A Bayesian Perspective on EM Covariances

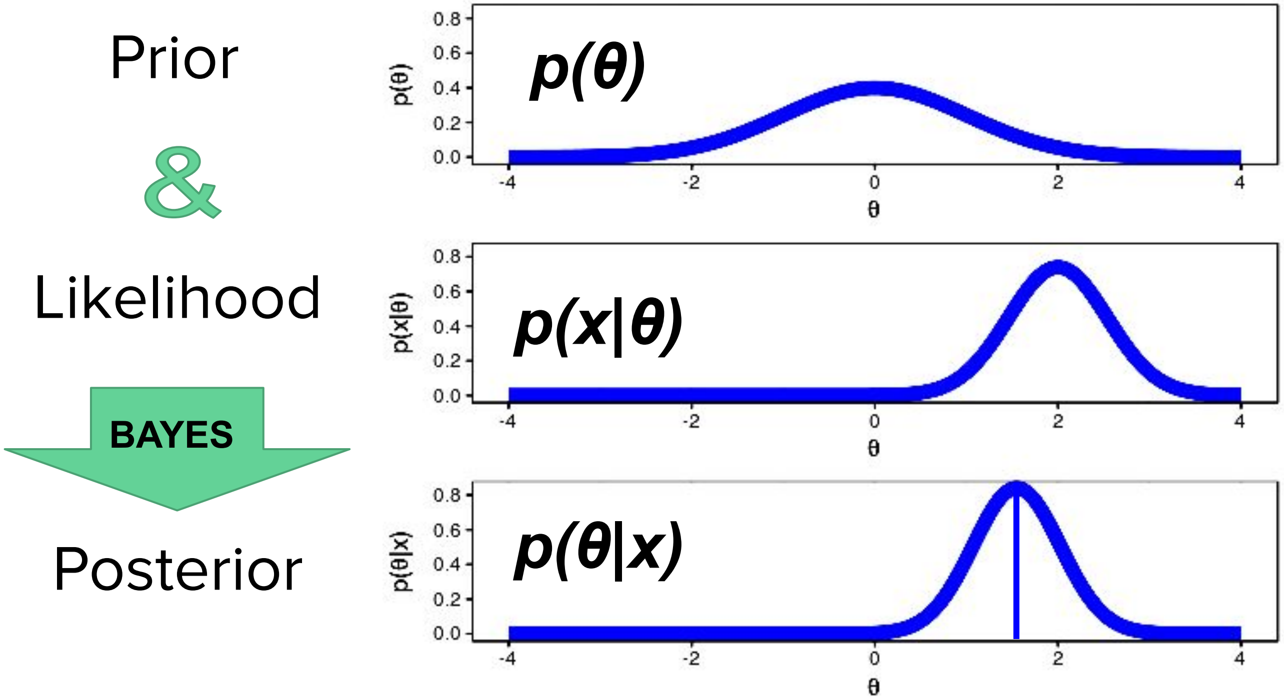
Ryan Giordano, March 2019

Outline

- **Prelude:** Linear response covariances.
- **Part 1:** The standard view of EM.
- **Part 2:** A Bayesian view of EM.
- **Part 3:** Covariance asymptotics.
- **Postscript:** Tools.

Prelude:
Linear response covariances.

The Bayesian Machinery



The Bayesian Machinery

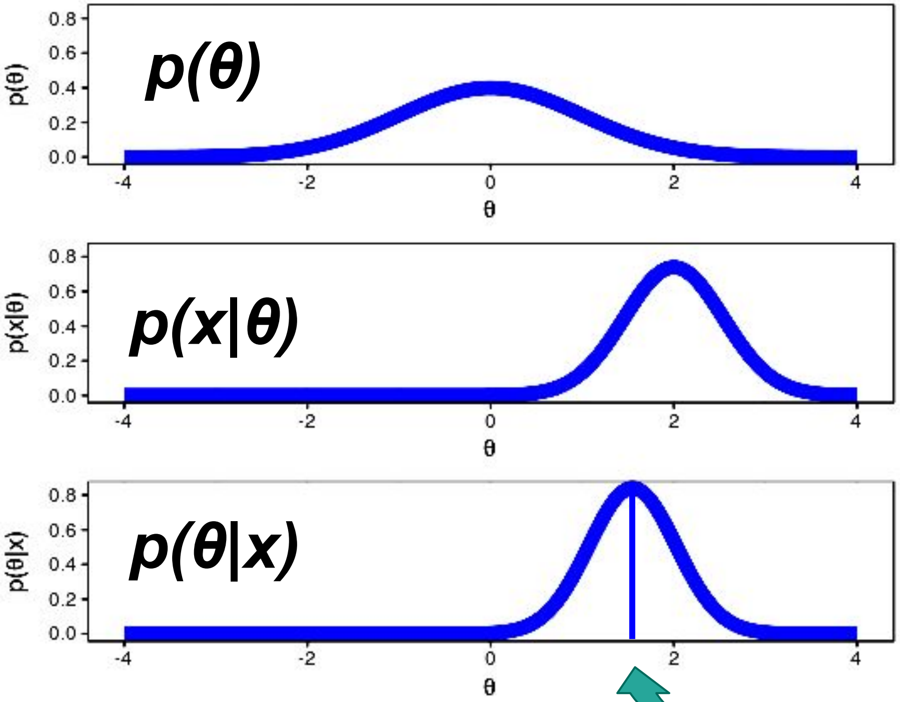
Prior

&

Likelihood

BAYES

Posterior



Summarize posteriors with a "posterior expectation":

$$\mathbb{E}_{p(\theta|X)}[\theta]$$

Everything has “hyperparameters”

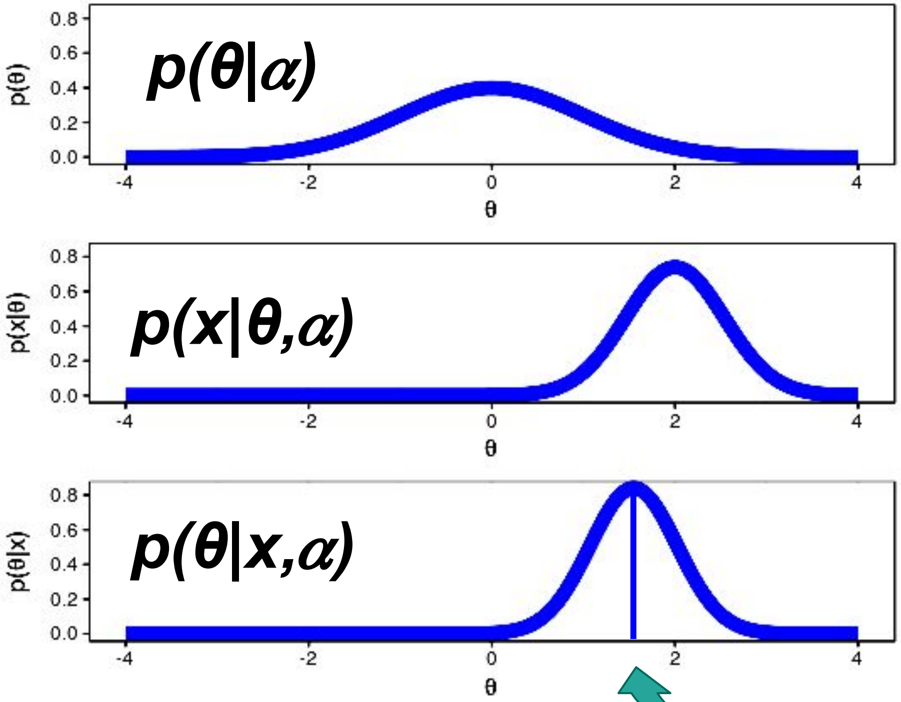
Prior

&

Likelihood

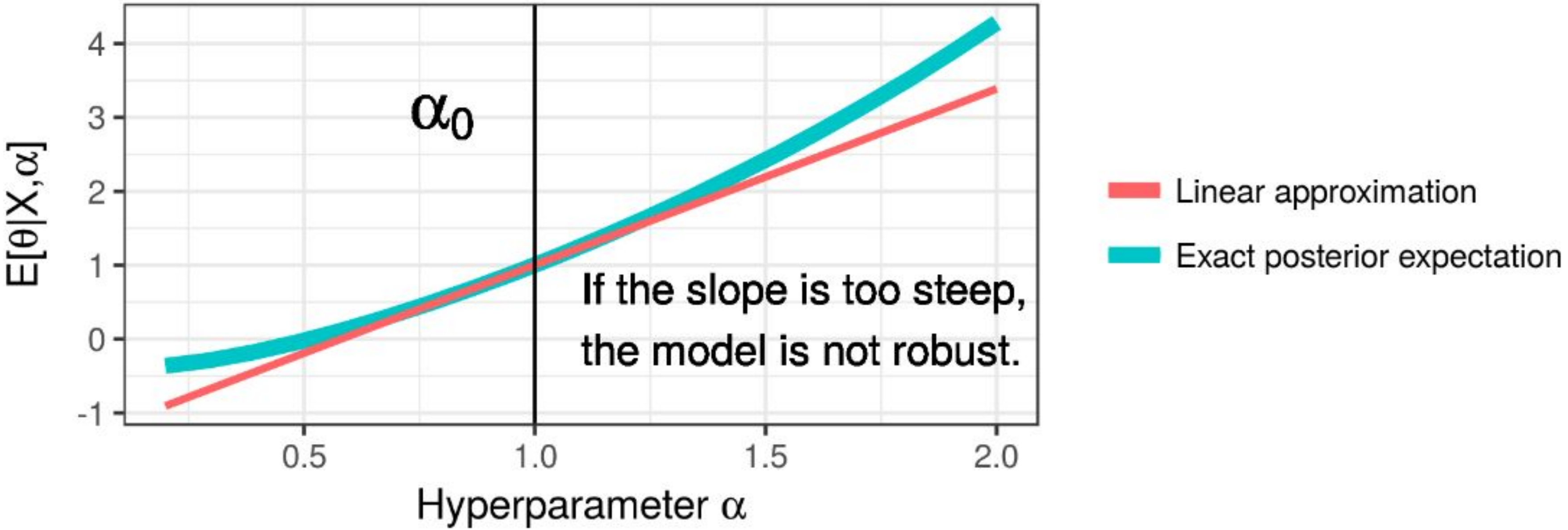
BAYES

Posterior



$$\mathbb{E}_{p(\theta|X,\alpha)}\left[\theta\right]$$

Sensitivity



Actual results calculated with
`https://github.com/rgiordan/StanSensitivity`

Sensitivity = Covariance

(exchange differentiation and integration)

$$\frac{d\mathbb{E}\left[\theta|\alpha, X\right]}{d\alpha} = \mathrm{Cov}_{p(\theta|\alpha,X)}\left(\theta, \underbrace{\frac{\partial}{\partial\alpha}\log p\left(\theta|\alpha, X\right)}_{\text{Some nasty derivative}}\right)$$

Classical Bayesian robustness:
Calculate the covariance to estimate the sensitivity

$$\frac{d\mathbb{E}\left[\theta|\alpha, X\right]}{d\alpha} = \mathrm{Cov}_{p(\theta|\alpha,X)}\left(\theta, \frac{\partial}{\partial\alpha}\log p\left(\theta|\alpha, X\right)\right)$$

Local Robustness in Bayesian Analysis
Paul Gustafson

Linear response covariances:

Calculate the sensitivity to estimate the covariance.

$$\frac{d\mathbb{E}\left[\theta|\alpha, X\right]}{d\alpha} = \mathrm{Cov}_{p(\theta|\alpha,X)}\left(\theta, \frac{\partial}{\partial\alpha}\log p\left(\theta|\alpha, X\right)\right)$$

Standard result from sensitivity analysis:

$$\hat{\theta}(t) = \underset{\theta}{\operatorname{argmin}} \left(f(\theta) + t\theta \right)$$

Standard result from sensitivity analysis:

$$\hat{\theta}(t) = \underset{\theta}{\operatorname{argmin}} \left(f(\theta) + t\theta \right)$$

$$\left.\frac{d\hat{\theta}}{dt}\right|_{t=0} = \left(\left.\frac{\partial^2 f(\theta)}{\partial\theta\partial\theta}\right|_{\hat{\theta}} \right)^{-1}$$

An inverse Hessian!

