9.2 Multiple Coin From a Single Mint

Zongyi Liu

2023-06-20

9.2 Multiple Coin From a Single Mint

The previous section considered a scenario in which we flip a single coin and make inferences about the bias θ and the hyperparameter μ of the coin. In this section we will consider another case. When we collect data from more than one coin. If each coin has its own distinct bias θ_j, then we are estimating a distinct parameter value for each coin. For now, we assume that all the coins have come from the same mint. This means that we have the same prior belief about$\mu$for all the coins. We also assume that each coin is minted independently of the others. This means that each coin's bias is independent of the others (conditional on μ), in our prior belief distribution.

We set up the experiment so that subjects don't interact with each other, and so we assume that individual biases are independent of each other.1 We "flip the coin" by measuring Bernoulli responses from the subject. For example, suppose that the drug is supposed to affect memory. We can test memory by giving the subject a list of random words to study, and then checking how many words can be recalled several minutes later.

The scenario is summarized in Figure 9.4, which is very like Figure 9.1, but with a small change.

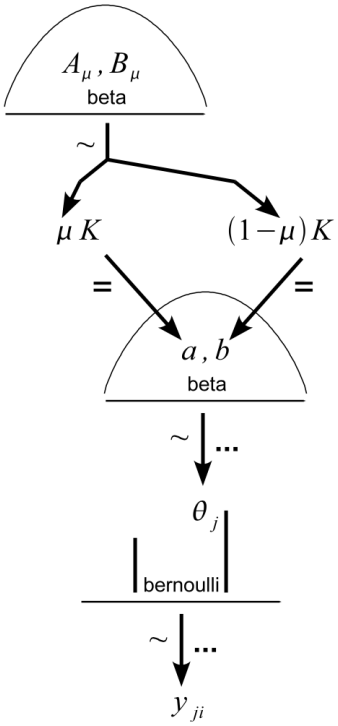


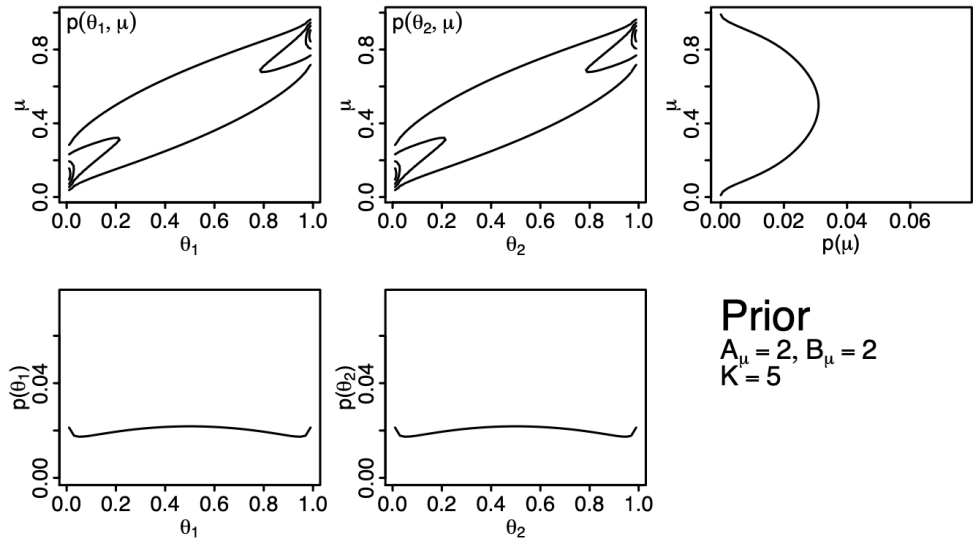
Figure 9.4: A model of hierarchical dependencies for data from J coins created independently from the same mint. A datum y_{ji}, from the i^{th} flip of the j^{th} coin, depends on the value of the bias parameter θ_j for the coin. The values of θ_j depend on the value of the hyperparameter μ for the mint that created the coins. The ellipsis on the dependency arrows denotes the repetition of the dependency across flips (for y_{ji}) or coins (for θ_j). The μ parameter has a prior belief distributed as a beta distribution with shape parameters A_μ and B_μ. We simultaneously estimate the $J + 1$ parameters: $\theta_1, ..., \theta_J$, and μ. This case is discussed in Section 9.2.

