Part I
Standard Empirical Bayes

1 Model and problem statement

Given a parameter θ, a meta-parameter α, and data x, we want the posterior:

$$p(\theta, \alpha | x) = \frac{p(x|\theta, \alpha) p(\theta) p(\alpha)}{\int p(x|\theta, \alpha) p(\theta) p(\alpha) d\theta d\alpha}$$

We will do empirical Bayes on α and MCMC on θ. Specifically, given

$$\begin{aligned}
\hat{\alpha} &:= \underset{\alpha}{\operatorname{argmax}} \{M(\alpha)\} \\
M(\alpha) &:= \log \int p(x|\theta, \alpha) p(\theta) p(\alpha) d\theta \\
\theta_n &\sim p(\theta | x, \hat{\alpha})
\end{aligned}$$

A common question is how uncertainty in α would translate into uncertainty in θ if it were not fixed at $\hat{\alpha}$. We can try to answer this question with linear response. Recall that if we define the tilted posterior $p(\theta, \alpha | x, t)$ as

$$p(\theta, \alpha | x, t) \propto p(\theta, \alpha | x) \exp\left(\theta^T t\right)$$

Then

$$\left.\frac{d}{dt^T} \mathbb{E}[\theta | x, t]\right|_{t=0} = \operatorname{Cov}(\theta)$$

Under the empirical Bayes assumption that in a neighborhood of $t = 0$,

$$\mathbb{E}[\theta | x, t] \approx \mathbb{E}[\theta | x, \hat{\alpha}(t), t]$$

it may be reasonable to also assume that

$$\operatorname{Cov}(\theta) = \frac{d}{dt^T} \mathbb{E}[\theta | x, t] \approx \frac{d}{dt^T} \mathbb{E}[\theta | x, \hat{\alpha}(t), t] =: \hat{\Sigma}_\theta$$

Note that we have written $\hat{\alpha}(t)$, since the optimal choice of α also depends on t. The right side can be evaluated explicitly without ever sampling from the full model, and this constitutes the linear response estimate of $\operatorname{Cov}(\theta)$.

2 Evaluating $\frac{d}{dt} \mathbb{E}[\theta | x, \hat{\alpha}(t), t]$

Let us write the tilted posterior explicitly:

$$
\begin{aligned}
p\left(x,\theta,\alpha,t\right) &= \frac{p\left(x|\theta\right)p\left(\theta|\alpha\right)p\left(\alpha\right)e^{\theta t}}{\int p\left(x|\theta\right)p\left(\theta|\alpha\right)p\left(\alpha\right)e^{\theta t}d\theta d\alpha} \\
&= \frac{p\left(x|\theta\right)\frac{p(\theta|\alpha)e^{\theta t}}{\int p(\theta'|\alpha)e^{\theta' t}d\theta'}\int p\left(\theta'|\alpha\right)e^{\theta' t}d\theta' p\left(\alpha\right)}{\int p\left(x|\theta\right)p\frac{p(\theta|\alpha)e^{\theta t}}{\int p(\theta'|\alpha)e^{\theta' t}d\theta'}\int p\left(\theta'|\alpha\right)e^{\theta' t}d\theta' p\left(\alpha\right)d\theta d\alpha} \\
&= \frac{p\left(x|\theta\right)p\left(\theta|\alpha,t\right)\int p\left(\theta'|\alpha\right)e^{\theta' t}d\theta' p\left(\alpha\right)}{\int p\left(x|\theta\right)p\left(\theta|\alpha,t\right)\int p\left(\theta'|\alpha\right)e^{\theta' t}d\theta' p\left(\alpha\right)d\theta d\alpha} \\
&= p\left(x|\theta\right)p\left(\theta|\alpha,t\right)p\left(\alpha|t\right)\aleph\left(t\right)
\end{aligned}
$$

Taking

$$
\begin{aligned}
\aleph\left(\alpha,t\right) &:= \int p\left(\theta'|\alpha\right)e^{\theta' t}d\theta' \\
p\left(\theta|\alpha,t\right) &= \frac{p\left(\theta|\alpha\right)e^{\theta t}}{\aleph\left(\alpha,t\right)} \\
p\left(\alpha|t\right) &:= \frac{\aleph\left(\alpha,t\right)p\left(\alpha\right)}{\int \aleph\left(\alpha',t\right)p\left(\alpha'\right)d\alpha'}
\end{aligned}
$$

Given this the tilted posterior is given by

$$
\begin{aligned}
p\left(\theta,\alpha|x,t\right) &= \frac{p\left(x|\theta\right)p\left(\theta|\alpha,t\right)p\left(\alpha|t\right)}{\int p\left(x|\theta\right)p\left(\theta|\alpha,t\right)p\left(\alpha|t\right)d\theta d\alpha}\cdot\frac{\int \aleph\left(\alpha',t\right)p\left(\alpha'\right)d\alpha'}{\int \aleph\left(\alpha',t\right)p\left(\alpha'\right)d\alpha'} \\
&= \frac{p\left(x|\theta\right)p\left(\theta|\alpha,t\right)p\left(\alpha|t\right)}{\int p\left(x|\theta\right)p\left(\theta|\alpha,t\right)p\left(\alpha|t\right)d\theta d\alpha}
\end{aligned}
$$

Then

$$
\Sigma_\theta := \frac{d}{dt^T}\,\mathbb{E}_p\left[\theta|x,t\right]\Big|_{t=0} \quad = \quad Cov_p\left(\theta\right)
$$

Under the assumption that the Empirical Bayes approximation is good, then

$$
\begin{aligned}
\mathbb{E}_p\left[\theta|x,t\right] &\approx \mathbb{E}_p\left[\theta|x,\alpha,t\right] \Rightarrow \\
\hat{\Sigma}_\theta &:= \frac{d}{dt^T}\mathbb{E}_p\left[\theta|x,\alpha,t\right] \approx \frac{d}{dt^T}\mathbb{E}_p\left[\theta|x,t\right]
\end{aligned}
$$