Part 1:
The standard view of EM.

Parameters and data:

Data: $Y = (Y_1, ..., Y_N)$

Latent variables: $Z = (Z_1, ..., Z_N)$

Parameters: $\theta \in \Omega_\theta \subseteq \mathbb{R}^D$

Parameters and data:

Data:	$Y = (Y_1, ..., Y_N)$	Observed, grows with N
Latent variables:	$Z = (Z_1, ..., Z_N)$	Unobserved, grows with N
Parameters:	$\theta \in \Omega_\theta \subseteq \mathbb{R}^D$	Unobserved, fixed dimension

Parameters and data:

Data: $Y = (Y_1, ..., Y_N)$

Latent variables: $Z = (Z_1, ..., Z_N)$

Parameters: $\theta \in \Omega_\theta \subseteq \mathbb{R}^D$

Generative process:

$p(\theta) = \text{A given prior.}$

Parameters and data:

Data: $Y = (Y_1, ..., Y_N)$

Latent variables: $Z = (Z_1, ..., Z_N)$

Parameters: $\theta \in \Omega_\theta \subseteq \mathbb{R}^D$

Generative process:

$$p(Z|\theta) = \prod_{n=1}^{N} p(Z_n|\theta)$$

$$p(\theta) = \text{A given prior.}$$

Parameters and data:

Data: $Y = (Y_1, ..., Y_N)$

Latent variables: $Z = (Z_1, ..., Z_N)$

Parameters: $\theta \in \Omega_\theta \subseteq \mathbb{R}^D$

Generative process:

$$p(Y|\theta, Z) = \prod_{n=1}^{N} p(Y_n|Z_n, \theta)$$

$$p(Z|\theta) = \prod_{n=1}^{N} p(Z_n|\theta)$$

$$p(\theta) = \text{A given prior.}$$

Parameters and data:

Equivalent generative process:

Data: $Y = (Y_1, ..., Y_N)$

$$p(Y|\theta) = \prod_{n=1}^{N} \int p(Y_n|Z_n, \theta)\, p(Z_n|\theta)\, dZ_n$$

Latent variables: $Z = (Z_1, ..., Z_N)$

$$p(Z|\theta) = \prod_{n=1}^{N} p(Z_n|\theta)$$

Parameters: $\theta \in \Omega_\theta \subseteq \mathbb{R}^D$

$$p(\theta) = \text{A given prior.}$$

Why not estimate with $$\hat{\theta}, \hat{Z} = \underset{\theta, Z}{\operatorname{argmax}}\, p\,(Y, Z|\theta)$$ **?**

Why not estimate with $\hat{\theta}, \hat{Z} = \underset{\theta, Z}{\operatorname{argmax}}\, p\left(Y, Z | \theta\right)$ **?**

$$p\left(Y_n | \theta\right) = \int p\left(Y_n, Z_n | \theta\right) dZ_n \neq p\left(Y_n, \hat{Z}_n | \theta\right)$$

...unless $p\left(Z_n | Y_n, \theta\right)$ is concentrated.
(Roughly speaking.)

We do EM when:

Hard:

$$p\left(Y|\theta\right) = \prod_{n=1}^{N} \int p\left(Y_n|Z_n, \theta\right) p\left(Z_n|\theta\right) dZ_n$$

Easy:

$$p\left(Y|\theta, Z\right) = \prod_{n=1}^{N} p\left(Y_n|Z_n, \theta\right)$$

Dispersed:

$$p\left(Z_n|Y_n, \theta\right)$$

We do EM when:

Hard:

$$p\left(Y|\theta\right) = \prod_{n=1}^{N} \int p\left(Y_n|Z_n, \theta\right) p\left(Z_n|\theta\right) dZ_n$$

Easy:

$$p\left(Y|\theta, Z\right) = \prod_{n=1}^{N} p\left(Y_n|Z_n, \theta\right)$$

Dispersed:
(and easy)

$$p\left(Z_n|Y_n, \theta\right)$$

Notation for log probabilities.

$$\ell\left(Y|\theta\right) = \log p\left(Y|\theta\right)$$

(and in general)

Assume the MLE is nice.

Consistent:

$$\hat{\theta} = \underset{\theta \in \Omega_\theta}{\operatorname{argmax}}\, \ell\left(Y|\theta\right)$$

$$\hat{\theta} \xrightarrow[N \to \infty]{} \theta_0$$

Assume the MLE is nice.

Consistent:

$$\hat{\theta} = \underset{\theta \in \Omega_\theta}{\operatorname{argmax}} \, \ell\left(Y \middle| \theta\right)$$

$$\hat{\theta} \xrightarrow[N \to \infty]{} \theta_0$$

Asymptotically normal:

$$\hat{\mathcal{I}}_{\theta\theta} := -\frac{1}{N} \left. \frac{\partial^2 \ell\left(Y \middle| \theta\right)}{\partial\theta\partial\theta} \right|_{\hat{\theta}}$$

$$\hat{\Sigma} := \hat{\mathcal{I}}_{\theta\theta}^{-1}$$

$$\sqrt{N}\hat{\Sigma}^{-1/2}\left(\hat{\theta} - \theta_0\right) \xrightarrow[N \to \infty]{} \mathcal{N}\left(0, I_D\right)$$

Hard to calculate
(by assumption)

$$\hat{\theta} = \underset{\theta \in \Omega_\theta}{\operatorname{argmax}} \, \ell\left(Y | \theta\right)$$

$$\hat{\theta} \xrightarrow[N \to \infty]{} \theta_0$$

$$\hat{\mathcal{I}}_{\theta\theta} := -\frac{1}{N} \left. \frac{\partial^2 \ell\left(Y | \theta\right)}{\partial\theta\partial\theta} \right|_{\hat{\theta}}$$

$$\hat{\Sigma} := \hat{\mathcal{I}}_{\theta\theta}^{-1}$$

$$\sqrt{N}\hat{\Sigma}^{-1/2}\left(\hat{\theta} - \theta_0\right) \xrightarrow[N \to \infty]{} \mathcal{N}\left(0, I_D\right)$$

The “EM Identity”

$$\underbrace{\ell\left(Y_n|\theta\right)}_{\text{Hard}} = \mathbb{E}_{p(Z_n|Y_n,\theta)}\left[\ell\left(Y_n|Z_n,\theta\right) + \ell\left(Z_n|\theta\right) - \ell\left(Z_n|Y_n,\theta\right)\right]$$

The "EM Identity"

$$\underbrace{\ell\left(Y_n|\theta\right)}^{\text{Hard}} = \mathbb{E}_{\underbrace{p(Z_n|Y_n,\theta)}_{\text{Easy}}}\left[\underbrace{\ell\left(Y_n|Z_n,\theta\right) + \ell\left(Z_n|\theta\right)}_{\text{Easy}} - \underbrace{\ell\left(Z_n|Y_n,\theta\right)}_{\text{Easy}}\right]$$

Proof.

$$\ell\left(Y_n|\theta\right) = \mathbb{E}_{p(Z_n|Y_n,\theta)}\left[\overbrace{\ell\left(Y_n|Z_n,\theta\right)}^{\text{Apply Bayes' rule}} + \ell\left(Z_n|\theta\right) - \ell\left(Z_n|Y_n,\theta\right)\right]$$

Proof.

$$
\begin{aligned}
\ell\left(Y_n|\theta\right) &= \mathbb{E}_{p(Z_n|Y_n,\theta)}\left[\overbrace{\ell\left(Y_n|Z_n,\theta\right)}^{\text{Apply Bayes' rule}} + \ell\left(Z_n|\theta\right) - \ell\left(Z_n|Y_n,\theta\right)\right] \\
&= \mathbb{E}_{p(Z_n|Y_n,\theta)}\left[\left(\ell\left(Z_n|Y_n,\theta\right) + \ell\left(Y_n|\theta\right) - \ell\left(Z_n|\theta\right)\right) + \ell\left(Z_n|\theta\right) - \ell\left(Z_n|Y_n,\theta\right)\right]
\end{aligned}
$$

Proof.

$$
\begin{aligned}
\ell\left(Y_n|\theta\right) &= \mathbb{E}_{p(Z_n|Y_n,\theta)}\left[\overbrace{\ell\left(Y_n|Z_n,\theta\right)}^{\text{Apply Bayes' rule}} + \ell\left(Z_n|\theta\right) - \ell\left(Z_n|Y_n,\theta\right)\right] \\
&= \mathbb{E}_{p(Z_n|Y_n,\theta)}\left[\left(\ell\left(Z_n|Y_n,\theta\right) + \ell\left(Y_n|\theta\right) - \ell\left(Z_n|\theta\right)\right) + \ell\left(Z_n|\theta\right) - \ell\left(Z_n|Y_n,\theta\right)\right] \\
&= \mathbb{E}_{p(Z_n|Y_n,\theta)}\left[\ell\left(Y_n|\theta\right)\right]
\end{aligned}
$$

Proof.

$$
\begin{aligned}
\ell\left(Y_n|\theta\right) &= \mathbb{E}_{p(Z_n|Y_n,\theta)}\left[\overbrace{\ell\left(Y_n|Z_n,\theta\right)}^{\text{Apply Bayes' rule}} + \ell\left(Z_n|\theta\right) - \ell\left(Z_n|Y_n,\theta\right)\right] \\
&= \mathbb{E}_{p(Z_n|Y_n,\theta)}\left[\left(\ell\left(Z_n|Y_n,\theta\right) + \ell\left(Y_n|\theta\right) - \ell\left(Z_n|\theta\right)\right) + \ell\left(Z_n|\theta\right) - \ell\left(Z_n|Y_n,\theta\right)\right] \\
&= \mathbb{E}_{p(Z_n|Y_n,\theta)}\left[\ell\left(Y_n|\theta\right)\right] \\
&= \ell\left(Y_n|\theta\right)
\end{aligned}
$$

Proof.

$$\begin{aligned}
\ell\left(Y_n|\theta\right) &= \mathbb{E}_{p(Z_n|Y_n,\theta)}\left[\ell\left(Y_n|Z_n,\theta\right) + \ell\left(Z_n|\theta\right) - \ell\left(Z_n|Y_n,\theta\right)\right] \\
&= \mathbb{E}_{p(Z_n|Y_n,\theta)}\left[\left(\ell\left(Z_n|Y_n,\theta\right) + \ell\left(Y_n|\theta\right) - \ell\left(Z_n|\theta\right)\right) + \ell\left(Z_n|\theta\right) - \ell\left(Z_n|Y_n,\theta\right)\right] \\
&= \mathbb{E}_{p(Z_n|Y_n,\theta)}\left[\ell\left(Y_n|\theta\right)\right] \\
&= \ell\left(Y_n|\theta\right)
\end{aligned}$$

This view of EM can simplify some EM proofs.