Instead of there being a single θ value, there is now a different θ value for each coin, with the bias of the jth coin denoted θ_j. Because the individual flips of the coins come from different coins, the flip results are double-subscripted, such that the ith flip of the jth coin is denoted y_{ji}.

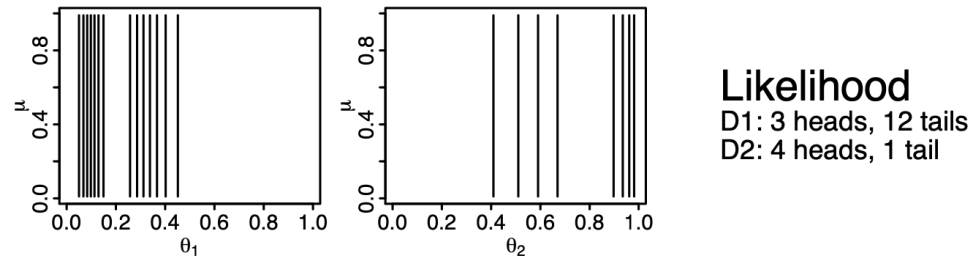
9.2.1 Posterior via Grid Approximation

As a concrete example, suppose we have two coins from the same mint. We want to estimate the biases θ_1 and θ_2 of the two coins, and simultaneously estimate μ of the mint that created them.

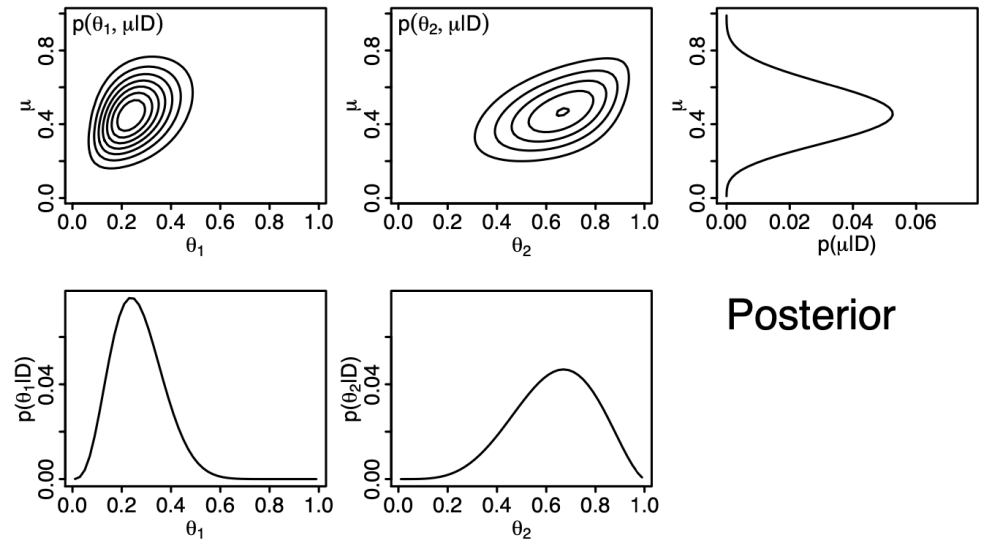
In Figure 9.5, the prior on μ is gently peaked over $\mu = 0.5$, in the form of a $beta(\mu|2, 2)$ distribution; that is, $A_\mu = B_\mu = 2$ in the top-level formula of Figure 9.4. The biases of the coins are only weakly dependent on μ according to the prior $p(\theta_j|\mu) = beta(\theta_j|\mu * 5, (1-\mu) * 5)$, which means $K = 5$ in the middle-level formula of Figure 9.4. The full prior distribution is a joint distribution over three parameters: μ, θ_1, and θ_2. In a grid approximation, the prior is specified as a 3D array that holds the prior probability at various grid points in the 3D space.