Note that α depends on t since, as a function of t, we have

$$
\begin{aligned}
\hat{\alpha}_t &:= \underset{\alpha}{\operatorname{argmax}}\left\{\log p\left(\alpha|x,t\right)\right\} \\
&= \underset{\alpha}{\operatorname{argmax}}\left\{\log\int p\left(x|\theta,\alpha\right)p\left(\theta|t\right)p\left(\alpha|t\right)d\theta - \log\int\int p\left(x|\theta,\alpha'\right)p\left(\theta|t\right)p\left(\alpha'|t\right)d\theta d\alpha'\right\} \\
&= \underset{\alpha}{\operatorname{argmax}}\left\{\log\int p\left(x|\theta,\alpha\right)p\left(\theta|t\right)p\left(\alpha|t\right)d\theta\right\} \\
\hat{\alpha} &= \hat{\alpha}_0
\end{aligned}
$$

Note at this point that $\hat{\alpha}_t$ is a function of t, and $\mathbb{E}_p[\theta|x, \hat{\alpha}_t, t]$ is a function of both α and t. Taking the total derivative,

$$\begin{aligned}
\hat{\Sigma}_\theta &= \frac{\partial \mathbb{E}_p[\theta|x, \alpha, t]}{\partial \alpha^T} \frac{d\alpha_t}{dt^T}\bigg|_{\alpha=\hat{\alpha}_0, t=0} + \frac{\partial \mathbb{E}_p[\theta|x, \hat{\alpha}_0, t]}{\partial t}\bigg|_{t=0} \\
&= \frac{\partial \mathbb{E}_p[\theta|x, \alpha]}{\partial \alpha^T} \frac{d\alpha_t}{dt^T}\bigg|_{\alpha=\hat{\alpha}_0, t=0} + \mathrm{Cov}(\theta|x, \alpha)
\end{aligned}$$

In order to evaluate $\frac{d}{dt^T}\mathbb{E}_p[\theta|x, \alpha_t, t]$ we will need two terms: $\frac{\partial}{\partial \alpha^T}\mathbb{E}_p[\theta|x, \alpha]$ and $\frac{d\alpha_t}{dt}$. Note that $\frac{\partial \mathbb{E}_p[\theta|x,\alpha,t]}{\partial t} = \mathrm{Cov}(\theta|x, \alpha)$, so this has the form of the conditional variance plus a correction term.

2.1 $d\alpha/dt$

First, we will calculate $d\alpha/dt$. Recall that

$$\begin{aligned}
\hat{\alpha}_t &:= \underset{\alpha}{\operatorname{argmax}} \{p(\alpha|x, t)\} \\
&= \underset{\alpha}{\operatorname{argmax}} \left\{ \frac{p(x|\alpha, t)\, p(\alpha|t)}{\int p(x|\alpha', t)\, p(\alpha'|t)\, d\alpha'} \right\} \\
&= \underset{\alpha}{\operatorname{argmax}} \left\{ \frac{\int p(x|\alpha)\, p(\theta|\alpha, t)\, p(\alpha|t)\, d\theta}{\int\int p(x|\alpha')\, p(\theta'|\alpha', t)\, p(\alpha'|t)\, d\alpha' d\theta'} \right\} \\
&= \underset{\alpha}{\operatorname{argmax}} \left\{ \log \int p(x|\theta, \alpha)\, p(\theta|\alpha, t)\, p(\alpha|t)\, d\theta - \log \int\int p(x|\theta, \alpha')\, p(\theta|\alpha', t)\, p(\alpha'|t)\, d\theta d\alpha' \right\} \\
&= \underset{\alpha}{\operatorname{argmax}} \left\{ \log \int p(x|\theta, \alpha)\, p(\theta|\alpha, t)\, p(\alpha|t)\, d\theta \right\} \text{ (the constant doesn't depend on } \alpha)
\end{aligned}$$

Let us define

$$\begin{aligned}
M(\alpha, t) &:= \log \int p(x|\theta, \alpha)\, p(\theta|\alpha, t)\, p(\alpha|t)\, d\theta \\
&= \log \int p(x|\theta, \alpha)\, p(\theta|\alpha, t)\, d\theta + \log p(\alpha|t)
\end{aligned}$$

Expanding in terms of the original distributions,

$$\begin{aligned}
\log \int p(x|\theta, \alpha)\, p(\theta|\alpha, t)\, d\theta + \log p(\alpha|t) &= \log \int p(x|\theta, \alpha) \frac{p(\theta|\alpha)\, e^{\theta t}}{\aleph(\alpha, t)} d\theta + \log \frac{\aleph(\alpha, t)\, p(\alpha)}{\int \aleph(\alpha', t)\, d\alpha'} \\
&= \log \int p(x|\theta, \alpha)\, p(\theta|\alpha)\, e^{\theta t} d\theta + \log p(\alpha) - \log \int \aleph(\alpha', t)\, d\alpha
\end{aligned}$$

Interestingly, note that if you perform the MLE version of EB, then the term $\aleph(\alpha, t)$ does not cancel and the result is not necessarily symmetric. The last term does not depends only on t, and so does not affect the optimization (partials with respect to α are zero). So

$$\begin{aligned}
\left.\frac{\partial^2 M}{\partial\alpha\partial\alpha^T}\right|_{t=0} &= \frac{\partial^2 M}{\partial\alpha\partial\alpha^T}\left(\log\int p\left(x|\theta,\alpha\right)p\left(\theta|\alpha\right)d\theta+\log p\left(\alpha\right)\right)\\
\left.\frac{\partial}{\partial\alpha}\,\frac{\partial M}{\partial t^T}\right|_{t=0} &= \frac{\partial}{\partial\alpha}\left(\frac{\int p\left(x|\theta,\alpha\right)p\left(\theta|\alpha\right)\theta^T d\theta}{\int p\left(x|\theta,\alpha\right)p\left(\theta|\alpha\right)d\theta}\right)\\
&= \frac{\partial}{\partial\alpha}\mathbb{E}\left[\theta|x,\alpha\right]
\end{aligned}$$