The "EM Identity"

$$\left.\hat{\theta} = \underset{\theta \in \Omega_\theta}{\operatorname{argmax}}\, \ell\left(Y|\theta\right)\right\} \text{Hard}$$

$$= \underset{\theta \in \Omega_\theta}{\operatorname{argmax}}\, \underbrace{\mathbb{E}_{p(Z|Y,\theta)}}_{\text{Easy}} \left[\underbrace{\ell\left(Y|Z,\theta\right) + \ell\left(Z|\theta\right)}_{\text{Easy}} - \underbrace{\ell\left(Z|Y,\theta\right)}_{\text{Easy}}\right]$$

The "EM algorithm"

Given iteration k, $\theta^{(k)}$:

Step 1k. "E step":

1. Calculate $$Q^{(k)}(\theta) = \mathbb{E}_{\underbrace{p\left(Z|Y,\theta^{(k)}\right)}_{\text{Fixed}}} \left[\underbrace{\ell\left(Y|Z,\theta\right) + \ell\left(Z|\theta\right)}_{\text{A function of } \theta} - \underbrace{\ell\left(Z|Y,\theta^{(k)}\right)}_{\text{Fixed (typically omitted)}} \right]$$

The "EM algorithm"

Given iteration k, $\theta^{(k)}$:

Step 1k. "E step":

1. Calculate $Q^{(k)}(\theta) = \mathbb{E}_{p\left(Z|Y,\theta^{(k)}\right)}\left[\underbrace{\ell(Y|Z,\theta) + \ell(Z|\theta)}_{\text{A function of } \theta} - \ell\left(Z|Y,\theta^{(k)}\right)\right]$

Step 2k. "M step":

Calculate $\theta^{(k+1)} = \underset{\theta \in \Omega_\theta}{\operatorname{argmax}}\, Q^{(k)}(\theta)$

...repeat.

EM algorithm ≠ EM identity

Under nice conditions, the EM algorithm solves the same optimization problem as the MLE.

$$\theta^{(k)} \xrightarrow[N \to \infty]{} \hat{\theta}$$

But what about covariances?

What about covariances?

We want the Hessian of the marginal log likelihood:

$$\left.\frac{\partial^2 \ell\left(Y|\theta\right)}{\partial\theta\partial\theta}\right|_{\hat{\theta}}$$

What about covariances?

We want the Hessian of the marginal log likelihood:

$$\left.\frac{\partial^2 \ell\left(Y|\theta\right)}{\partial\theta\partial\theta}\right|_{\hat{\theta}}$$

All we have is the Q function:

$$\hat{Q}\left(\theta\right) = \mathbb{E}_{p\left(Z|Y,\hat{\theta}\right)}\left[\ell\left(Y|Z,\theta\right) + \ell\left(Z|\theta\right) - \ell\left(Z|Y,\hat{\theta}\right)\right]$$

What about covariances?

...but the Hessians are not the same.

$$\left.\frac{\partial^2 \ell\left(Y|\theta\right)}{\partial\theta\partial\theta}\right|_{\hat{\theta}} \neq \left.\frac{\partial^2 \hat{Q}\left(\theta\right)}{\partial\theta\partial\theta}\right|_{\hat{\theta}}$$

$$\hat{Q}\left(\theta\right) = \mathbb{E}_{p\left(Z|Y,\hat{\theta}\right)}\left[\ell\left(Y|Z,\theta\right) + \ell\left(Z|\theta\right) - \ell\left(Z|Y,\hat{\theta}\right)\right]$$

...but the Hessians are not the same.

$$\left.\frac{\partial^2 \ell\left(Y|\theta\right)}{\partial\theta\partial\theta}\right|_{\hat{\theta}} = \left.\frac{\partial^2}{\partial\theta\partial\theta}\mathbb{E}_{p(Z|Y,\theta)}\left[\ell\left(Y|Z,\theta\right) + \ell\left(Z|\theta\right) - \ell\left(Z|Y,\theta\right)\right]\right|_{\hat{\theta}}$$

(by the EM identity)

...but the Hessians are not the same.

$$
\begin{aligned}
\left.\frac{\partial^2 \ell\left(Y|\theta\right)}{\partial\theta\partial\theta}\right|_{\hat{\theta}} &= \left.\frac{\partial^2}{\partial\theta\partial\theta}\mathbb{E}_{p(Z|Y,\theta)}\left[\ell\left(Y|Z,\theta\right)+\ell\left(Z|\theta\right)-\ell\left(Z|Y,\theta\right)\right]\right|_{\hat{\theta}} \\
&\neq \left.\frac{\partial^2}{\partial\theta\partial\theta}\mathbb{E}_{p\left(Z_n|Y_n,\hat{\theta}\right)}\left[\ell\left(Y|Z,\theta\right)+\ell\left(Z|\theta\right)-\ell\left(Z|Y,\hat{\theta}\right)\right]\right|_{\hat{\theta}}
\end{aligned}
$$

(fix the “E step”)

...but the Hessians are not the same.

$$
\begin{aligned}
\left.\frac{\partial^2 \ell\left(Y|\theta\right)}{\partial\theta\partial\theta}\right|_{\hat{\theta}} &= \left.\frac{\partial^2}{\partial\theta\partial\theta}\mathbb{E}_{p(Z|Y,\theta)}\left[\ell\left(Y|Z,\theta\right)+\ell\left(Z|\theta\right)-\ell\left(Z|Y,\theta\right)\right]\right|_{\hat{\theta}} \\
&\neq \left.\frac{\partial^2}{\partial\theta\partial\theta}\mathbb{E}_{p\left(Z_n|Y_n,\hat{\theta}\right)}\left[\ell\left(Y|Z,\theta\right)+\ell\left(Z|\theta\right)-\ell\left(Z|Y,\hat{\theta}\right)\right]\right|_{\hat{\theta}} \\
&= \left.\frac{\partial^2 \hat{Q}\left(\theta\right)}{\partial\theta\partial\theta}\right|_{\hat{\theta}}
\end{aligned}
$$

(by definition)

...but the Hessians are not the same.

$$\left.\frac{\partial^2 \ell\left(Y|\theta\right)}{\partial\theta\partial\theta}\right|_{\hat{\theta}} \neq \left.\frac{\partial^2 \hat{Q}\left(\theta\right)}{\partial\theta\partial\theta}\right|_{\hat{\theta}}$$

Standard work-arounds are kinda complicated*.

And what about uncertainty in *Z*?

- Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm, Meng et al., 2001
- Direct calculation of the information matrix via the EM algorithm, Oakes, 1999
- The EM algorithm and extensions, McLachlan, G. and T. Krishnan, 2007

A Bayesian view of EM.

Let’s calculate the posterior.

$$p\left(\theta, Z|Y\right) = \frac{p\left(Y|Z, \theta\right) p\left(Z|\theta\right) p\left(\theta\right)}{p\left(Y\right)}$$

Let's calculate the posterior.

$$p(\theta, Z|Y) = \frac{\overbrace{p(Y|Z,\theta)\, p(Z|\theta)\, p(\theta)}^{\text{Hard}}}{p(Y)}$$

Let's ~~calculate~~ approximate the posterior.

$$p\left(\theta, Z \middle| Y\right) = \frac{\overbrace{p\left(Y \middle| Z, \theta\right) p\left(Z \middle| \theta\right) p\left(\theta\right)}^{\text{Hard}}}{p\left(Y\right)}$$

Let's ~~calculate~~ approximate the posterior.

$$p\left(\theta, Z|Y\right) = \frac{p\left(Y|Z,\theta\right)p\left(Z|\theta\right)p\left(\theta\right)}{p\left(Y\right)}$$

Variational Bayes (VB): find a *q(θ,Z)* that is

(a) Easy to deal with and

(b) Close to *p(θ,Z|Y)* in some sense

Variational Bayes.

Define a class of approximating distribution.

$$q\left(\theta, Z|\vartheta, \zeta\right) := \delta\left(\theta - \vartheta\right) q\left(Z|\zeta\right)$$

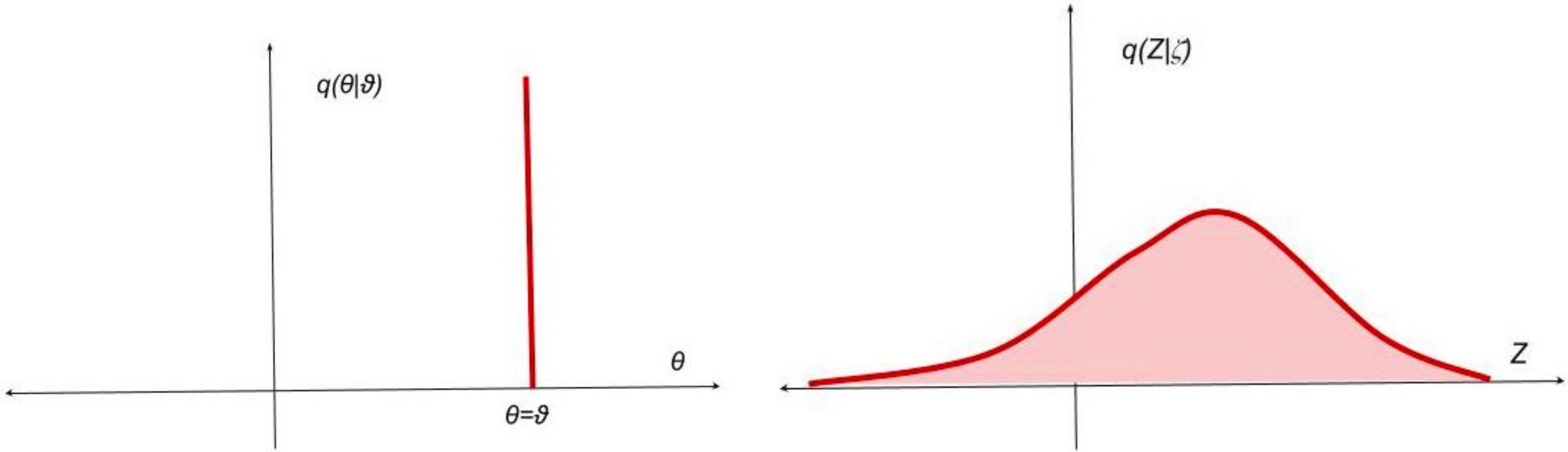
Variational Bayes.