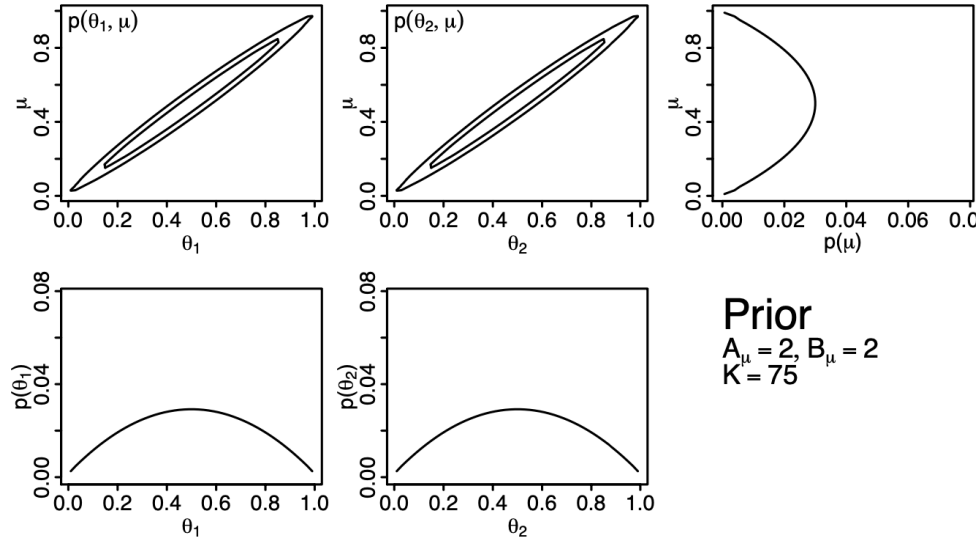
The middle row of Figure 9.5 shows the likelihood function for the data, which comprise 3 heads out of 15 flips of the first coin, and 4 heads out of 5 flips of the second coin.



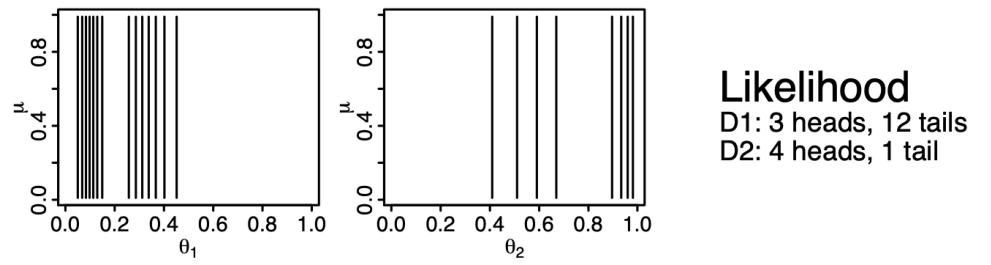
The lower two rows of Figure 9.5 show the posterior distribution. Notice that the po terior on θ_1 is centered not far from the proportion $3/15 = .2$ in its coin's data, and the posterior on θ_2 is centered not far from the proportion $4/5 = .8$ in its coin's data.



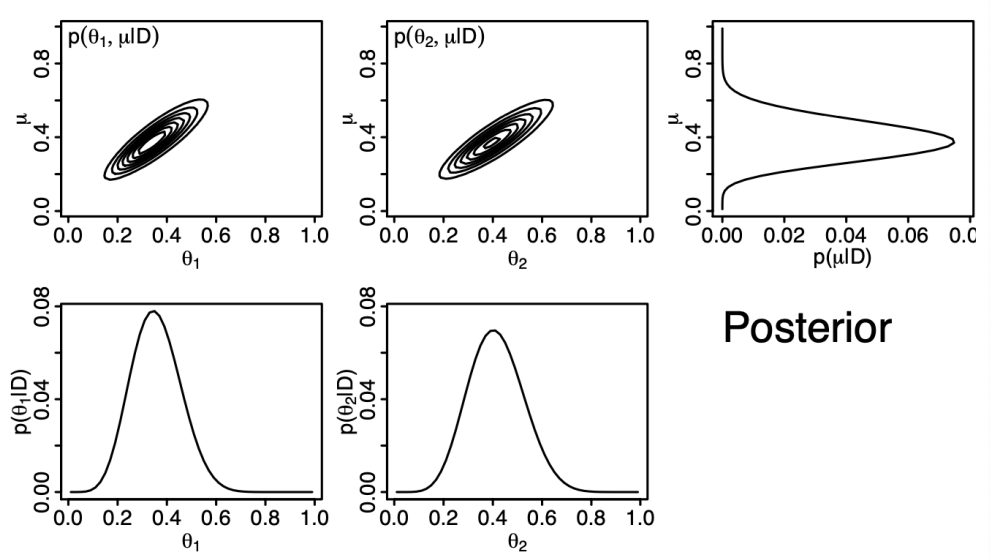
In Figure 9.6, we use the same data but a different prior. In Figure 9.6, the prior on$\mu$is the same gentle peak, but the prior dependency of θ_j on μ is much stronger. The dependency can be seen graphically in the top two panels of Figure 9.6, which show contour plots of the marginals $p(f_j, \mu)$. The contours reveal that when θ_j is not very close to μ, away from the diagonal, then the probability $p(\theta_j, \mu)$ is very small.



