19. Model Identifiability

This chapter discusses the frequentist notion of model identifiability and the Bayesian notion of identifiability in terms of well-behaved posteriors, concentrating on computational issues arising from model formulation and computation in Stan.

19.1. Identifiability of Likelihoods

In traditional frequentist statistics, a model is said to be identifiable if and only if different parameters produce different likelihood functions. More formally, a likelihood function $f_\theta(y) = p(y|\theta)$ is identifiable if and only if

$$\theta \neq \theta' \text{ implies } f_\theta \neq f_{\theta'}.$$

The inequality on the right hand side of this definition is between functions of y, meaning that $f_\theta \neq f_{\theta'}$ if there is some y such that $f_\theta(y) \neq f_{\theta'}(y)$.

Model identifiability is a necessary (but not sufficient) condition for the existence of maximum likelihood estimates (MLE). Without identifiability, the maximum likelihood estimate

$$\hat{\theta} = \text{argmax}_\theta \; p(y|\theta)$$

might not be unique.

Examples

A simple normal model with a location parameter μ, a scale parameter $\sigma > 0$, and likelihood

$$p(y|\mu,\sigma) = \prod_{n=1}^{N} \text{Normal}(y_n|\mu,\sigma).$$

is identifiable because every distinct value of μ and σ produces a different likelihood function $p(y|\mu,\sigma)$.

A similar model[1] with two location parameters, λ_1 and λ_2, a scale $\sigma > 0$, and likelihood function

$$p(y|\lambda_1,\lambda_2,\sigma) = \prod_{n=1}^{N} \text{Normal}(y_n|\lambda_1+\lambda_2,\sigma)$$

[1] This example was raised by Richard McElreath on the Stan users group in a query about the behavior of the no-U-turn sampler (NUTS).

is not identifiable because for any non-zero quantity q,

$$p(y|\lambda_1, \lambda_2, \sigma) = p(y|\lambda_1 + q, \lambda_2 - q, \sigma).$$

Another example of a non-identifiable model is a normal mixture model, with two location parameters μ_1 and μ_2, a shared scale $\sigma > 0$, a mixture ratio $\theta \in [0, 1]$, and likelihood

$$p(y|\theta, \mu_1, \mu_2, \sigma) = \prod_{n=1}^{N} \left(\theta \times \mathsf{Normal}(y_n|\mu_1, \sigma) + (1 - \theta) \times \mathsf{Normal}(y_n|\mu_2, \sigma)\right).$$

The issue here is exchangeability of the labels 1 and 2, because

$$p(y|\theta, \mu_1, \mu_2, \sigma) = p(y|(1 - \theta), \mu_2, \mu_1, \sigma).$$

19.2. Bayesian "Identifiability"

In the broadest sense, a Bayesian model is identified if the posterior distribution,

$$p(\theta|y) \propto p(y|\theta)\, p(\theta),$$

is proper, i.e.,

$$\int p(\theta|y)\, d\theta = 1.$$

Mathematically speaking, with a proper posterior, one can do Bayesian inference and that's that. There is not even a need to require a finite variance or even a finite mean—all that's needed is a finite integral.

Examples (continued)

Consider again the non-identifiable model from the previous section with two mean parameters. Now suppose we create a Bayesian model with the likelihood $p(y|\lambda_1, \lambda_2, \sigma)$ and an improper uniform prior $p(\lambda_1, \lambda_2, \sigma) = 1$. The result is an improper posterior

$$p(\lambda_1, \lambda_2, \sigma|y) \;\propto\; p(y|\lambda_1, \lambda_2, \sigma)\, p(\lambda_1, \lambda_2, \sigma) \;=\; p(y|\lambda_1, \lambda_2, \sigma).$$