Define a class of approximating distribution.

$$q\left(\theta, Z|\vartheta, \zeta\right) := \overbrace{\delta\left(\theta - \vartheta\right)}^{\text{Degenerate at } \vartheta} \underbrace{q\left(Z|\zeta\right)}_{\text{Non-degenerate, in some parametric family:}}$$

$$q\left(Z|\zeta\right) \in \mathcal{Q}$$

$$q\left(\theta, Z|\vartheta, \zeta\right) := \delta\left(\theta - \vartheta\right) q\left(Z|\zeta\right)$$



Variational Bayes.

Estimate using a Kullback-Leibler (KL)-like divergence.

$$\hat{\vartheta}, \hat{\zeta} = \underset{\vartheta, \zeta}{\operatorname{argmin}} \widetilde{KL} \left(q \left(\theta, Z | \vartheta, \zeta \right) || p \left(\theta, Z | Y \right) \right)$$

$$\widetilde{KL}\left(q\left(\theta, Z|\vartheta, \zeta\right)||p\left(\theta, Z|Y\right)\right)$$

$$= \mathbb{E}_{q(\theta,Z|\vartheta,\zeta)}\underbrace{\left[\log q\left(Z|\zeta\right)\right]} - \mathbb{E}_{q(\theta,Z|\vartheta,\zeta)}\left[\ell\left(\theta, Z|Y\right)\right]$$

Entropy of $\boldsymbol{\theta}$ is missing -- in this sense it's not a real KL divergence

$$
\begin{aligned}
&\widetilde{KL}\left(q\left(\theta, Z|\vartheta, \zeta\right)||p\left(\theta, Z|Y\right)\right)\\
&\qquad = \mathbb{E}_{q(\theta, Z|\vartheta,\zeta)}\left[\log q\left(Z|\zeta\right)\right] - \mathbb{E}_{q(\theta, Z|\vartheta,\zeta)}\left[\ell\left(\theta, Z|Y\right)\right]\\
&\qquad = \mathbb{E}_{q(Z|\zeta)}\left[\log q\left(Z|\zeta\right) - \ell\left(\vartheta, Z|Y\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&\widetilde{KL}\left(q\left(\theta, Z|\vartheta, \zeta\right)||p\left(\theta, Z|Y\right)\right) \\
&\qquad = \mathbb{E}_{q(\theta, Z|\vartheta, \zeta)}\left[\log q\left(Z|\zeta\right)\right] - \mathbb{E}_{q(\theta, Z|\vartheta, \zeta)}\left[\ell\left(\theta, Z|Y\right)\right] \\
&\qquad = \mathbb{E}_{q(Z|\zeta)}\left[\log q\left(Z|\zeta\right) - \ell\left(\vartheta, Z|Y\right)\right] \\
&\qquad = \mathbb{E}_{q(Z|\zeta)}\left[\log q\left(Z|\zeta\right) - \left(\ell\left(Y|Z, \vartheta\right) + \ell\left(Z|\vartheta\right) - \ell\left(\vartheta\right) + \ell\left(Y\right)\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&\widetilde{KL}\left(q\left(\theta, Z|\vartheta, \zeta\right)||p\left(\theta, Z|Y\right)\right) \\
&\qquad = \mathbb{E}_{q(\theta, Z|\vartheta,\zeta)}\left[\log q\left(Z|\zeta\right)\right] - \mathbb{E}_{q(\theta, Z|\vartheta,\zeta)}\left[\ell\left(\theta, Z|Y\right)\right] \\
&\qquad = \mathbb{E}_{q(Z|\zeta)}\left[\log q\left(Z|\zeta\right) - \ell\left(\vartheta, Z|Y\right)\right] \\
&\qquad = \mathbb{E}_{q(Z|\zeta)}\left[\log q\left(Z|\zeta\right) - \left(\ell\left(Y|Z, \vartheta\right) + \ell\left(Z|\vartheta\right) - \ell\left(\vartheta\right) + \ell\left(Y\right)\right)\right] \\
&\qquad = -\mathbb{E}_{q(Z|\zeta)}\left[\ell\left(Y|Z, \vartheta\right) + \ell\left(Z|\vartheta\right) - \ell\left(\vartheta\right) - \log q\left(Z|\zeta\right)\right] + \ell\left(Y\right)
\end{aligned}
$$

Contrast with the EM identity.

$$
\ell\left(Y|\theta\right) = \mathbb{E}_{p(Z|Y,\theta)}\left[\ell\left(Y|Z, \theta\right) + \ell\left(Z|\theta\right) - \ell\left(Z|Y, \theta\right)\right]
$$

Proposition: VB = EM
(Neal and Hinton, 1998)

Suppose that $p\left(Z|\theta, Y\right) \in \mathcal{Q}, \forall\theta \in \Omega_{\theta}$

Then the VB and EM optima are the same:

$$\hat{\vartheta} = \hat{\theta}$$

$$q\left(Z|\hat{\zeta}\right) = p\left(Z|Y, \hat{\theta}\right)$$

Proposition: VB = EM
(Neal and Hinton, 1998)

We can always find a parametric class Q that satisfies this condition.
(Why?)

$$p\left(Z|\theta, Y\right) \in \mathcal{Q}, \forall \theta \in \Omega_\theta$$

Proposition: VB = EM
(Neal and Hinton, 1998)

From now on *q* will be used for the conditional distribution of Z.

$$q\left(Z|\hat{\zeta}\right) = p\left(Z|Y,\hat{\theta}\right)$$

Proposition: VB = EM

(Neal and Hinton, 1998):

In fact, the EM algorithm is coordinate ascent in ϑ, ζ .

Proposition: VB = EM

(Neal and Hinton, 1998):

Given iteration k, $\vartheta^{(k)}, \zeta^{(k)}$

Step 1k. “E step”:

$$\text{Calculate } \hat{\zeta}^{(k+1)} = \underset{\zeta}{\operatorname{argmin}} \widetilde{KL}\left(q\left(\theta, Z|\hat{\vartheta}^{(k)}, \zeta\right) || p\left(\theta, Z|Y\right)\right)$$

Proposition: VB = EM
(Neal and Hinton, 1998):

Given iteration k, $\vartheta^{(k)}, \zeta^{(k)}$

Step 1k. "E step":

$$\text{Calculate } \hat{\zeta}^{(k+1)} = \underset{\zeta}{\operatorname{argmin}} \widetilde{KL}\left(q\left(\theta, Z|\hat{\vartheta}^{(k)}, \zeta\right) || p\left(\theta, Z|Y\right)\right)$$

Step 2k. "M step":

$$\text{Calculate } \hat{\vartheta}^{(k+1)} = \underset{\vartheta}{\operatorname{argmin}} \widetilde{KL}\left(q\left(\theta, Z|\vartheta, \hat{\zeta}^{(k+1)}\right) || p\left(\theta, Z|Y\right)\right)$$

...repeat.

Who uses coordinate ascent?

$$\hat{\vartheta}, \hat{\zeta} = \underset{\vartheta, \zeta}{\operatorname{argmin}} \widetilde{KL} \left(q \left(\theta, Z | \vartheta, \zeta \right) || p \left(\theta, Z | Y \right) \right)$$

Part 3:
Covariance asymptotics.

Assumption : Bayesian CLT.
Bernstein-von Mises (BVM) theorem

$$\mathcal{I}_{\theta\theta} = -\lim_{n\to\infty} \frac{1}{N} \left.\frac{\partial^2 \ell\left(Y|\theta\right)}{\partial\theta\partial\theta}\right|_{\hat{\theta}}$$

Covariance from the Laplace approximation

Asymptotic Statistics, van der Vaart, 2007.

Assumption : Bayesian CLT.
Bernstein-von Mises (BVM) theorem

$$\mathcal{I}_{\theta\theta} = - \lim_{n \to \infty} \frac{1}{N} \left. \frac{\partial^2 \ell \left(Y | \theta \right)}{\partial \theta \partial \theta} \right|_{\hat{\theta}}$$

$$p \left(\sqrt{N} \mathcal{I}_{\theta\theta}^{1/2} \left(\theta - \hat{\theta} \right) | Y \right) \xrightarrow[N \to \infty]{} \mathcal{N} \left(0, I_D \right)$$

Asymptotic Statistics, van der Vaart, 2007.

Assumption : Bayesian CLT.
Bernstein-von Mises (BVM) theorem

The posterior on θ goes to a degenerate distribution

$$p\left(\underbrace{\sqrt{N}\mathcal{I}_{\theta\theta}^{1/2}\left(\theta-\hat{\theta}\right)}_{\text{The posterior on }\theta\text{ goes to a degenerate distribution}}|Y\right)\xrightarrow[N\to\infty]{}\mathcal{N}\left(0,I_D\right)$$

Asymptotic Statistics, van der Vaart, 2007.

How good are the VB approximation's covariances?

$$\mathrm{Cov}_{p(\theta|Y)}(\theta) \approx \frac{1}{N}\mathcal{I}_{\theta\theta}^{-1} = o_p\left(\sqrt{N}\right)$$

Bayesian CLT

How good are the VB approximation's covariances?

$$\mathrm{Cov}_{p(\theta|Y)}\left(\theta\right) \approx \frac{1}{N}\mathcal{I}_{\theta\theta}^{-1} = o_p\left(\sqrt{N}\right)$$

Bayesian CLT

$$\mathrm{Cov}_{q(\theta|\hat{\vartheta})}\left(\theta\right) \equiv 0$$

Degenerate approximation

How good are the VB approximation's covariances?

$$\mathrm{Cov}_{p(\theta|Y)}\left(\theta\right) \approx \frac{1}{N}\mathcal{I}_{\theta\theta}^{-1} = o_p\left(\sqrt{N}\right)$$

Bayesian CLT

$$\mathrm{Cov}_{q(\theta|\hat{\vartheta})}\left(\theta\right) \equiv 0$$

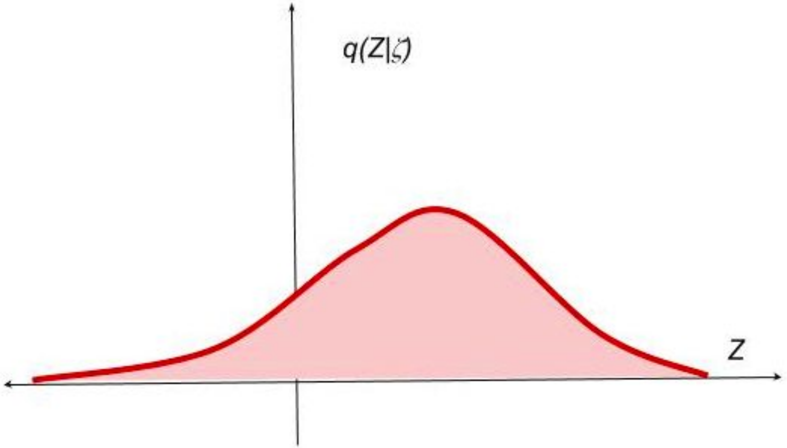
Degenerate approximation

$$\mathrm{Cov}_{p(\theta|Y)}\left(\theta\right) - \mathrm{Cov}_{q(\theta|\hat{\vartheta})}\left(\theta\right) = o_p\left(\sqrt{N}\right)$$

Consistent!
(But trivial.)

