

So

$$\hat{\theta} \xrightarrow{n \rightarrow \infty} \frac{1}{4} 2\theta = \frac{\theta}{2} \neq \theta$$



- We failed to account for uncertainty in mean z_n : we only measure it from 2 data points
- We need to marginalize over z_n

MAP/MLE vs. Bayes

- We see that MAP/MLE is strongly biased here even in large N limit: one has to be careful with asymptotics theorems
- Full Bayesian analysis (e.g. MCMC) gives posterior of θ marginalized over z_n and automatically takes care of the problem. Bayesian analysis gives correct answer, i.e. it gives $M/(M-1)$ correction without “thinking”.
- EM also solves this problem correctly: it gives point estimator of θ averaging over z_n . So frequentist analyses that perform marginals over latent variables can also be correct “without thinking” (or without simulations telling us there is a problem).
- VB solves it too, and converges to the correct answer
- Lesson: sometimes we need to account for uncertainty in latent variables by marginalizing over them, even if we just want point estimators

Summary

- MCMC is great, but slow
- EM is a point estimator (like MAP/MLE) which marginalizes over latent variables
- Its Bayesian generalization is VB
- Both of these are able to perform marginalization and solve Neyman-Scott paradox, while MLE/MAP fails
- VB is not perfect and can provide wrong means or variances, and is never used for full posteriors

Literature

- *Bayesian Data Analysis*, Gelman et al. , Chapter 13
- D. Mackay, *Information Theory, Inference, and Learning Algorithms* (See course website), Chapter 33