The posterior contains a ridge with a peak along the line where the sum $\lambda_1 + \lambda_2$ is equal to the maximum likelihood estimate $\hat{\mu}$ from the identified model

$$p(y, \mu, \sigma) = \prod_{n=1}^{N} \mathsf{Normal}(y_n|\mu, \sigma).$$

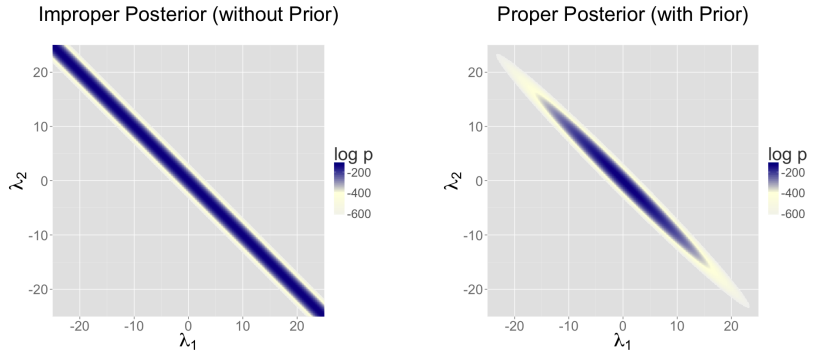

Figure 19.1: *Posteriors of an unidentified model and the same model identified with a prior. In both models, the likelihood function is* $y_n \sim \text{Normal}(\lambda_1 + \lambda_2, \sigma)$. *Both plots show the posterior for the same 100 data points simulated from a unit normal distribution. The left-hand plot shows the posterior for the model with no priors for* λ_1 *or* λ_2 *(or* σ*). The resulting posterior has a ridge extending infinitely to the northwest and southeast, and is thus improper (does not integrate to 1). The posterior density in the right-hand plot is from a model that adds unit normal priors* $\lambda_1, \lambda_2 \sim \text{Normal}(0, 1)$*, resulting in a proper posterior.*

The posterior ridge is illustrated in the plot in the left-hand figure of Figure 19.1, which shows the posterior resulting from sampling 100 data points $y_n \sim \text{Normal}(0, 1)$. With 100 data points, the posterior for σ is proper, even in the model parameterized by λ_1 and λ_2. With improper posteriors, sampling becomes impossible. A "proper" posterior sample should spend as much time in the neighborhood of $\lambda_1 = 1000000000$ and $\lambda_2 = -1000000000$ as it does in the neighborhood of $\lambda_1 = 0$ and $\lambda_2 = 0$, and so on for ever larger values. This is, of course, impossible in finite amounts of time, not to mention with limited maximum and minimum values for double-precision floating point numbers as used in computers.

By way of contrast, consider what happens when we add proper priors for λ_1 and λ_2, say

$$p(\lambda_1, \lambda_2) \;=\; \text{Normal}(\lambda_1 | 0, 1) \times \text{Normal}(\lambda_2 | 0, 1).$$

With proper priors on λ_1 and λ_2, the posterior is now proper. The effect on the posterior is to convert the ridge into a hill, as illustrated in the right-hand plot in Figure 19.1. With the same 100 data points y_n, the probability density quickly falls off as λ_1 and λ_2 move away from the origin $(0, 0)$.

19.3. What Goes Wrong Sampling in Non-Identifiable Models

With an improper posterior, it is theoretically impossible to properly explore the posterior. However, Gibbs sampling as performed by BUGS and JAGS behaves quite dif-

ferently the Hamiltonian Monte Carlo sampling performed by Stan when faced with the two-location model discussed in the previous section.

Gibbs Sampling

Gibbs sampling, as performed by BUGS and JAGS, may appear to be efficient and well behaved for this unidentified model, but as discussed in the previous subsection, will not actually explore the posterior properly.