What about the variance of *Z*?

$$\mathrm{Cov}_{p(Z_n|Y)}(Z_n) \xrightarrow[N\to\infty]{} \mathrm{Cov}_{p\left(Z_n|Y,\theta=\hat{\theta}\right)}(Z_n) \neq 0$$



What about the variance of *Z*?

$$\mathrm{Cov}_{p(Z_n|Y)}\left(Z_n\right) \xrightarrow[N\to\infty]{} \mathrm{Cov}_{p\left(Z_n|Y,\theta=\hat{\theta}\right)}\left(Z_n\right) \neq 0$$

$$\Vert$$

$$\mathrm{Cov}_{q\left(Z_n|\hat{\zeta}\right)}\left(Z_n\right)$$

What about the variance of *Z*?

$$\mathrm{Cov}_{p(Z_n|Y)}\left(Z_n\right) \xrightarrow[N\to\infty]{} \mathrm{Cov}_{p\left(Z_n|Y,\theta=\hat{\theta}\right)}\left(Z_n\right) \neq 0$$

$$\parallel$$

$$\mathrm{Cov}_{q\left(Z_n|\hat{\zeta}\right)}\left(Z_n\right)$$

Therefore, trivially,

$$\mathrm{Cov}_{p(Z_n|Y)}\left(Z_n\right) - \mathrm{Cov}_{q\left(Z_n|\hat{\zeta}\right)}\left(Z_n\right) = o_p\left(1\right)$$

What about the variance of *Z*?

$$\mathrm{Cov}_{p(Z_n|Y)}\left(Z_n\right) \xrightarrow[N\to\infty]{} \mathrm{Cov}_{p\left(Z_n|Y,\theta=\hat{\theta}\right)}\left(Z_n\right) \neq 0$$

$$\Vert$$

$$\mathrm{Cov}_{q\left(Z_n|\hat{\zeta}\right)}\left(Z_n\right)$$

Only slightly less trivially:

$$\mathrm{Cov}_{p(Z_n|Y)}\left(Z_n\right) - \mathrm{Cov}_{q\left(Z_n|\hat{\zeta}\right)}\left(Z_n\right) = o_p\left(\sqrt{N}\right)$$

Can these naive approximations be improved?

Let’s try linear response covariances.

Linear response covariances reminder.

$$\frac{d\mathbb{E}\left[\theta|\alpha, X\right]}{d\alpha} = \mathrm{Cov}_{p(\theta|\alpha,X)}\left(\theta, \frac{\partial}{\partial\alpha}\log p\left(\theta|\alpha, X\right)\right)$$

$$\frac{d\mathbb{E}\left[\theta|\alpha, X\right]}{d\alpha} = \mathrm{Cov}_{p(\theta|\alpha, X)}\left(\theta, \frac{\partial}{\partial\alpha}\log p\left(\theta|\alpha, X\right)\right)$$

$$\hat{\theta}\left(t\right) = \underset{\theta}{\mathrm{argmin}}\;\left(f\left(\theta\right) + t\theta\right)$$

$$\left.\frac{d\hat{\theta}}{dt}\right|_{t=0} = \left(\left.\frac{\partial^2 f\left(\theta\right)}{\partial\theta\partial\theta}\right|_{\hat{\theta}}\right)^{-1}$$

$$\frac{d\mathbb{E}\left[\theta|\alpha, X\right]}{d\alpha} = \mathrm{Cov}_{p(\theta|\alpha, X)}\left(\theta, \frac{\partial}{\partial\alpha}\log p\left(\theta|\alpha, X\right)\right)$$

Define a perturbation:

$$p\left(\theta, Z|Y, t\right) \propto p\left(\theta, Z|Y\right)\exp\left(t\theta\right)$$

$$\frac{d\mathbb{E}\left[\theta|\alpha, X\right]}{d\alpha} = \mathrm{Cov}_{p(\theta|\alpha,X)}\left(\theta, \frac{\partial}{\partial\alpha}\log p\left(\theta|\alpha, X\right)\right)$$

Define a perturbation:

$$p\left(\theta, Z|Y, t\right) \propto p\left(\theta, Z|Y\right)\exp\left(t\theta\right)$$

$$\left.\frac{d\mathbb{E}_{p(\theta,Z|Y,t)}\left[\theta\right]}{dt}\right|_{t=0} = \mathrm{Cov}_{p(\theta,Z|Y)}\left(\theta\right)$$

$$\frac{d\mathbb{E}\left[\theta|\alpha, X\right]}{d\alpha} = \mathrm{Cov}_{p(\theta|\alpha,X)}\left(\theta, \frac{\partial}{\partial\alpha}\log p\left(\theta|\alpha, X\right)\right)$$

Define a perturbation:

$$p\left(\theta, Z|Y, t\right) \propto p\left(\theta, Z|Y\right)\exp\left(t\theta\right)$$

$$\left.\frac{d\hat{\theta}\left(t\right)}{dt}\right|_{t=0} \approx \left.\frac{d\mathbb{E}_{p(\theta,Z|Y,t)}\left[\theta\right]}{dt}\right|_{t=0} = \mathrm{Cov}_{p(\theta,Z|Y)}\left(\theta\right)$$

Sensitivity requires the Hessian of the optimum.

$$H := \left. \frac{d^2 \widehat{\mathrm{KL}}(\eta)}{d\eta d\eta^\intercal} \right|_{\hat{\eta}} = \begin{pmatrix} \overbrace{H_{\theta\theta}}^{\text{Fixed dimension}} & H_{\theta\zeta} \\ H_{\zeta\theta} & \underbrace{H_{\zeta\zeta}}_{\text{Grows with N}} \end{pmatrix}$$

$$\eta := \begin{pmatrix} \vartheta \\ \zeta \end{pmatrix}$$

Linear response covariances for theta.

$$H := \left. \frac{d^2 \widehat{\mathrm{KL}}\,(\eta)}{d\eta d\eta^\intercal} \right|_{\hat{\eta}} = \left(\begin{array}{cc} H_{\theta\theta} & H_{\theta\zeta} \\ H_{\zeta\theta} & H_{\zeta\zeta} \end{array} \right)$$

Linear response covariances for theta.

$$H := \left. \frac{d^2 \widehat{\mathrm{KL}}\left(\eta\right)}{d\eta d\eta^{\intercal}} \right|_{\hat{\eta}} = \begin{pmatrix} H_{\theta\theta} & H_{\theta\zeta} \\ H_{\zeta\theta} & H_{\zeta\zeta} \end{pmatrix}$$

$$\Sigma = H^{-1} = \begin{pmatrix} \Sigma_{\theta\theta} & \Sigma_{\theta\zeta} \\ \Sigma_{\zeta\theta} & \Sigma_{\zeta\zeta} \end{pmatrix}$$

Linear response covariances for theta.

$$H := \left.\frac{d^2\widehat{\mathrm{KL}}(\eta)}{d\eta d\eta^\intercal}\right|_{\hat{\eta}} = \begin{pmatrix} H_{\theta\theta} & H_{\theta\zeta} \\ H_{\zeta\theta} & H_{\zeta\zeta} \end{pmatrix}$$

$$\Sigma = H^{-1} = \begin{pmatrix} \Sigma_{\theta\theta} & \Sigma_{\theta\zeta} \\ \Sigma_{\zeta\theta} & \Sigma_{\zeta\zeta} \end{pmatrix}$$

$$\widehat{\mathrm{Cov}}_{LR}(\theta) = \Sigma_{\theta\theta}$$

Proposition 1:

With a flat prior, linear response posterior covariances for θ and classical frequentist covariance estimates are the same.

$$\widehat{\mathrm{Cov}}_{LR}(\theta) = \Sigma_{\theta\theta} = -\left(\left. \frac{d^2 \ell \left(Y | \theta \right)}{d\theta d\theta^\intercal} \right|_{\hat{\theta}} \right)^{-1}$$

Proposition 1 proof.

$$\widehat{\mathrm{KL}}\left(\vartheta, \zeta\right) = -\mathbb{E}_{q(Z|\zeta)}\left[\ell\left(Y|Z, \vartheta\right) + \ell\left(Z|\vartheta\right) - \log q\left(Z|\zeta\right)\right] + \ell\left(Y\right)$$

Definition, flat prior.

Proposition 1 proof.

$$\widehat{\mathrm{KL}}\left(\vartheta,\zeta\right)=-\mathbb{E}_{q(Z|\zeta)}\left[\ell\left(Y|Z,\vartheta\right)+\ell\left(Z|\vartheta\right)-\log q\left(Z|\zeta\right)\right]+\ell\left(Y\right)$$

Define

$$\hat{\zeta}\left(\vartheta\right)=\underset{\zeta}{\operatorname{argmin}}\,\widehat{\mathrm{KL}}\left(\vartheta,\zeta\right)$$

$$q\left(Z|\hat{\zeta}\left(\theta\right)\right)=p\left(Z|Y,\theta\right)$$

Completeness of VB approximation.

Proposition 1 proof.

$$\widehat{\mathrm{KL}}\left(\vartheta, \hat{\zeta}\left(\vartheta\right)\right) = -\mathbb{E}_{q\left(Z|\hat{\zeta}(\vartheta)\right)}\left[\ell\left(Y|Z,\vartheta\right) + \ell\left(Z|\vartheta\right) - \log q\left(Z|\hat{\zeta}\left(\vartheta\right)\right)\right] + \ell\left(Y\right)$$

Definition.

Proposition 1 proof.

$$\widehat{\mathrm{KL}}\left(\vartheta, \hat{\zeta}(\vartheta)\right) = -\mathbb{E}_{q\left(Z|\hat{\zeta}(\vartheta)\right)}\left[\ell(Y|Z,\vartheta) + \ell(Z|\vartheta) - \log q\left(Z|\hat{\zeta}(\vartheta)\right)\right] + \ell(Y)$$

$$= -\mathbb{E}_{p(Z|Y,\vartheta)}\left[\ell(Y|Z,\vartheta) + \ell(Z|\vartheta) - \ell(Z|Y,\vartheta)\right] + \ell(Y)$$

Completeness of VB approximation.