For any t, $\hat{\alpha}_t$ is chosen so that

$$\left.\frac{\partial M}{\partial\alpha}\right|_{t,\hat{\alpha}_t} = 0$$

we can differentiate using the chain rule, giving

$$\begin{aligned}
\frac{\partial^2 M}{\partial\alpha\partial\alpha^T}\frac{d\alpha_t}{dt}+\frac{\partial^2 M}{\partial\alpha\partial t} &= 0 \Rightarrow\\
\frac{d\alpha}{dt} &= -\left(\frac{\partial^2 M}{\partial\alpha\partial\alpha^T}\right)^{-1}\frac{\partial^2 M}{\partial\alpha\partial t^T}
\end{aligned}$$

Note that since M is maximized, its Hessian will be negative definite at $\hat{\alpha}_t$.

2.2 $\partial\mathbb{E}_p\left[\theta|x,\alpha\right]/\partial\alpha$

In some cases, this derivative can be computed exactly (both of our examples below have this property). In general, it can be evaluated with MCMC draws from the EB model by differentiating under the integral:

$$\begin{aligned}
\frac{\partial}{\partial\alpha}\mathbb{E}_p\left[\theta|x,\alpha\right] &= \frac{\partial}{\partial\alpha}\int\theta p\left(\theta|x,\alpha,t\right)d\theta\\
&= \int\theta\frac{\partial}{\partial\alpha}\log p\left(\theta|x,\alpha,t\right)p\left(\theta|x,\alpha,t\right)d\theta\\
&= \int\theta\left(\frac{\partial}{\partial\alpha}\log p\left(\theta|x,\alpha,t\right)-\mathbb{E}\left[\frac{\partial}{\partial\alpha}\log p\left(\theta|x,\alpha,t\right)|x,\alpha,t\right]\right)p\left(\theta|x,\alpha,t\right)d\theta\\
&= \int\left(\theta-\mathbb{E}\left[\theta|x,\alpha,t\right]\right)\left(\frac{\partial}{\partial\alpha}\log p\left(\theta|x,\alpha,t\right)-\mathbb{E}\left[\frac{\partial}{\partial\alpha}\log p\left(\theta|x,\alpha,t\right)|x,\alpha,t\right]\right)p\left(\theta|x,\alpha,t\right)d\theta\\
&= \mathrm{Cov}\left(\theta,\frac{\partial}{\partial\alpha}\log p\left(\theta|x,\alpha,t\right)\right)
\end{aligned}$$

where we have used the fact that

$$\mathbb{E}\left[\frac{\partial}{\partial\alpha}\log p\left(\theta|x,\alpha,t\right)|x,\alpha,t\right] = 0$$

This covariance can be estimated with MCMC draws.

2.3 Putting it all together

Putting together the above results and using the fact that $\frac{\partial^2 M}{\partial \alpha^T \partial t} = \frac{\partial \mathbb{E}_p[\theta|x,\alpha]}{\partial \alpha^T}$, we see that

$$\begin{aligned}
\hat{\Sigma}_\theta &= \left.\frac{\partial \mathbb{E}_p\left[\theta|x,\alpha\right]}{\partial \alpha^T}\frac{d\alpha_t}{dt^T}\right|_{\alpha=\hat{\alpha}_0,t=0} + \mathrm{Cov}\left(\theta|x,\alpha\right) \\
&= \left.-\frac{\partial \mathbb{E}_p\left[\theta|x,\alpha\right]}{\partial \alpha^T}\left(\frac{\partial^2 M}{\partial \alpha \partial \alpha^T}\right)^{-1}\left(\frac{\partial \mathbb{E}_p\left[\theta|x,\alpha\right]}{\partial \alpha}\right)^T\right|_{\hat{\alpha}_0} + \mathrm{Cov}\left(\theta|x,\hat{\alpha}_0\right)
\end{aligned}$$

Since M is maximized at $\hat{\alpha}_0$, this has the form of a positive definite correction to the conditional covariance matrix in terms that can be evaluated only at the EB solution. Note that it is exactly the naive form that you would make with a Laplace approximation to $p\left(\alpha|x\right)$ at $\hat{\alpha}_0$ and a linear assumption on $\mathbb{E}_p\left[\theta|x,\alpha\right]$, though the assumptions seem superficially to have quite a different form. It is interesting to ask whether the key assumption

$$\frac{d}{dt^T}\mathbb{E}\left[\theta|x,t\right] \approx \frac{d}{dt^T}\mathbb{E}\left[\theta|x,\hat{\alpha}\left(t\right),t\right]$$

has made an implicit normal approximation somehow.

3 Gamma Poisson Example

Let's take a classic EB problem, a simple Gamma-Poisson model.

$$\begin{aligned}
y_i|\lambda_i &\sim \mathrm{Poisson}\left(\lambda_i\right) \text{ for } i=1,...,N \\
\lambda_i &\sim \mathrm{Gamma}\left(\gamma,\beta\right) \\
\mathbb{E}\left[\lambda_i\right] &= \frac{\gamma}{\beta}
\end{aligned}$$

Additionally, we may put gamma priors on γ and β (at the risk of having too many gammas around):

$$\begin{aligned}
\gamma &\sim \mathrm{Gamma}\left(a_\gamma,b_\gamma\right) \\
\beta &\sim \mathrm{Gamma}\left(a_\beta,b_\beta\right)
\end{aligned}$$

In the above notation, we have

$$\begin{aligned}
x &= y \\
\theta &= \left(\lambda_1,...,\lambda_N\right)^T \\
\alpha &= \left(\gamma,\beta\right)^T
\end{aligned}$$