Consider what happens with initial values $\lambda_1^{(0)}, \lambda_2^{(0)}$. Gibbs sampling proceeds in iteration m by drawing

$$\begin{aligned}
\lambda_1^{(m)} &\sim p(\lambda_1 \mid \lambda_2^{(m-1)}, \sigma^{(m-1)}, y) \\
\lambda_2^{(m)} &\sim p(\lambda_2 \mid \lambda_1^{(m)}, \sigma^{(m-1)}, y) \\
\sigma^{(m)} &\sim p(\sigma \mid \lambda_1^{(m)}, \lambda_2^{(m)}, y).
\end{aligned}$$

Now consider the draw for λ_1 (the draw for λ_2 is symmetric), which is conjugate in this model and thus can be done very efficiently. In this model, the range from which the next λ_1 can be drawn is highly constrained by the current values of λ_2 and σ. Gibbs will run very quickly and provide excellent inference for $\lambda_1 + \lambda_2$. But it will not explore the full range of the posterior; it will merely take a slow random walk from the initial values. This random walk behavior is typical of Gibbs sampling when posteriors are highly correlated and the primary reason to prefer Hamiltonian Monte Carlo to Gibbs sampling for models with parameters correlated in the posterior.

Hamiltonian Monte Carlo Sampling

Hamiltonian Monte Carlo (HMC), as performed by Stan, is much more efficient at exploring posteriors in models where parameters are correlated in the posterior. In this particular example, the Hamiltonian dynamics (i.e., the motion of a fictitioous particle given random momentum in the field defined by the negative log posterior) is going to run up and down along the valley defined by the potential energy (ridges in log posteriors correspond to valleys in potential energy). In practice, even with a random momentum for λ_1 and λ_2, the gradient of the log posterior is going to adjust for the correlation and the simulation will run along the valley corresponding to the ridge in the posterior log density.

No-U-Turn Sampling

The no-U-turn sampler (NUTS), the default form of HMC used in Stan, shows even more pathological behavior in the face of non-identifiability. Because NUTS tries to

simulate the motion of the fictitious particle representing the parameter values until it makes a U-turn, it will be defeated in most cases, as it will just move down the potential energy valley indefinitely without making a U-turn. What happens in practice is that the maximum number of leapfrog steps in the simulation will be hit in many of the iterations, causing a very large number of log probability and gradient evaluations (1000 if the max tree depth is set to 10, as in the default). Thus sampling will appear to be very slow. But like the case for HMC and Gibbs, sampling is not just slow, it's impossible in this case. NUTS continues to explore more of the posterior density than HMC with few leapfrog steps and much more of the posterior than a Gibbs sampler. The problem is that it's never possible to explore all of it because it's improper.

Examples: Fits in Stan

To illustrate the issues with sampling from non-identified and only weakly identified models, we fit three models with increasing degrees of identification of their parameters. The first model is the unidentified model with two location parameters and no priors.

```
data {
  int N;
  real y[N];
}
parameters {
  real lambda1;
  real lambda2;
  real<lower=0> sigma;
}
transformed parameters {
  real mu;
  mu <- lambda1 + lambda2;
}
model {
  y ~ normal(mu, sigma);
}
```

The second adds priors to the model block for `lambda1` and `lambda2` to the previous model.

```
  lambda1 ~ normal(0,10);
  lambda2 ~ normal(0,10);
```

The third involves a single location parameter, but no priors.

Two Scale Parameters, Improper Prior

```
Inference for Stan model: improper_stan
Warmup took (2.7, 2.6, 2.9, 2.9) seconds, 11 seconds total
Sampling took (3.4, 3.7, 3.6, 3.4) seconds, 14 seconds total
```

	Mean	MCSE	StdDev	5%	95%	N_Eff	N_Eff/s	R_hat
lp__	-5.3e+01	7.0e-02	8.5e-01	-5.5e+01	-5.3e+01	150	11	1.0
n_leapfrog__	1.4e+03	1.7e+01	9.2e+02	3.0e+00	2.0e+03	2987	212	1.0
lambda1	1.3e+03	1.9e+03	2.7e+03	-2.3e+03	6.0e+03	2.1	0.15	5.2
lambda2	-1.3e+03	1.9e+03	2.7e+03	-6.0e+03	2.3e+03	2.1	0.15	5.2
sigma	1.0e+00	8.5e-03	6.2e-02	9.5e-01	1.2e+00	54	3.9	1.1
mu	1.6e-01	1.9e-03	1.0e-01	-8.3e-03	3.3e-01	2966	211	1.0