Proposition 1 proof.

$$\begin{aligned}\widehat{\mathrm{KL}}\left(\vartheta, \hat{\zeta}(\vartheta)\right) &= -\mathbb{E}_{q\left(Z|\hat{\zeta}(\vartheta)\right)}\left[\ell(Y|Z,\vartheta) + \ell(Z|\vartheta) - \log q\left(Z|\hat{\zeta}(\vartheta)\right)\right] + \ell(Y) \\ &= -\mathbb{E}_{p(Z|Y,\vartheta)}\left[\ell(Y|Z,\vartheta) + \ell(Z|\vartheta) - \ell(Z|Y,\vartheta)\right] + \ell(Y) \\ &= -\ell(Y|\vartheta) + \ell(Y).\end{aligned}$$

The EM identity.

Proposition 1 proof.

All these perturbed optimization problems are the same:

$$\hat{\vartheta}(t), \hat{\zeta}(t) = \underset{\vartheta,\zeta}{\operatorname{argmax}} \; -\widehat{\mathrm{KL}}(\vartheta,\zeta) + t\vartheta$$

$$\hat{\hat{\vartheta}}(t) = \underset{\vartheta}{\operatorname{argmax}} \; -\widehat{\mathrm{KL}}\left(\vartheta, \hat{\zeta}(\vartheta)\right) + t\vartheta$$

$$\hat{\theta}(t) = \underset{\theta}{\operatorname{argmax}} \, \ell(Y|\theta) + t\theta$$

$$\hat{\vartheta}(t) = \hat{\hat{\vartheta}}(t) = \hat{\theta}(t)$$

Recall our Q-function

$$\hat{Q}(\theta) = \mathbb{E}_{p\left(Z|Y,\hat{\theta}\right)}\left[\ell\left(Y|Z,\theta\right) + \ell\left(Z|\theta\right) - \ell\left(Z|Y,\hat{\theta}\right)\right]$$

Recall our Q-function

$$\hat{Q}\left(\theta\right) = \mathbb{E}_{p\left(Z|Y,\hat{\theta}\right)}\left[\ell\left(Y|Z,\theta\right) + \ell\left(Z|\theta\right) - \ell\left(Z|Y,\hat{\theta}\right)\right]$$

This is the Hessian of the Q-function

$$H = \begin{pmatrix} \overbrace{H_{\theta\theta}} & H_{\theta\zeta} \\ H_{\zeta\theta} & H_{\zeta\zeta} \end{pmatrix}$$

Recall our Q-function

$$\hat{Q}(\theta) = \mathbb{E}_{p(Z|Y,\hat{\theta})}\left[\ell(Y|Z,\theta) + \ell(Z|\theta) - \ell\left(Z|Y,\hat{\theta}\right)\right]$$

$$H = \begin{pmatrix} \overbrace{H_{\theta\theta}}^{\text{This is the Hessian of the Q-function}} & H_{\theta\zeta} \\ H_{\zeta\theta} & H_{\zeta\zeta} \end{pmatrix} \qquad \Sigma = H^{-1} = \begin{pmatrix} \overbrace{\Sigma_{\theta\theta}}^{\text{This is the linear response covariance}} & \Sigma_{\theta\zeta} \\ \Sigma_{\zeta\theta} & \Sigma_{\zeta\zeta} \end{pmatrix}$$

$$\left.\frac{\partial^2 \ell\left(Y|\theta\right)}{\partial\theta\partial\theta}\right|_{\hat{\theta}} \neq \left.\frac{\partial^2 \hat{Q}\left(\theta\right)}{\partial\theta\partial\theta}\right|_{\hat{\theta}}$$

$$\Sigma_{\theta\theta}^{-1} \neq H_{\theta\theta}^{-1}$$

$$H = \begin{pmatrix} \overbrace{H_{\theta\theta}}^{\text{This is the Hessian of the Q-function}} & H_{\theta\zeta} \\ H_{\zeta\theta} & H_{\zeta\zeta} \end{pmatrix}$$

$$\Sigma = H^{-1} = \begin{pmatrix} \overbrace{\Sigma_{\theta\theta}}^{\text{This is the linear response covariance}} & \Sigma_{\theta\zeta} \\ \Sigma_{\zeta\theta} & \Sigma_{\zeta\zeta} \end{pmatrix}$$

Theorem 1:

The linear response covariances add a root-N order of accuracy to the following covariances.

$$\mathrm{Cov}_{p(\theta|Y)}(\theta) - \widehat{\mathrm{Cov}}_{LR}(\theta) = o_p\left(\frac{1}{N}\right)$$

Theorem 1:

The linear response covariances add a root-N order of accuracy to the following covariances.

$$\mathrm{Cov}_{p(\theta|Y)}(\theta) - \widehat{\mathrm{Cov}}_{LR}(\theta) = o_p\left(\frac{1}{N}\right)$$

$$\mathrm{Cov}_{p(\theta|Y)}(\theta, Z_n) - \widehat{\mathrm{Cov}}_{LR}(\theta, Z_n) = o_p\left(\frac{1}{N}\right)$$

Theorem 1:

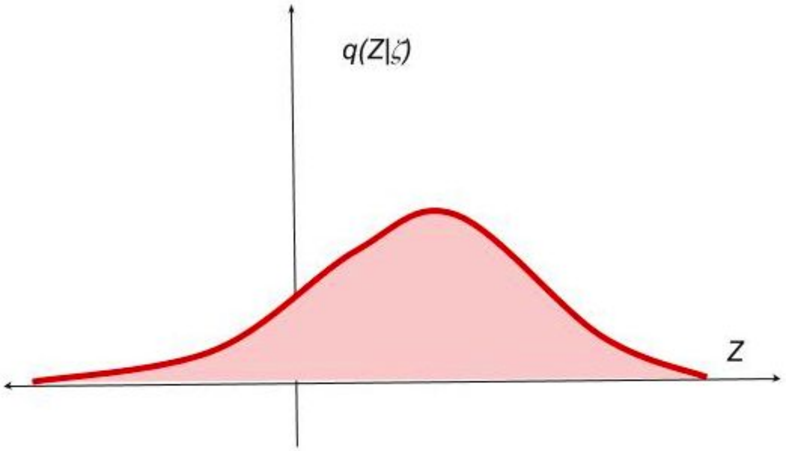
The linear response covariances add a root-N order of accuracy to the following covariances.

$$\mathrm{Cov}_{p(\theta|Y)}(\theta) - \widehat{\mathrm{Cov}}_{LR}(\theta) = o_p\left(\frac{1}{N}\right)$$

$$\mathrm{Cov}_{p(\theta|Y)}(\theta, Z_n) - \widehat{\mathrm{Cov}}_{LR}(\theta, Z_n) = o_p\left(\frac{1}{N}\right)$$

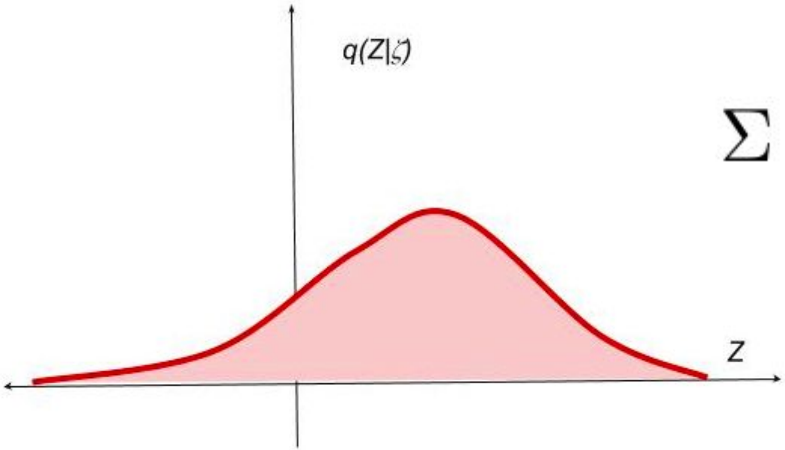
$$\mathrm{Cov}_{p(\theta|Y)}(Z_n, Z_m) - \widehat{\mathrm{Cov}}_{LR}(Z_n, Z_m) = o_p\left(\frac{1}{N}\right) \quad (\text{for } n \neq m)$$

What about Z?



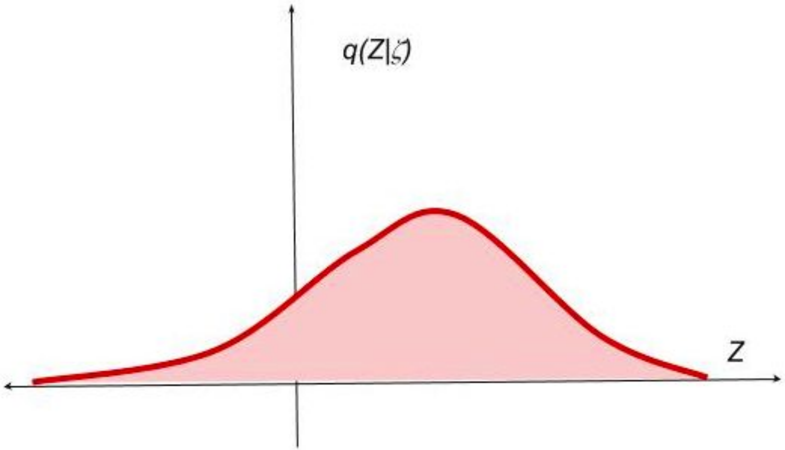
What about Z?

And what about these other Hessian terms?



$$\Sigma = H^{-1} = \begin{pmatrix} \Sigma_{\theta\theta} & \Sigma_{\theta\zeta} \\ \Sigma_{\zeta\theta} & \Sigma_{\zeta\zeta} \end{pmatrix}$$

What about Z?

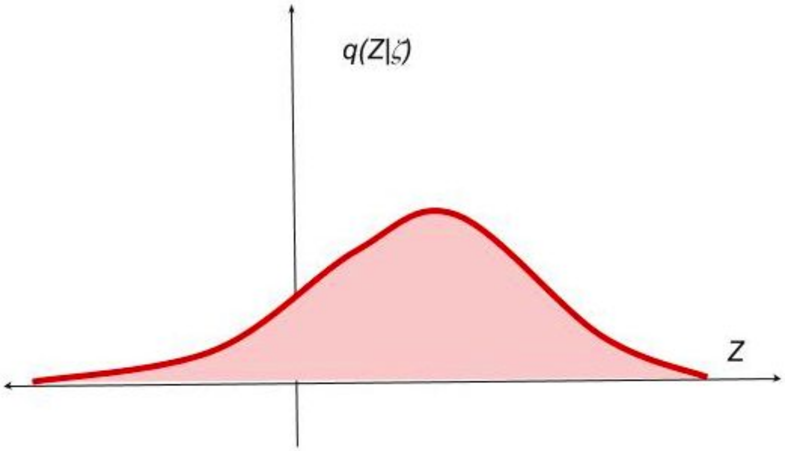


Results are best expressed in terms of the score function and its variance.

$$\hat{\gamma}\left(Z_n\right) := \left.\frac{\partial \log q\left(Z_n|\zeta\right)}{\partial \zeta}\right|_{\hat{\zeta}}$$

$$V_n := \mathrm{Cov}_{q\left(Z_n|\hat{\zeta}\right)}\left(\hat{\gamma}\left(Z_n\right)\right)$$

What about Z?