3.1 Variance of λ

Note that it is a standard result that the EB objective can be marginalized exactly :

$$\begin{aligned}
\log p(y_i|\lambda_i) &= -\lambda_i + y_i \log \lambda_i - \log y_i! \\
\log p(\lambda_i|\gamma,\beta) &= \gamma \log \beta - \log \Gamma(\gamma) - \beta\lambda_i + (\gamma-1)\log \lambda_i \\
\log p(y,\lambda|\gamma,\beta) &= N\gamma \log \beta - N \log \Gamma(\gamma) - \sum_i (\beta+1)\lambda_i + \sum_i (\gamma+y_i-1)\log\lambda_i - \sum_i \log y_i! \\
&= N\gamma \log \beta - N \log \Gamma(\gamma) - \sum_i (\beta+1)\lambda_i + \sum_i (\gamma+y_i-1)\log\lambda_i - \sum_i \log y_i! \\
&= N\gamma \log \beta - N \log \Gamma(\gamma) - \sum_i \log y_i! \\
&\quad + \sum_i \left[-(\beta+1)\lambda_i + (\gamma+y_i-1)\log\lambda_i + (\gamma+y_i)\log(\beta+1) - \log\Gamma(\gamma+y_i)\right] \\
&\quad - \sum_i \left[(\gamma+y_i)\log(\beta+1) - \log\Gamma(\gamma+y_i)\right] \Rightarrow \\
\log p(y|\gamma,\beta) &= \sum_i \gamma \log\beta - N\log\Gamma(\gamma) - \sum_i (\gamma+y_i)\log(\beta+1) + \sum_i \log\Gamma(\gamma+y_i) - \sum_i \log y_i!
\end{aligned}$$

3.1.1 Hessian of M

First, we need to find the Hessian of the marginal posterior at the optimum:

$$M = \log p(y|\gamma,\beta) + \log p(\gamma) + \log p(\beta)$$

From above, we have

$$\begin{aligned}
\log p(y|\gamma,\beta) &= N\gamma\log(\beta) - N\log\Gamma(\gamma) - \sum_i (\gamma+y_i)\log(\beta+1) + \sum_i \log\Gamma(\gamma+y_i) - \sum_i \log y_i! \\
\frac{\partial}{\partial\gamma}\log p(y|\gamma,\beta) &= \sum_i \log(\beta) - N\psi(\gamma) - \sum_i \log(\beta+1) + \sum_i \psi(\gamma+y_i) \\
\frac{\partial}{\partial\beta}\log p(y|\gamma,\beta) &= \sum_i \left(\frac{\gamma}{\beta} - \frac{\gamma+y_i}{1+\beta}\right)
\end{aligned}$$

So

$$\frac{\partial^2 \log p(y|\gamma,\beta)}{\partial\alpha\partial\alpha^T} = N \begin{pmatrix} \frac{1}{N}\sum_i (\psi'(\gamma+y_i) - \psi'(\gamma)) & \frac{1}{\beta} - \frac{1}{1+\beta} \\ \frac{1}{\beta} - \frac{1}{1+\beta} & \frac{1}{N}\sum_i \left(\frac{\gamma+y_i}{(1+\beta)^2} - \frac{\gamma}{\beta^2}\right) \end{pmatrix}$$

Recall that here N refers to observations within a sample, not the number of MCMC draws. The prior's contribution to the Hessian is given by

$$\begin{aligned}
\gamma &\sim \mathrm{Gamma}(a_\gamma, b_\gamma) \\
\log p(\gamma) &= -b_\gamma\gamma + \log(\gamma)(a_\gamma - 1) + C \\
\frac{\partial^2 \log p(\gamma)}{\partial\gamma^2} &= -\frac{a_\gamma - 1}{\gamma^2}
\end{aligned}$$

And similarly

$$\frac{\partial^2 \log p(\beta)}{\partial \beta^2} = -\frac{a_\beta - 1}{\beta^2}$$

Putting this together,

$$\frac{\partial^2 M}{\partial \alpha \partial \alpha^T} = N \begin{pmatrix} \frac{1}{N}\sum_i \left(\psi'(\gamma + y_i) - \psi'(\gamma)\right) & \frac{1}{\beta} - \frac{1}{1+\beta} \\ \frac{1}{\beta} - \frac{1}{1+\beta} & \frac{1}{N}\sum_i \left(\frac{\gamma+y_i}{(1+\beta)^2} - \frac{\gamma}{\beta^2}\right) \end{pmatrix} - \frac{a_\gamma - 1}{\gamma^2} - \frac{a_\beta - 1}{\beta^2}$$

3.2 Exact moments

Instead of using MCMC, in this case we can use exact moments. Specifically,

$$\begin{aligned} \log p(y_i|\lambda_i) &= -\lambda_i + y_i \log \lambda_i - \log y_i! \\ \log p(\lambda_i|\gamma, \beta) &= \gamma \log(\beta) - \log \Gamma(\gamma) - (\beta)\lambda_i + (\gamma - 1)\log \lambda_i \\ \log p(\lambda_i|y_i, \gamma, \beta) &= (y_i + \gamma)\log(\beta + 1) - \log \Gamma(y_i + \gamma) - (\beta + 1)\lambda_i + (y_i + \gamma - 1)\log \lambda_i \end{aligned}$$

$$\begin{aligned} \mathbb{E}_{p_t}[\lambda_i|\alpha, y_i] &= \frac{\gamma + y_i}{\beta + 1} := m_i(\alpha) \\ \left.\frac{\partial m_i}{\partial \gamma}\right|_{t_i=0} &= \frac{1}{1+\beta} \\ \left.\frac{\partial m_i}{\partial \beta}\right|_{t_i=0} &= -\frac{\gamma + y_i}{(1+\beta)^2} \end{aligned}$$

So

$$\mathrm{Var}_t(\lambda_i) = -\begin{pmatrix} \frac{1}{1+\beta} \\ -\frac{\gamma+y_i}{(1+\beta)^2} \end{pmatrix}^T \left(\frac{\partial^2 M}{\partial \alpha \partial \alpha^T}\right)^{-1} \begin{pmatrix} \frac{1}{1+\beta} \\ -\frac{\gamma+y_i}{(1+\beta)^2} \end{pmatrix} + \frac{y_i + \gamma}{(\beta + 1)^2}$$