Likelihood is the same as before



The plots of the posterior distribution, in the lower rows of Figure 9.6, reveal some very interesting results.



Because the biases and the hyperparameter are being simultaneously estimated, and the biases are strongly dependent on the hyperparameter, the posterior estimates are fairly tightly constrained, especially in comparison with Figure 9.5. Essentially, because the prior emphasizes a relatively narrow spindle within the 3D box, the posterior is restricted to a zone within that spindle. Not only does this cause the posterior to be relatively peaked on all the parameters, it also pulls all the estimates in toward the focal zone.

9.2.2 Posterior via Monte Carlo Sampling

The previous sections have used a simplified model (believe it or not) for the purpose of being able to graphically display the parameter space and gain clear intuitions about how Bayesian inference works. In this section, the first thing we'll do is include one more parameter in the model, to make it more realistic. The previous examples arbitrarily fixed the degree of dependency of θ on μ. The degree of dependency was specified as the value of K, such that when K was large, the individualθ_j values stayed close to μ, but when K was small, the individual θ_j values could spread quite far from μ.

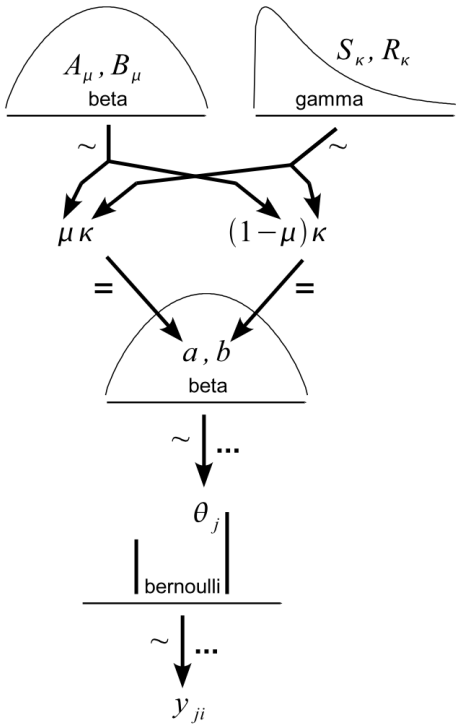


Figure 9.7: A model of hierarchical dependencies for data from J coins created independently from the same mint, with the uncertainty of the mint parameterized by μ and κ. The datum y_{ji} from the i^{th} flip of the j^{th} coin depends on the value of the coin's bias parameter θ_j. The values of θ_j depend on the value of the hyperparameters μ and κ for the mint that created the coins. The μ parameter has a prior belief distributed as a beta distribution with shape parameters A_μ and B_μ, while the κ parameter has a prior belief distributed as a gamma distribution with shape and rate parameters of S_κ and R_κ. We simultaneously estimate the $J+2$ parameters, $\theta_1, ..., \theta_J, \mu$, and κ. This case is discussed in Section 9.2.2.

In real situations, we don't know the value of K in advance, and instead we let the data inform us regarding its credible values. Intuitively, when the proportions of heads in the different coins are all very similar to each other, we have evidence that K is high. But when the proportions of heads in the different coins are very diverse, then we have evidence that K is small.