If q is in the exponential family, $\gamma(Z)$ are sufficient statistics.

$$\hat{\gamma}\left(Z_n\right) := \left.\frac{\partial \log q\left(Z_n|\zeta\right)}{\partial \zeta}\right|_{\hat{\zeta}}$$

$$V_n := \mathrm{Cov}_{q\left(Z_n|\hat{\zeta}\right)}\left(\hat{\gamma}\left(Z_n\right)\right)$$

Proposition 2:

The linear response covariances of the normalized score function are given by the inverse Hessian of the KL divergence.

$$\widehat{\mathrm{Cov}}_{LR}\left(V_n^{-1}\hat{\gamma}\left(Z_n\right)\right) = \Sigma_{\zeta_n \zeta_n}$$

$$\hat{\gamma}\left(Z_n\right) := \left.\frac{\partial \log q\left(Z_n \mid \zeta\right)}{\partial \zeta}\right|_{\hat{\zeta}}$$

$$V_n := \mathrm{Cov}_{q\left(Z_n \mid \hat{\zeta}\right)}\left(\hat{\gamma}\left(Z_n\right)\right)$$

Theorem 2:

The linear response covariances improves the constant, but not the rate, of the covariances of the score function.

Theorem 2:

The linear response covariances improves the constant, but not the rate, of the covariances of the score function.

Error when using the degenerate approximation for p(θ):

$$\mathrm{Cov}_{p(\theta,Z|Y)}\left(\hat{\gamma}\left(Z\right)\right)-\mathrm{Cov}_{q\left(Z|\hat{\zeta}\right)}\left(\hat{\gamma}\left(Z\right)\right)=\left(\mathbb{E}_{q\left(Z_{n}|\hat{\zeta}\right)}\left[\hat{\gamma}\left(Z\right)^{3}\right]+1\right)o_{p}\left(\frac{1}{\sqrt{N}}\right)+\mathrm{Cov}_{p(\theta|Y)}\left(\mathbb{E}_{p(Z|Y,\theta)}\left[\hat{\gamma}\left(Z\right)\right]\right)$$

Theorem 2:

The linear response covariances improves the constant, but not the rate, of the covariances of the score function.

$$\mathrm{Cov}_{p(\theta,Z|Y)}\left(\hat{\gamma}\left(Z\right)\right) - \mathrm{Cov}_{q\left(Z|\hat{\zeta}\right)}\left(\hat{\gamma}\left(Z\right)\right) = \left(\overbrace{\mathbb{E}_{q\left(Z_n|\hat{\zeta}\right)}\left[\hat{\gamma}\left(Z\right)^3\right]}^{\textbf{Skewness (small when q is approximately symmetric).}} + 1\right) o_p\left(\frac{1}{\sqrt{N}}\right) + \underbrace{\mathrm{Cov}_{p(\theta|Y)}\left(\mathbb{E}_{p(Z|Y,\theta)}\left[\hat{\gamma}\left(Z\right)\right]\right)}_{o_p\left(\frac{1}{\sqrt{N}}\right)}$$

Theorem 2:

The linear response covariances improves the constant, but not the rate, of the covariances of the score function.

$$\mathrm{Cov}_{p(\theta,Z|Y)}\left(\hat{\gamma}\left(Z\right)\right) - \mathrm{Cov}_{q\left(Z|\hat{\zeta}\right)}\left(\hat{\gamma}\left(Z\right)\right) = \left(\mathbb{E}_{q\left(Z_n|\hat{\zeta}\right)}\left[\hat{\gamma}\left(Z\right)^3\right] + 1\right) o_p\left(\frac{1}{\sqrt{N}}\right) + \mathrm{Cov}_{p(\theta|Y)}\left(\mathbb{E}_{p(Z|Y,\theta)}\left[\hat{\gamma}\left(Z\right)\right]\right)$$

Error when using linear response covariances:

$$\mathrm{Cov}_{p(\theta,Z|Y)}\left(\hat{\gamma}\left(Z\right)\right) - \widehat{\mathrm{Cov}}_{LR}\left(\hat{\gamma}\left(Z\right)\right) = \left(\mathbb{E}_{q\left(Z_n|\hat{\zeta}\right)}\left[\hat{\gamma}\left(Z\right)^3\right] + 1\right) o_p\left(\frac{1}{\sqrt{N}}\right)$$

Theorem 3:

For covariances of functions of Z other than linear combinations of the score functions, linear response is inconsistent.

$$\mathrm{Cov}_{p(\theta,Z|Y)}\left(h\left(Z_n\right)\right) - \widehat{\mathrm{Cov}}_{LR}\left(h\left(Z_n\right)\right) = O_p\left(1\right)$$

Theorem 3:

For covariances of functions of Z other than linear combinations of the score functions, linear response is inconsistent.

$$\mathrm{Cov}_{p(\theta,Z|Y)}\left(h\left(Z_n\right)\right)-\widehat{\mathrm{Cov}}_{LR}\left(h\left(Z_n\right)\right)=O_p\left(1\right)$$

Practical workarounds:

- Increase the expressivity of *q*
- Use Monte Carlo instead of linear response

Tools.

“However, analytical evaluation of the second-order derivatives of the incomplete-data log likelihood may be difficult or at least tedious. Indeed, often it is for reasons of this nature that the EM algorithm is used to compute the MLE in the first place.”

- The EM Algorithm and Extensions, McLachlan (2008)

“However, analytical evaluation of the second-order derivatives of the incomplete-data log likelihood may be difficult or at least tedious. Indeed, often it is for reasons of this nature that the EM algorithm is used to compute the MLE in the first place.”

- The EM Algorithm and Extensions, McLachlan (2008)

Automatic differentiation makes everything easy.

Automatic e-steps (for conjugate Z distributions):

```
def log_likelihood(z, log_z, theta, y):
    # ... return scalar log posterior value.

# Implemented with autodiff. Takes as an argument a function whose first
# two arguments are z and log_z and which evaluates the log probability..
get_e_step = get_gamma_e_step_funs(log_posterior)

y = load_data()
done = False
theta_k = # ... some init value
while not done:
    # This is automatic.
    e_z, log_z = get_e_step(theta_k, y)

    # Your favorite M-step routine here.
    theta_k = optimize(lambda theta: log_posterior(e_z, e_log_z, theta, y))

    # Check convergence and set done
```

But why bother with an e-step?

```python
def log_likelihood(z, log_z, theta, y):
    # .... return scalar log posterior value.

# Implemented with autodiff. Takes as an argument a function whose first
# two arguments are z and log_z and which evaluates the log probability.
get_marginal_log_lik = get_gamma_marginal_log_lik(log_posterior)

y = load_data()
theta_hat = optimize(lambda theta: get_marginal_log_lik(theta, y))
```

This is what I actually do in practice.

And covariances are now easy.

```
def log_likelihood(z, log_z, theta, y):
    # .... return scalar log posterior value.

# Implemented with autodiff. Takes as an argument a function whose first
# two arguments are z and log_z and which evaluates the log probability.
get_marginal_log_lik = get_gamma_marginal_log_lik(log_posterior)

y = load_data()
theta_hat = optimize(lambda theta: get_marginal_log_lik(theta, y))

cov_theta_k = -1 * np.linalg.inv(
    autograd.hessian(get_marginal_log_lik)(theta_hat, y))
```

(...and quite a lot of academic literature is obsolete.)

Paragami: “Parameter origami”

https://github.com/rgiordan/paragami

Converts parameter dictionaries and the functions that consume or return them between “folded” and “flat” representations.

All transformations are differentiable by autograd.

Paragami: "Parameter origami" https://github.com/rgiordan/paragami

Define "patterns" that describe your structured parameter sets.

[3]
```
mvn_pattern = paragami.PatternDict(free_default=True)
mvn_pattern['mean'] = paragami.NumericVectorPattern(length=dim)
mvn_pattern['cov'] = paragami.PSDSymmetricMatrixPattern(size=dim)
```

Paragami: "Parameter origami" https://github.com/rgiordan/paragami

[36

```
true_mvn_par = dict()
true_mvn_par['mean'] = mean_true
true_mvn_par['cov'] = cov_true

print('\nA dictionary of MVN parameters:\n{}'.format(
      true_mvn_par))

mvn_par_free = mvn_pattern.flatten(true_mvn_par)
print('\nA flat representation:\n{}'.format(
    mvn_pattern.flatten(true_mvn_par)))

print('\nFolding recovers the original parameters:\n{}'.format(
    mvn_pattern.fold(mvn_par_free)))
```

```
A dictionary of MVN parameters:
{'cov': array([[1.1, 1. ],
       [1. , 1.1]]), 'mean': array([0.87367236, 0.21280422])}

A flat representation:
[ 0.87367236  0.21280422  0.04765509  0.95346259 -0.82797896]

Folding recovers the original parameters:
OrderedDict([('mean', array([0.87367236, 0.21280422])), ('cov', array([[1.1, 1. ],
       [1. , 1.1]]))])
```

Vittles:
“Variational inference tools to leverage estimator sensitivity”

https://github.com/rgiordan/vittles

Calculates Taylor series approximations (to arbitrary order) of the dependence of optima on hyperparameters.

Linear response covariances are a special case.

Vittles:
“Variational inference tools to leverage estimator sensitivity”

https://github.com/rgiordan/vittles

```
def get_flat_kl_divergence(parameters):
    # Return the KL divergence

def get_posterior_moments_from_parameters(parameters):
    # Return a vector of posterior moments

optimal_parameters = optimize(get_flat_kl_divergence)

get_lrvb_cov = vittles.LinearResponseCovariances(
    get_flat_kl_divergence,
    optimal_parameters)

lrvb_cov = get_lrvb_cov.get_lr_covariance(
    get_posterior_moments_from_parameters)
```

Thank you for your attention!

Extra topics (in case anyone asks)

The Laplace approximation and linear response covariances.