3.3 Gamma moments

For reference, here are some moments of a gamma distribution:

$$\begin{aligned} x &\sim \mathrm{Gamma}(\alpha, \beta) \\ p(x) &= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \end{aligned}$$

$$\begin{aligned} \mathbb{E}\left[x^k\right] &= \int \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} x^k dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(k+\alpha)}{\beta^{k+\alpha}} \int \frac{\beta^{k+\alpha}}{\Gamma(k+\alpha)} x^{k+\alpha-1} e^{-\beta x} dx \\ &= \frac{\Gamma(k+\alpha)}{\Gamma(\alpha)} \beta^{-k} \end{aligned}$$

The variance of the sample variance is given by

$$
\begin{aligned}
\mathrm{Var}\left(s^2\right) &= \frac{(N-1)^2}{N^3}\mu_4 - \frac{(N-1)(N-3)\mu_2^2}{N^3} \\
\mu_4 &= \mathbb{E}\left[(x-\mu)^4\right] \\
&= \mathbb{E}\left[x^4 - 4x^3\mu + 6x^2\mu^2 - 4x\mu^3 + \mu^4\right] \\
&= m_4 - 4m_3\mu + 6m_2\mu^2 - 4\mu^4 + \mu^4 \\
&= m_4 - 4m_3\mu + 6m_2\mu^2 - 3\mu^4 \\
\mu_2 &= m_2 - \mu^2
\end{aligned}
$$

Then by the delta method

$$
\mathrm{Var}\left(\sqrt{x}\right) \approx \frac{\mathrm{Var}(x)}{4\mathbb{E}[x]}
$$

3.4 Rao-Blackwellization of variance estimates

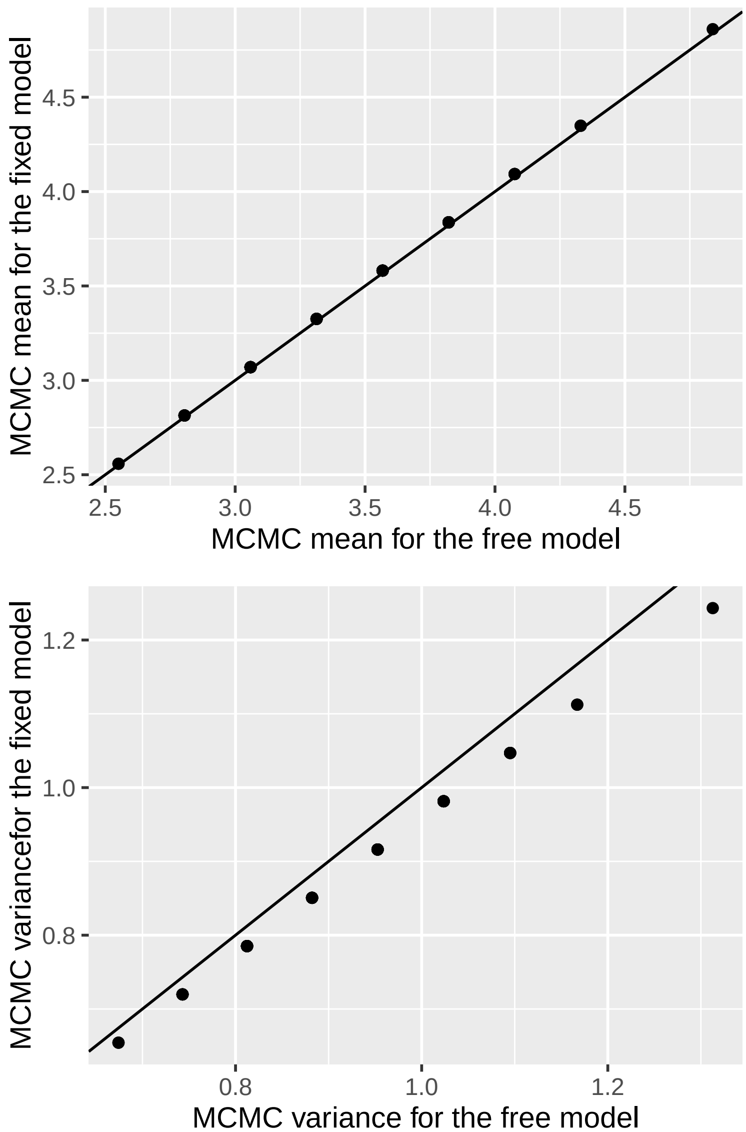
Note that we can use the law of total variance to avoid sampling variance for λ standard deviations. Suppose we know exactly (e.g. because of conjugacy) $\mathbb{E}[\theta|\alpha]$ and $\mathrm{Var}(\theta|\alpha)$. Then under α sampling

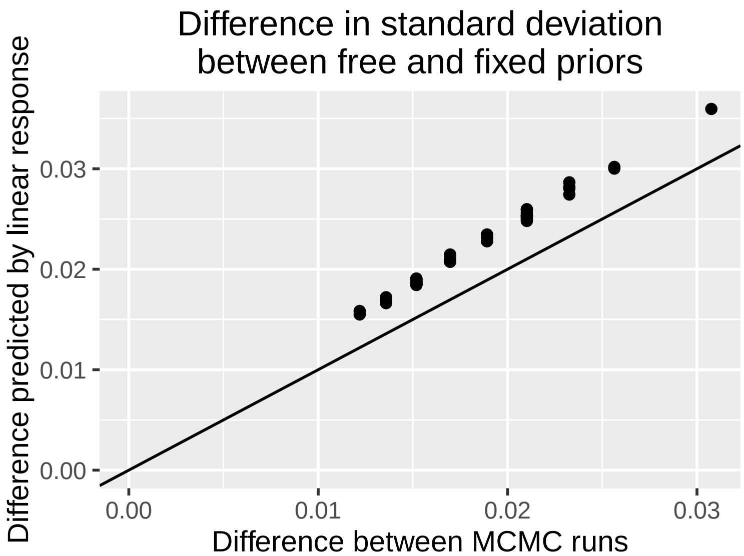
$$
\mathrm{Var}(\theta|X) = \mathbb{E}\left[\mathrm{Var}(\theta|\alpha, X)|X\right] + \mathrm{Var}\left(\mathbb{E}[\theta|\alpha, X]|X\right)
$$

...where the outer expectations can be done by Monte Carlo. This can reduce the number of MCMC samples necessary to accurately estimate variances.