Two Scale Parameters, Weak Prior

```
Warmup took (0.40, 0.44, 0.40, 0.36) seconds, 1.6 seconds total
Sampling took (0.47, 0.40, 0.47, 0.39) seconds, 1.7 seconds total
```

	Mean	MCSE	StdDev	5%	95%	N_Eff	N_Eff/s	R_hat
lp__	-54	4.9e-02	1.3e+00	-5.7e+01	-53	728	421	1.0
n_leapfrog__	157	2.8e+00	1.5e+02	3.0e+00	511	3085	1784	1.0
lambda1	0.31	2.8e-01	7.1e+00	-1.2e+01	12	638	369	1.0
lambda2	-0.14	2.8e-01	7.1e+00	-1.2e+01	12	638	369	1.0
sigma	1.0	2.6e-03	8.0e-02	9.2e-01	1.2	939	543	1.0
mu	0.16	1.8e-03	1.0e-01	-8.1e-03	0.33	3289	1902	1.0

One Scale Parameter, Improper Prior

```
Warmup took (0.011, 0.012, 0.011, 0.011) seconds, 0.044 seconds total
Sampling took (0.017, 0.020, 0.020, 0.019) seconds, 0.077 seconds total
```

	Mean	MCSE	StdDev	5%	50%	95%	N_Eff	N_Eff/s	R_hat
lp__	-54	2.5e-02	0.91	-5.5e+01	-53	-53	1318	17198	1.0
n_leapfrog__	3.2	2.7e-01	1.7	1.0e+00	3.0	7.0	39	507	1.0
mu	0.17	2.1e-03	0.10	-3.8e-03	0.17	0.33	2408	31417	1.0
sigma	1.0	1.6e-03	0.071	9.3e-01	1.0	1.2	2094	27321	1.0

Figure 19.2: *Results of Stan runs with default parameters fit to $N = 100$ data points generated from $y_n \sim \text{Normal}(0, 1)$. On the top is the non-identified model with improper uniform priors and likelihood $y_n \sim \text{Normal}(\lambda_1 + \lambda_2, \sigma)$. In the middle is the same likelihood as the middle plus priors $\lambda_k \sim \text{Normal}(0, 10)$. On the bottom is an identified model with an improper prior, with likelihood $y_n \sim \text{Normal}(\mu, \sigma)$. All models estimate μ at roughly 0.16 with very little Monte Carlo standard error, but a high posterior standard deviation of 0.1; the true value $\mu = 0$ is within the 90% posterior intervals in all three models.*

```
data {
  int N;
  real y[N];
}
parameters {
  real mu;
  real<lower=0> sigma;
}
model {
  y ~ normal(mu, sigma);
}
```

All three of the example models models were fit in CmdStan 2.1.0 with default parameters (1000 warmup iterations, 1000 sampling iterations, NUTS sampler with max tree depth of 10). The results are shown in Figure 19.2. The key statistics from these outputs are the following.

- As indicated by `R_hat` column, all parameters have converged other than λ_1 and λ_2 in the non-identified model.
- The average number of leapfrog steps is roughly 3 in the identified model, 150 in the model identified by a weak prior, and 1400 in the non-identified model.
- The number of effective samples per second for μ is roughly 31,000 in the identified model, 1900 in the model identified with weakly informative priors, and 200 in the non-identified model; the results are similar for σ.
- In the non-identified model, the 95% interval for λ_1 is (-2300,6000), whereas it is only (-12,12) in the model identified with weakly informative priors.
- In all three models, the simulated value of $\mu = 0$ and $\sigma = 1$ are well within the posterior 90% intervals.