Example: the Laplace approximation.

$$\hat{\theta} = \underset{\theta \in \Omega_\theta}{\operatorname{argmax}} \; \log p\left(\theta | x\right)$$

Example: the Laplace approximation.

$$\hat{\theta} = \underset{\theta \in \Omega_\theta}{\operatorname{argmax}} \; \log p\left(\theta | x\right)$$

$$\mathbb{E}_{p(\theta | X)}\left[\theta\right] \approx \hat{\theta}$$

$$\hat{\theta}(\alpha) = \underset{\theta \in \Omega_\theta}{\operatorname{argmax}} \ \log p(\theta|x) + \overbrace{\alpha\theta}^{\text{Cleverly chosen perturbation}}$$

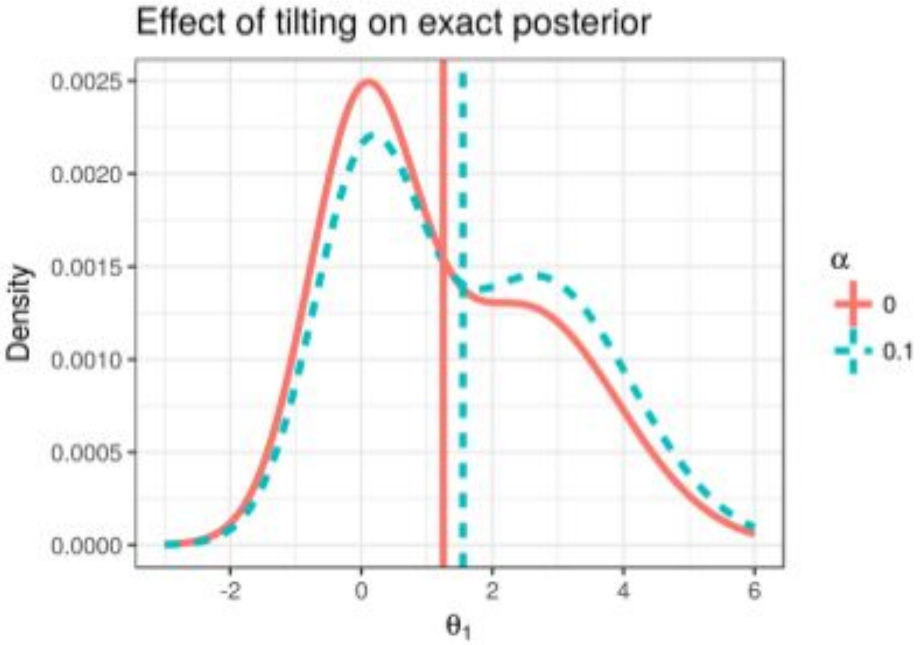
$$\hat{\theta}(\alpha) = \underset{\theta \in \Omega_\theta}{\operatorname{argmax}} \log p(\theta|x) + \overbrace{\alpha\theta}^{\text{Cleverly chosen perturbation}}$$

$$\frac{d\mathbb{E}[\theta|\alpha, X]}{d\alpha} = \operatorname{Cov}_{p(\theta|\alpha,X)}\left(\theta, \overbrace{\frac{\partial}{\partial\alpha}\log p(\theta|\alpha, X)}^{\text{Chosen so that this term becomes }\theta}\right)$$

$$\hat{\theta}(\alpha) = \underset{\theta \in \Omega_\theta}{\operatorname{argmax}} \; \log p(\theta | x) + \overbrace{\alpha\theta}^{\textbf{Cleverly chosen perturbation}}$$

$$\mathbb{E}_{p(\theta | X, \alpha)}[\theta] \approx \hat{\theta}(\alpha)$$

$$\hat{\theta}(\alpha) = \underset{\theta \in \Omega_\theta}{\operatorname{argmax}} \log p(\theta | x) + \alpha\theta$$

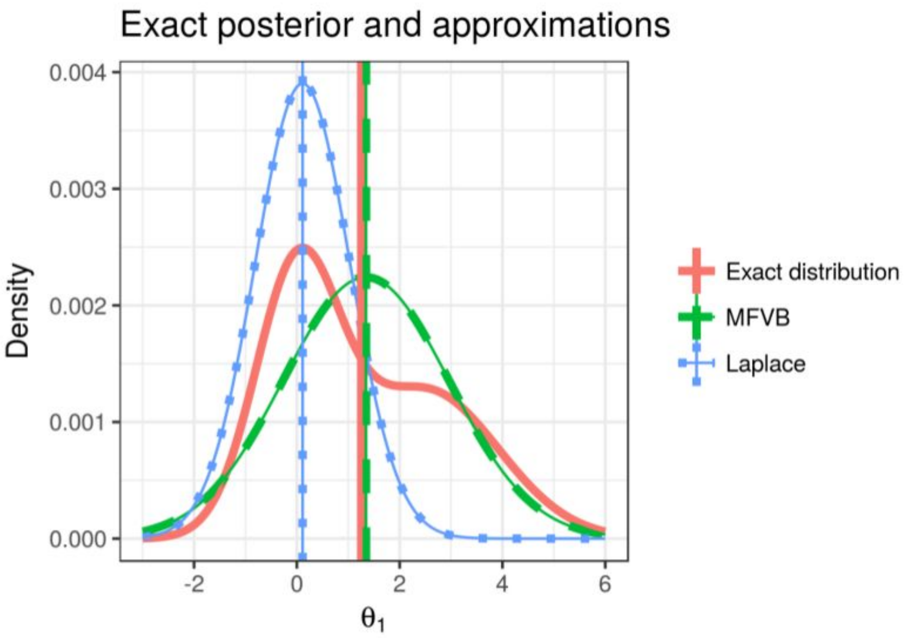


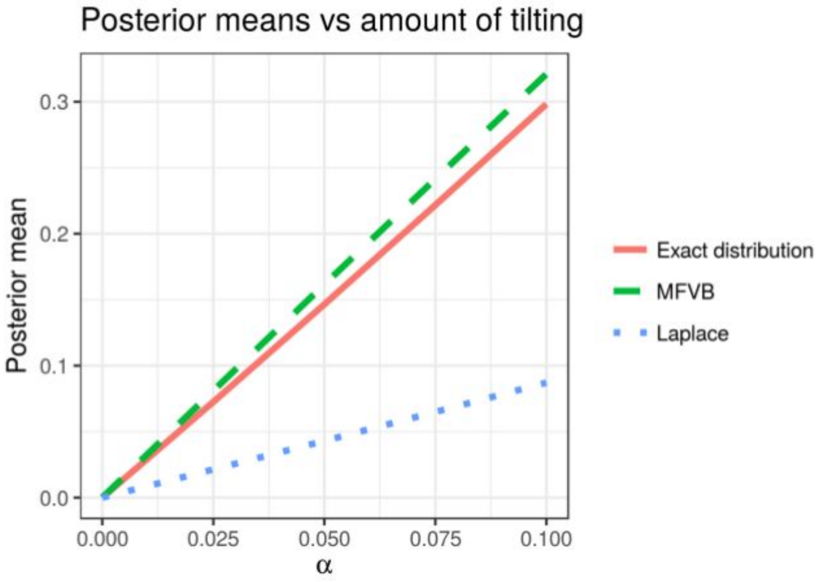
$$\mathrm{Cov}_{p(\theta|X)}\left(\theta\right) = \left. \frac{d\mathbb{E}_{p(\theta|X,\alpha)}\left[\theta\right]}{d\alpha} \right|_{\alpha=0}$$

$$\begin{aligned}
\mathrm{Cov}_{p(\theta|X)}\left(\theta\right) &= \left.\frac{d\mathbb{E}_{p(\theta|X,\alpha)}\left[\theta\right]}{d\alpha}\right|_{\alpha=0} \\
&\approx \left.\frac{d\hat{\theta}\left(\alpha\right)}{d\alpha}\right|_{\alpha=0}
\end{aligned}$$

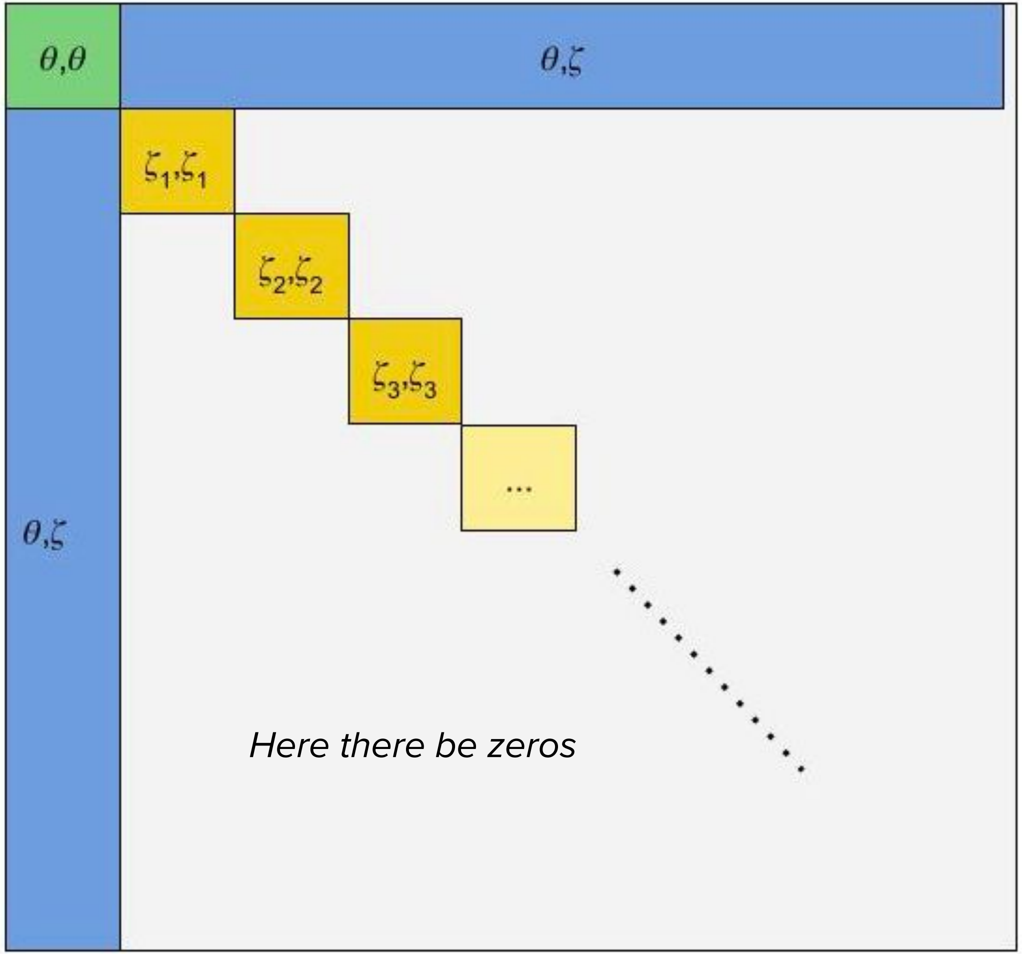
$$
\begin{aligned}
\operatorname{Cov}_{p(\theta|X)}(\theta) &= \left.\frac{d\mathbb{E}_{p(\theta|X,\alpha)}[\theta]}{d\alpha}\right|_{\alpha=0} \\
&\approx \left.\frac{d\hat{\theta}(\alpha)}{d\alpha}\right|_{\alpha=0} \\
&= -\left(\left.\frac{\partial^2 \log p(\theta|x)}{\partial\theta\partial\theta}\right|_{\hat{\theta}}\right)^{-1}
\end{aligned}
$$

In principle, linear response covariances can be calculated for *any optimization-based posterior approximation*.





Hessian picture



$$\begin{pmatrix} \overbrace{H_{\theta\theta}}^{\text{Fixed dimension}} & H_{\theta\zeta} \\ H_{\zeta\theta} & \underbrace{H_{\zeta\zeta}}_{\text{Grows with N}} \end{pmatrix}$$