3.5 Results

Here I show some simple results. The "free model" refers to a full MCMC model marginalized over α, where the "fixed model" refers to the model where α is fixed at its EB estimate, $\hat{\alpha}_0$. Here I have chosen a set of parameters where the linear response works well, as shown in the third plot. However, when the posterior on α is less concentrated it does not work very well.


MCMC mean for the fixed model
MCMC mean for the free model
2.5
3.0
3.5
4.0
4.5
MCMC variancefor the fixed model
MCMC variance for the free model
0.8
1.0
1.2



4 Normal-Normal Example

Here, I show that the correction is exact for Normal mixtures of Normals. This is not surprising, but was helpful in tidying up the above results.

Let's take an even more classic example:

$$\begin{aligned} x_i|\theta_i &\sim \mathcal{N}\left(\theta_i, 1\right) \\ \theta_i|\alpha &\sim \mathcal{N}\left(\alpha, \tau^{-1}\right) \\ \alpha &\sim \mathcal{N}\left(0, \tau_\alpha^{-1}\right) \\ \tau, \tau_\alpha & \quad \text{are known} \end{aligned}$$

Then

$$\begin{aligned} \log p\left(x_i|\theta_i\right) &= -\frac{1}{2}\left(x_i - \theta_i\right)^2 + C \\ &= -\frac{1}{2}x_i^2 + x_i\theta_i - \frac{1}{2}\theta_i^2 \\ \log p\left(\theta_i|\alpha\right) &= -\frac{1}{2}\left(\tau\left(\theta_i - \alpha\right)^2 - \log\tau\right) + C \\ &= -\frac{1}{2}\tau\theta_i^2 + \tau\theta_i\alpha - \frac{1}{2}\tau\alpha^2 - \frac{1}{2}\log\tau + C \\ \log p\left(\alpha\right) &= -\frac{1}{2}\tau_\alpha\alpha^2 \end{aligned}$$

4.1 True posterior: integrating out α

Here, we know that marginally over α,

$$\begin{aligned}\theta_i &\sim \mathcal{N}\left(0,\tau_\theta^{-1}\right)\\ \tau_\theta &:= \left(\tau^{-1}+\tau_\alpha^{-1}\right)^{-1}\\ &= \frac{\tau\tau_\alpha}{\tau+\tau_\alpha}\end{aligned}$$

Note that

$$\begin{aligned}\tau_\theta^{-1} &= \tau^{-1}+\tau_\alpha^{-1} > \tau^{-1} \Rightarrow\\ \tau &> \tau_\theta\end{aligned}$$

Thus the posterior is given by

$$\begin{aligned}\mathbb{E}\left[\theta_i|x_i\right] &= \frac{1}{\tau_\theta+1}\left(0\cdot\tau_\theta+x_i\right)\\ &= \frac{x_i}{\tau_\theta+1}\end{aligned}$$

$$\begin{aligned}\theta_i|x_i &\sim \mathcal{N}\left(\frac{1}{\tau_\theta+1}\cdot x_i,\left(\tau_\theta+1\right)^{-1}\right)\\ \frac{1}{\tau_\theta+1} &= \frac{\tau_\alpha+\tau}{\tau_\alpha+\tau+\tau\tau_\alpha}\end{aligned}$$

This is the variance we hope to recover with linear response.

4.2 Integrating out θ_i for Empirical Bayes

Integrating out θ_i,

$$\begin{aligned}x_i|\alpha &\sim \mathcal{N}\left(\alpha,\tau_x^{-1}\right)\\ \tau_x &:= \left(1+\tau^{-1}\right)^{-1}=\frac{\tau}{1+\tau}\\ \alpha &\sim \mathcal{N}\left(0,\tau_\alpha^{-1}\right)\\ \mathbb{E}\left[\alpha|x_i\right] &= \frac{\tau_x}{\tau_\alpha+\tau_x}x_i\end{aligned}$$

Note that this is maximized at $\hat{\alpha}=\frac{\tau_x}{\tau_\alpha+\tau_x}x_i$. Thus the EB posterior of θ_i is given by

$$\begin{aligned}\theta_i|\hat{\alpha} &\sim \mathcal{N}\left(\frac{\tau_x}{\tau_\alpha+\tau_x}x_i,\tau^{-1}\right)\\ x_i|\theta_i &\sim \mathcal{N}\left(\theta_i,1\right)\\ \mathrm{Var}\left(\theta_i|\hat{\alpha},x_i\right) &= \left(1+\tau\right)^{-1}\\ \mathbb{E}\left[\theta_i|\hat{\alpha},x_i\right] &= \frac{1}{1+\tau}\left(\frac{\tau_x\tau}{\tau_\alpha+\tau_x}+1\right)x_i\end{aligned}$$

Dig that

$$\begin{aligned}
\frac{\tau_x \tau}{\tau_\alpha + \tau_x} &= \frac{\tau^2}{\tau_\alpha + \tau_\alpha \tau + \tau} \\
\frac{\tau_x \tau}{\tau_\alpha + \tau_x} + 1 &= \frac{\tau^2 + \tau_\alpha + \tau_\alpha \tau + \tau}{\tau_\alpha + \tau_\alpha \tau + \tau} \\
&= \frac{\tau \left(\tau + \tau_\alpha\right) + \tau_\alpha + \tau}{\tau_\alpha + \tau_\alpha \tau + \tau} \\
&= \frac{\left(1 + \tau\right)\left(\tau + \tau_\alpha\right)}{\tau_\alpha + \tau_\alpha \tau + \tau} \\
&= \frac{1 + \tau}{1 + \tau_\theta}
\end{aligned}$$