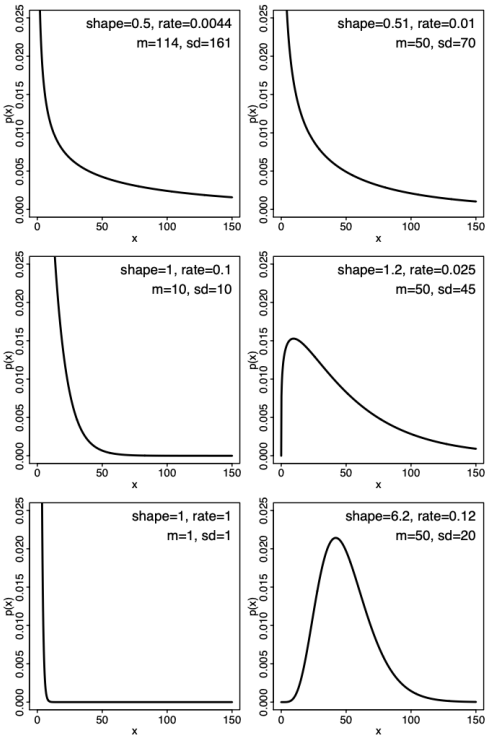


Figure 9.8: Examples of the gamma density distribution. The gamma$(x|s,r)$ *distribution* is a probability density for $x \geq 0$, given by gamma$(x|s,r) = \frac{r^s}{\Gamma(s)} x^{s-1} e^{-rx}$, where $\Gamma(s)$ is the gamma *function*: $\Gamma(s) = \int_0^\infty dt\, t^{s-1} e^{-t}$. The gamma function is a generalization of the factorial, because for positive integers, $\Gamma(s) = (s-1)!$. In the specification of the distribution, s is called the "shape" parameter and r is called the "rate" (or "inverse scale") parameter. The mean of the gamma distribution is $m = s/r$, and the standard deviation of the gamma distribution is $sd = \sqrt{s}/r$. Hence $s = m^2/sd^2$ and $r = m/sd^2$. In R, the gamma density is provided by `dgamma(x,shape=s,rate=r)`, and the gamma function is provided by `gamma(s)`. Conveniently, BUGS parameterizes the gamma distribution the same way as R, i.e., with shape and rate parameters in that order.

9.2.2.1 Doing It with BUGS

Here is a BUGS model specification that corresponding to Figure 9.7

```
model {
    # Likelihood:
    for ( t in 1:nTrialTotal ) {
        y[t] ~ dbern( theta[ coin[ t ] ] )
    }
    # Prior:
    for ( j in 1:nCoins ) {
        theta[j] ~ dbeta( a , b )I(0.0001,0.9999)
    }
    a <- mu * kappa
    b <- ( 1.0 - mu ) * kappa
    mu ~ dbeta( Amu , Bmu )
    kappa ~ dgamma( Skappa , Rkappa )
    Amu <- 2.0
    Bmu <- 2.0
    Skappa <- pow(10,2)/pow(10,2)
    Rkappa <- 10/pow(10,2)
}
```

The BUGS model specification used for loops that repeat the dependencies for each flip of each coin. The for loops implement the ellipsis symbols next to the arrows in Figure 9.7. The loop for the θ_j values.

In Figure 9.5, we assumed that the dependency of θ on μ had a fixed value, namely K = 5. We will capture that assumption in the present BUGS program by making the prior on κ be a very narrow spike over $\kappa = 5$. This is achieved by setting the mean of its gamma distribution to be 5.0, and the standard deviation of its gamma distribution to be 0.01. The corresponding shape and rate parameter values can be computed as shown in the caption of Figure 9.8.

Figure 9.10 shows what happens when the prior on κ restricts it to values extremely close to 75.0.

Now that we are convinced that BUGS is performing properly, we consider some new cases. These continue to be "toy" examples, intended to train our intuition about how Bayesian inference works for this hierarchical prior.