19.4. What can Go Wrong with Proper Posteriors

Multimodal Posteriors

Even with proper priors, the normal mixture model discussed earlier will have a posterior $p(\theta, \mu_1, \mu_2, \sigma | y)$ that can be difficult in practice due to multiple modes that swap the indexes and value of θ. The model is not identified in any real sense.

Theoretically, this should not present a problem for inference because all of the integrals involved in posterior predictive inference will be well behaved. The problem in practice is computation. Even if the posterior is proper, MCMC samplers are

notoriously ineffective at exploring multiple modes efficiently, especially when the number of modes grows exponentially as it does for mixture models with increasing numbers of components. In Gibbs sampling, it is unlikely for μ_1 to move to a new mode when sampled conditioned on the current values of μ_2 and θ. For HMC and NUTS, the problem is that the sampler gets stuck in one of the two "bowls" arounds the modes and cannot gather enough energy from random momentum assignment to move from one mode to another.

Supposing a sample could adequately explore multiple posterior modes, a further complication arises due to the fact that posterior means and standard deviations provide poor summaries of the posterior, thus complicating issues such as convergence monitoring and effective sample size estimation.

One possibility for dealing with multimodal posteriors that arise from label switching in mixture models is to somehow define an ordering on the values, such as requiring $\mu_1 < \mu_2$. This can be achieved with Stan's constraint language, but can lead to estimation bias if the posterior uncertainty of μ_1 overlaps with that of μ_2. In other cases, such as a discrimination parameter δ_j in an item-response theory model, restricting $\delta_j > 0$ can solve the identifiability problem.

Posteriors with Unbounded Densities

In some cases, the posterior density grows without bounds as parameters approach certain poles or boundaries. In thse cases, there are no maximum likelihood estimates. One such example is a binary mixture model with scales varying by component, σ_1 and σ_2 for locations μ_1 and μ_2. In this situation, the density grows without bound as $\sigma_1 \to 0$ and $\mu_1 \to y_n$ for some n; that is, one of the mixture components concentrates all of its mass around a single data item y_n.

Another example of unbounded densities arises with a posterior such as $\mathsf{Beta}(\phi|0.5, 0.5)$, which can arise if very "weak" beta priors are used for groups that have no data. This density is unbounded as $\phi \to 0$ and $\phi \to 1$. Similarly, a Bernoulli likelihood model coupled with a "weak" beta prior, leads to a posterior

$$\begin{aligned} p(\phi|y) &\propto \mathsf{Beta}(\phi|0.5, 0.5) \times \textstyle\prod_{n=1}^{N} \mathsf{Bernoulli}(y_n|\phi) \\ &= \mathsf{Beta}(\phi \mid 0.5 + \textstyle\sum_{n=1}^{N} y_n, \;\; 0.5 + N - \textstyle\sum_{n=1}^{N} y_n). \end{aligned}$$

If $N = 10$ and each $y_n = 1$, the posterior is $\mathsf{Beta}(\phi|10.5, 0, 5)$, which is unbounded as $\phi \to 1$.

Because the posterior is proper even in these cases (i.e., the posterior mass does not grow without bound), Bayesian inference can often overcome the difficulty with maximum likelihood inference and calculate proper posterior means. In other cases, when the unbounded posterior modes are attractive enough, the simulated particle falls down an infinitely deep well corresponding to unbounded negative log posterior

density and never gains enough random kinetic energy in future iterations to climb back out again.