So that

$$\mathbb{E}\left[\theta_i | \hat{\alpha}, x_i\right] = \frac{1}{1 + \tau_\theta} x_i$$

Here, in contrast to the correct model with α marginalized out, $\theta_i | \hat{\alpha}, x_i$ has an estimated variance that is too small since $\tau > \tau_m$. However, the shrinkage matches the marginal case exactly. The difference in the variances is given by

$$\begin{aligned}
\frac{1}{1 + \tau_\theta} - \frac{1}{1 + \tau} &= \frac{\tau + \tau_\alpha}{\tau + \tau \tau_\alpha + \tau_\alpha} - \frac{1}{1 + \tau} \\
&= \frac{\left(1 + \tau\right)\left(\tau + \tau_\alpha\right) - \tau - \tau \tau_\alpha - \tau_\alpha}{\left(\tau + \tau \tau_\alpha + \tau_\alpha\right)\left(1 + \tau\right)} \\
&= \frac{\tau + \tau_\alpha + \tau^2 + \tau \tau_\alpha - \tau - \tau \tau_\alpha - \tau_\alpha}{\left(\tau + \tau \tau_\alpha + \tau_\alpha\right)\left(1 + \tau\right)} \\
&= \frac{\tau^2}{\left(\tau + \tau \tau_\alpha + \tau_\alpha\right)\left(1 + \tau\right)} \\
&= \frac{\tau_x^2 \left(1 + \tau\right)}{\left(\tau + \tau \tau_\alpha + \tau_\alpha\right)} \\
&= \frac{\tau_x^2 \left(1 + \tau\right)}{\left(\tau + \left(\tau + 1\right) \tau_\alpha\right)} \\
&= \frac{\tau_x^2}{\frac{\tau}{(1 + \tau)} + \tau_\alpha} \\
&= \frac{\tau_x^2}{\tau_x + \tau_\alpha}
\end{aligned}$$

4.3 Tilted model

Suppose we are interested in the variance of θ_i. Then we perturb with $t_i \theta_i$:

$$
\begin{aligned}
p(x,\theta,\alpha,t) &= \frac{p(x|\theta)\,p(\theta|\alpha)\,p(\alpha)\exp(t\theta)}{\int\int\int p(x|\theta)\,p(\theta|\alpha)\,p(\alpha)\exp(t\theta)\,dxd\theta d\alpha} \\
&= \frac{p(x|\theta)\,p(\theta|\alpha)\,p(\alpha)\exp(t\theta)}{\int\int p(\theta|\alpha)\exp(t\theta)\,p(\alpha)\,d\theta d\alpha} \\
&= \frac{p(x|\theta)\,p(\theta|\alpha)\,p(\alpha)\exp(t\theta)}{\int\int p(\theta|\alpha)\exp(t\theta)\,d\theta p(\alpha)\,d\alpha}
\end{aligned}
$$

$$
p(\theta_i,\alpha|t_i) = \frac{p(\theta_i|\alpha,t_i)\,p(\alpha)\exp(t_i\theta_i)}{\int\int p(\theta_i|\alpha,t_i)\,p(\alpha)\exp(t_i\theta_i)\,d\theta_i d\alpha}
$$

$$
\begin{aligned}
p(\alpha|t_i) &= \frac{\int p(\theta_i|\alpha,t_i)\,p(\alpha)\exp(t_i\theta_i)\,d\theta_i}{\int\int p(\theta_i|\alpha,t_i)\,p(\alpha)\exp(t_i\theta_i)\,d\theta_i d\alpha} \\
&= \frac{p(\alpha)\int p(\theta_i|\alpha,t_i)\exp(t_i\theta_i)\,d\theta_i}{\int p(\alpha)\left[\int p(\theta_i|\alpha,t_i)\exp(t_i\theta_i)\,d\theta_i\right]d\alpha}
\end{aligned}
$$

$$
\begin{aligned}
\log p(\theta_i|\alpha,t_i) &= -\frac{1}{2}\tau\theta_i^2+\tau\theta_i\alpha-\frac{1}{2}\tau\alpha^2+t_i\theta_i+C \\
&= -\frac{1}{2}\tau\theta_i^2+\tau\left(\alpha+\tau^{-1}t_i\right)\theta_i-\frac{1}{2}\tau\left(\alpha+\tau^{-1}t_i\right)^2+\frac{1}{2}\tau\left(\alpha+\tau^{-1}t_i\right)^2-\frac{1}{2}\tau\alpha^2+C \\
&= -\frac{1}{2}\tau\left(\theta_i-\left(\alpha+\tau^{-1}t_i\right)\right)^2+\frac{1}{2}\tau\left(\alpha+\tau^{-1}t_i\right)^2-\frac{1}{2}\tau\alpha^2+C \\
&= \log p(\theta_i|\alpha_t,t_i)+\frac{1}{2}\tau\left(\alpha^2+2\tau^{-1}t_i\alpha+\tau^{-2}t_i^2\right)-\frac{1}{2}\tau\alpha^2+C \\
&= \log p(\theta_i|\alpha_t,t_i)+t_i\alpha+C \\
\alpha_t &:= \alpha+\tau^{-1}t_i \\
\log p(\alpha|t_i) &= \log p(\alpha)++t_i\alpha+C \\
&= -\frac{1}{2}\tau_\alpha\alpha^2+t_i\alpha+C \\
&= -\frac{1}{2}\tau_\alpha\alpha^2+\tau_\alpha\tau_\alpha^{-1}t_i\alpha+C \\
&= -\frac{1}{2}\tau_\alpha\left(\alpha-\tau_\alpha^{-1}t_i\right)^2+C
\end{aligned}
$$

