Math Appendix

Charles C. Margossian
charles.margossian@columbia.edu

1 Estimating Posteriors for Latent Gaussian Models

Latent Gaussian models are a popular class of models and typically have the following hierarchical structure:

$$\begin{aligned} \phi &\sim \pi(\phi) \\ \theta_i &\sim \text{Normal}(0, \Sigma_\phi)^{-1}) \\ y_{j \in g(i)} &\sim p(\theta_i, \phi) \end{aligned}$$

where ϕ is a global parameter and θ a local parameter. The observations y_j belong to local groups indexed by $g(i)$ and follow distributions parametrized by θ_i. In this article, I focus on the common case where ϕ is low dimensional and θ high dimensional.

Our goal is to make inference about ϕ. In a Bayesian setting, this amounts to computing the posterior:

$$\begin{aligned} p(\phi|y) &= \frac{p(\phi, y)}{p(y)} \\ &= \frac{p(\phi, y)p(\theta|\phi, y)}{p(y)p(\theta|\phi, y)} \\ &= \frac{p(\phi, \theta, y)}{p(\theta|y, \phi)p(y)} \\ &\propto \frac{p(y|\theta, \phi)p(\theta|\phi)p(\phi)}{p(\theta|y, \phi)} \end{aligned} \tag{1}$$

The model gives us the terms in the nominator but not in the denominator. A straightforward way to tackle this is to do a full Bayesian inference on both ϕ and θ. However, doing so significantly increases the dimension of our model's parameter space.

The alternative approach is to perform inference on ϕ only, and approximate the conditional density in the denominator as a Gaussian density. That is

$$p(\theta|y, \phi) \approx p_{\mathcal{G}}(\theta)$$

where $p_{\mathcal{G}}$ is a normal density centered at the mode of $p(\theta|y, \phi)$, which we denote θ^*. Moreover, we use the approximation, $\theta \approx \theta^*$, in our calculation of the posterior. The curvature, $\mathcal{H}$, of $p_{\mathcal{G}}$ matches that of $p(\theta|y, \phi)$.

The here discussed strategy was first proposed by (Tierney & Kadane, 1986), who showed that, under certain regularity conditions, the error of the approximation is given by:

$$p(\theta|y, \phi) = p_{\mathcal{G}}(\theta)(1 + \mathcal{O}(n^{-\frac{3}{2}}))$$

where n is the number of observations. Note the error is relative, and furthermore the rate of convergence is a factor n larger than what we get from the central limit theorem. The above-mentioned regularity conditions apply, among other cases, when y follows a normal, Poisson, binomial, or negative-binomial distribution.

The main benefit of using a Laplace approximation is that the Markov chain only explores the parameter space of ϕ, as opposed to the joint space of ϕ and θ. But finding the mode, θ^*, comes at a significant cost, as this requires solving a high-dimensional algebraic equation. This trade-off informs which models and regime the approximation works best.

1.1 Calculating the approximate Posterior

The mode, θ^*, is found with a numerical solver. The curvature $\mathcal{H}$ is evaluated either analytically or numerically, depending on the difficulty of the problem. In both cases, the details depend on the specifics of the model, in particular the distribution $p(y|\theta, \phi)$. As an example, I work out the objective function we need to optimize when fitting a log poisson model with a latent Gaussian parameter in section 1.2.

But first, let us derive some more general results.

$\mathcal{H}$ may be found using the following lemma:

Lemma 1. *Let $\mathcal{H}$ be the Hessian of $p(\theta|y, \phi)$ at $\theta = \theta^*$. Then*

$$\mathcal{H} = \Sigma_\phi^{-1} + H$$

where H is the Hessian of $\log p(\theta|y, \phi)$ at $\theta = \theta^*$.

Proof. Work out proof. □

Our Laplace approximation is then

$$p_G(\theta) = \text{Normal}(\theta^*, (\Sigma_\phi^{-1} + H)^{-1})$$

We can then explicitly write the multivariate Gaussian distributions in our approximation of the posterior (equation 1):

$$p(\theta|\phi) = \left(\frac{1}{2\pi \text{det}|\Sigma_\phi|}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\theta^{*T}\Sigma_\phi^{-1}\theta^*\right)$$

and

$$\begin{aligned} p_G(\theta^*) &= \left(\frac{1}{2\pi \text{det}|(\Sigma_\phi^{-1} + H)^{-1}|}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\theta^* - \theta^*)^T(\Sigma_\phi^{-1} + H)(\theta^* - \theta^*)\right) \\ &= \left(\frac{1}{2\pi}\text{det}|\Sigma_\phi^{-1} + H|\right)^{\frac{1}{2}} \end{aligned}$$