Uniform Posteriors

Suppose your model includes a parameter ψ that is defined on $[0, 1]$ and is given a flat prior $\mathsf{Uniform}(\psi|0, 1)$. Such a model is guaranteed to have a proper posterior for ψ no matter what the data looks like. But suppose the data don't tell us anything about ψ, so that our posterior is also $\mathsf{Uniform}(\psi|0, 1)$. The maximum likelihood estimate is ill defined, but the Bayesian posterior is proper. The posterior mean for ψ is well defined at $1/2$, but the model still feels non-identified. Nevertheless, posterior predictive inference may do the right thing by simply integrating (i.e., averaging) over the predictions for ψ at all points in $[0, 1]$.

19.5. Weak Identification

Suppose that with reasonable data, you'd have a posterior with a standard deviation of 1 (or that order of magnitude). But suppose you have sparse data or co-linearity of predictors, and so you have some dimension in your posterior that's really flat—essentially a "ridge" with a standard deviation of 1000. Then it makes sense to say that this parameter or linear combination of parameters is only weakly identified. Or one can say that it's identified from the prior but not the likelihood.

Although technically sound in the mathematical sense, a posterior that is only weakly identified by the data in this way can be problematic for practical inference. With very weak priors, the ridge illustrated in Figure 19.1 becomes a very gently sloping hill, and sampling remains problematic due to the large extent of the posterior that must be explored.

In general, identification depends not just on the model but also on the data. So, strictly speaking, one should not talk about an "identifiable model" but rather an "identifiable fitted model" or "identifiable parameters" within a fitted model.

That is, we can think of a weakly informative prior as being one that supplies relatively little information compared to the data about the posterior. A strongly informative prior, on the other hand, supplies relatively more information than the data about the posterior. The dividing line between "strong" and "weak" here is not well defined, and also depends on the amount.

The crucial notion computationally is that the data plus the prior together need to provide enough information about the posterior that it is not effectively a very long ridge of equal density.

It is common to see very diffuse priors applied to parameters in BUGS or JAGS models, such as $\mathsf{Uniform}(-20000, 20000)$ or $\mathsf{Normal}(0, 10000)$. Presumably such pri-

ors are motivated by the desire of modelers to produce Bayesian posterior mean estimates for a parameter that are very close to the maximum likelihood estimate obtained in a non-Bayesian setting.

Although these very diffuse priors do identify models in theory, they are not very effective in practice. When considering their effect on the geometry of the posterior, the reason is obvious—they flatten the ridge that would otherwise arise in a non-identified model, but only very gently. Adding a very diffuse prior to a non-identified model like $y_n \sim \mathsf{Normal}(\lambda_1 + \lambda_2, \sigma)$ causes major headaches and is not much better than just using the non-identified model without any prior at all.

It is important to keep in mind that Gibbs samplers like BUGS or JAGS do not effectively explore the full posterior entailed by these diffuse priors, instead devolving into a random walk. It may just seem like they are better behaved and exploring the posterior because of inferences for transformed parameters such as $\lambda_1 + \lambda_2$ in the models driving Figure 19.2.

On the other hand, Hamiltonian Monte Carlo samplers like Stan will do a better job at exploring the full posterior, but it will take them considerable time to explore the extent of the shallow valleys induced in the negative log posterior. As can be seen in the behavior of the non-identifed model in Figure 19.2, the posterior scale should be less than 1000 for the no-U-turn sampler to behave even remotely reasonably with Stan's default settings.

The ideal solution would be to employ informative priors based on knowledge of the data being modeled. The modeler will almost always have some grip on the scale of the parameters expected based on the data being modeled. If -25,000 or +25,000 are not reasonable values for a parameter, a prior of $\mathsf{Normal}(0, 10000)$ is not reasonable.

A kind of halfway solution is to use weakly informative priors that indicate the rough scale of the estimates without being too diffuse or exerting too much influence on the posterior. For example, if we expect a parameter to be in the range (-2,2), as we would for a logistic regression coefficient for a standardized (mean 0, deviation 1) predictor, then a prior for the coefficient of $\mathsf{Normal}(0, 5)$ will be enough to bring the posterior geometry under control while at the same time not providing a noticeable shrinking of the posterior toward 0.