Thus

$$
\begin{aligned}
\alpha|t_i &\sim \mathcal{N}\left(\tau_\alpha^{-1}t_i,\tau_\alpha^{-1}\right) \\
\theta_i|\alpha,t_i &\sim \mathcal{N}\left(\alpha+\tau^{-1}t_i,\tau^{-1}\right)
\end{aligned}
$$

Note that, as expected,

$$
\frac{d}{dt_i}\mathbb{E}\left[\theta_i|\alpha,t_i\right] = \tau^{-1}
$$

Note also that

$$\begin{aligned}
\mathrm{Cov}\left(\theta_i, \alpha\right) &= \mathbb{E}\left[\theta_i \alpha\right] - \mathbb{E}\left[\theta_i\right]\mathbb{E}\left[\alpha\right] \\
&= \mathbb{E}\left[\alpha \mathbb{E}\left[\theta_i | \alpha\right]\right] - \mathbb{E}\left[\alpha^2\right] \\
&= \mathbb{E}\left[\alpha^2\right] - \mathbb{E}\left[\alpha^2\right] \\
&= \mathrm{Var}\left(\alpha\right)
\end{aligned}$$

and that

$$\begin{aligned}
\frac{d}{dt_i}\mathbb{E}\left[\alpha | t_i\right] &= \tau_\alpha^{-1} \\
&= \mathrm{Var}\left(\alpha\right) \\
&= \mathrm{Cov}\left(\theta_i, \alpha\right)
\end{aligned}$$

...also as expected.

4.3.1 Exact tilted posterior

Given t_i we can again integrate out α. First observe that

$$\begin{aligned}
\tau_\alpha^{-1} + \tau^{-1} &= \frac{1}{\tau_\alpha} + \frac{1}{\tau} \\
&= \frac{\tau + \tau_\alpha}{\tau_\alpha \tau} \\
&= \tau_\theta^{-1}
\end{aligned}$$

$$\begin{aligned}
\theta_i | t_i &\sim \mathcal{N}\left(\tau_\theta^{-1} t_i, \tau_\theta^{-1}\right) \\
x_i | \theta_i, t_i &\sim \mathcal{N}\left(\theta_i, 1\right) \Rightarrow \\
\mathbb{E}\left[\theta_i | x_i, t_i\right] &= \frac{1}{1+\tau_\theta}\left(x_i + \tau_\theta \tau_\theta^{-1} t_i\right) \\
&= \frac{1}{1+\tau_\theta}\left(x_i + t_i\right)
\end{aligned}$$

Observe that

$$\left.\frac{d}{dt_i}\mathbb{E}\left[\theta_i | x_i, t_i\right]\right|_{t_i=0} = \frac{1}{1+\tau_\theta}$$

as expected.

4.3.2 EB with a tilted posterior

Now we will fix α and evaluate the EB posterior approximation as a function of t_i and α.

$$
\begin{aligned}
\theta_i|\alpha, t_i &\sim \mathcal{N}\left(\alpha + \tau^{-1}t_i, \tau^{-1}\right) \\
x_i|\theta_i, t_i &\sim \mathcal{N}\left(\theta_i, 1\right) \\
\mathbb{E}\left[\theta_i|x_i, t_i, \alpha\right] &= \frac{1}{1+\tau}\left(x_i + \tau\left(\alpha + \tau^{-1}t_i\right)\right) \\
&= \frac{1}{1+\tau}\left(x_i + \tau\alpha + t_i\right)
\end{aligned}
$$

So that

$$
\begin{aligned}
m &:= \mathbb{E}\left[\theta_i|\alpha, t_i, x_i\right] \\
\frac{\partial m}{\partial t_i} &= \frac{1}{1+\tau} \\
\frac{\partial m}{\partial \alpha} &= \frac{\tau}{1+\tau} = \tau_x
\end{aligned}
$$

4.3.3 $\hat{\alpha}$ as a function of t_i

Next, we see that

$$
\begin{aligned}
\alpha|t_i &\sim \mathcal{N}\left(\tau_\alpha^{-1}t_i, \tau_\alpha^{-1}\right) \\
x_i|\alpha, t_i &\sim \mathcal{N}\left(\alpha + \tau^{-1}t_i, \tau_x^{-1}\right) \Rightarrow \\
x_i - \tau^{-1}t_i|\alpha, t_i &\sim \mathcal{N}\left(\alpha, \tau_x^{-1}\right) \\
\mathbb{E}\left[\alpha|x_i, t_i\right] &= \frac{1}{\tau_\alpha + \tau_x}\left(\tau_\alpha\tau_\alpha^{-1}t_i + \tau_x\left(x_i - \tau^{-1}t_i\right)\right) \\
&= \frac{1}{\tau_\alpha + \tau_x}\left(t_i + \tau_x x_i - \tau_x\tau^{-1}t_i\right) \\
&= \frac{1}{\tau_\alpha + \tau_x}\left(\tau_x x_i + \left(1 - \frac{1}{1+\tau}\right)t_i\right) \\
&= \frac{\tau_x}{\tau_\alpha + \tau_x}\left(x_i + t_i\right)
\end{aligned}
$$

As a function of α, this is maximized at

$$
\hat{\alpha} = \frac{\tau_x}{\tau_\alpha + \tau_x}\left(x_i + t_i\right)
$$

Thus

$$
\frac{d\hat{\alpha}}{dt} = \frac{\tau_x}{\tau_\alpha + \tau_x}
$$

Putting this together, we get that

$$
\begin{aligned}
\frac{dm}{dt} &= \frac{\partial m}{\partial \alpha}\frac{d\alpha}{dt} + \frac{\partial m}{\partial t} \\
&= \frac{\tau_x^2}{\tau_\alpha + \tau_x} + \frac{1}{1+\tau}
\end{aligned}
$$

Note that the difference is exactly what is expected from the above calculation – the linear response variance adjustment is exact.