where we used the approximation $\theta \approx \theta^*$ and the fact that, for an invertible matrix A, $\text{det}|A^{-1}| = (\text{det}|A|)^{-1}$. Combining all our results, the approximate posterior becomes

$$\begin{aligned} p(\phi|y) &\approx p(\phi)p(y|\theta^*, \phi)\frac{p(\theta^*|\phi)}{p_G(\theta^*)} \\ &= p(\phi)p(y|\theta^*, \phi)\left(\frac{1}{\text{det}|\Sigma_\phi|\text{det}|\Sigma_\phi^{-1} + H|}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\theta^{*T}\Sigma_\phi^{-1}\theta^*\right) \end{aligned}$$

or on the log scale

$$\log p(\phi|y) \approx \log p(\phi) + \log p(y|\theta^*, \phi) - \frac{1}{2}\left(\log\det|\Sigma_\phi| + \log\det|\Sigma_\phi^{-1} + H| + \theta^{*T}\Sigma_\phi^{-1}\theta^*\right)$$

1.2 Log Poisson model with latent Gaussian parameter

To test the performance of the Laplace approximation, I construct a computer experiment in which a full Bayesian inference is performed on the following model:

$$\begin{aligned} \phi \sim & \quad \text{Normal}(0, 2) \\ \theta \sim & \quad \text{Normal}(0, \Sigma_\phi) \\ y_{j \in g(i)} \sim & \quad \text{Poisson}(e^{\theta_i}) \end{aligned} \tag{2}$$

where Σ_ϕ is a diagonal covariance matrix, which deterministically depends on ϕ.

To apply the Laplace approximation we first need to find the mode of $p(\theta|\theta, y)$. Applying Bayes' rule:

$$p(\theta|y, \phi) \propto p(y|\theta, \phi)p(\theta|\phi) \tag{3}$$

By equation 2, the right hand side is a product of poisson and normal distributions. Since our goal is to find the mode, i.e. is optimize the function for θ, we can ignore normalizing constants. Let

$$\begin{aligned} m_i &= \sum_{j \in g(i)} 1 \\ S_i &= \sum_{j \in g(i)} y_j \end{aligned}$$

respectively the total number of terms and the total number of counts in the i^{th} group. Then, on the log scale, the objective function is:

$$f(\theta) = \quad \left\{ \textstyle\sum_{i=1}^M S_i \theta_i - e^{\theta_i} \right\} - \tfrac{1}{2} \theta^T \Sigma^{-1} \theta \tag{4}$$

Using the fact Σ^{-1} is symmetric, the gradient is then:

$$\nabla f(\theta) = \mathcal{V} - \Sigma^{-1} \theta \tag{5}$$

where $\mathcal{V}_i = S_i - m_i e^{\theta_i}$.

Noting the normalizing constant can be dropped in the log scale, the Hessian H is easily worked out from equation 5 to be

$$H(\theta) = \mathcal{W} - \Sigma^{-1}$$

where $\mathcal{W}$ is a diagonal matrix with $\mathcal{W}_i = -m_i e^{\theta_i}$. The log posterior is thus:

$$\log p(\phi|y) \approx \log p(\phi) + \log p(y|\theta^*, \phi) - \frac{1}{2} \Big(\log \det |\Sigma_\phi| + \log \det |H| + \theta^{*T} \sigma_\phi^{-1} \theta^* \Big)$$

As a starting point, we consider the case where Σ_ϕ is a diagonal matrix with entries $\sigma_i = \phi^2$. Then equation 4 becomes:

$$\log p(\theta|y, \phi) \quad = \quad \sum_{i=1}^M \left(S_i \theta_i - e^{\theta_i} \right) - \frac{1}{2\phi^2} \sum_{i=1}^M \theta_i^2$$

and the gradient and the hessian are:

$$\begin{aligned} \frac{\partial}{\partial \theta_i} \log p(\theta_i|y, \phi) &= S_i - m_i e^{\theta_i} - \frac{\theta_i}{\phi^2} \\ \frac{\partial^2}{\partial \theta_i^2} \log p(\theta_i|y, \phi) &= -m_i e^{\theta_i} - \frac{1}{\phi^2} \end{aligned}$$

Note these partial derivatives fully define the gradient and the Hessian, as the θ_i's are uncorrelated. The corresponding log posterior is then

$$\log p(\phi|y) \approx \log p(\phi) + \log p(y|\theta^*, \phi) - \frac{1}{2} \left(M \log \sigma^2 + \sum_{i=1}^M (\log m_i + \theta_i^*) + \theta^{*T} \sigma_\phi^{-1} \theta^* \right)$$

which seems wrong.

In the more general case, we need to worry about off-diagonal terms and use linear algebra operators which can take advantage of matrix sparsity.

References

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*(393), 82-86. Retrieved from `https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1986.10478240` doi: 10.1080/01621459.1986.10478240