9.2.3 Outliers and Shrinkage of Individual Estimates

When estimating a bias in an individual coin, the estimate can be affected by the results of the other coins, if there is a belief that the coins depend on a shared mint parameter μ. If many coins yield similar data proportions, then the posterior estimate of the dependence, κ, will tend to be high, and that dependence in turn will tend to yield estimates of the individual biases that more closely resemble the mint parameter μ. This "shrinkage" of individual estimates toward the hyperparameter value is especially evident for outliers.

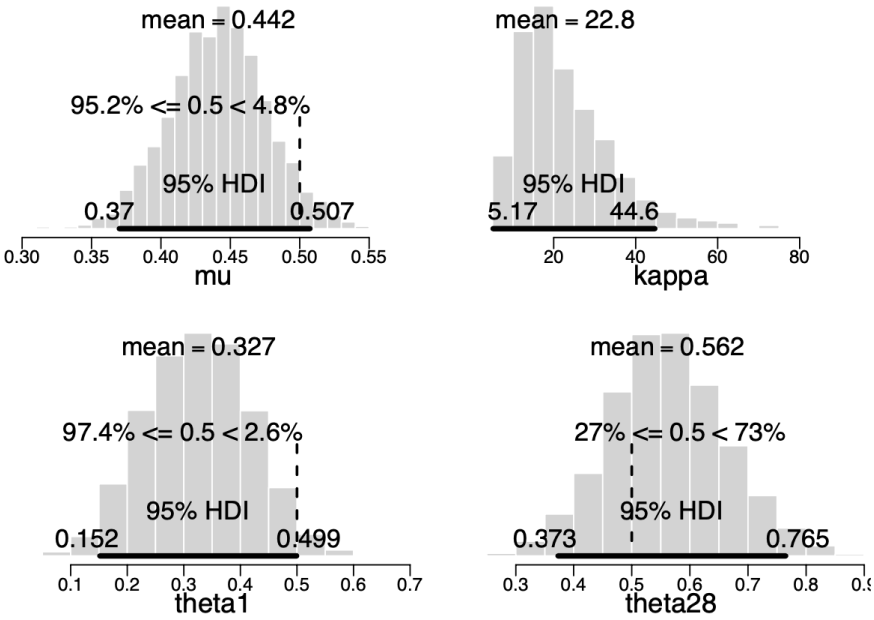


Figure 9.14: Posterior distribution for data from Rosa et al. (1998). Prior assumes $\mu \sim \text{beta}(\mu|1,1)$ and $\kappa \sim \text{gamma}(\kappa|1.0, 0.1)$, i.e., a prior κ mean of 10.0 and prior κ standard deviation of 10.0.

This shrinkage of the estimates of the individual biases is not a problem with the analysis; in fact, the shrinkage accurately reflects our beliefs: If we believe that all the coins have biases generated by the same mint, and that the biases depend on the mint, then the flips from the different coins should mutually inform each other's estimated biases. The hierarchical Bayesian analysis is especially nice because it tells us not only the estimates of the biases (θ_j) and the mint parameter (μ), but also the degree of dependence (κ) of the biases on the mint parameter.

9.2.4 Case Study: Therapeutic Touch

Some scholars investigated therapeutic touch (TT) among medical practitioners. TT is a technique to heal patients' bodies.

The crucial prerequisite for TT is sensing of the patient's energy field.

There were 28 TT practitioners who volunteered to participate (7 were repeated with a several-month separation, so these are counted as distinct subjects). Some authors suggest that the recruitment rate was aided by the fact that the experimenter was a 9-year old girl (the second author of the article). Results showed that the mean number correct, across practitioners, was 4.39, with the lowest being 1 and the highest being 8. Chance performance is 5 out of 10 correct.

These data are precisely of the form that can be modeled by Figure 9.7. Each practitioner corresponds to a "coin" being flipped 10 times, and the underlying ability of the jth practitioner is denoted θ_j. The practitioners are assumed to be randomly representative of the group of all practitioners, and the group has a mean ability denoted by μ. The dependency of the individual abilities on the group mean is measured by κ.

Figure 9.14 shows various marginal distributions of the 30-dimensional joint posterior. The posterior on μ (upper left panel) indicates that the chance value of $\mu = .5$ is among the 95% most believable values. The posterior indicates that the most believable group accuracies actually tend to be less than chance correct,

meaning that the therapists, if sensing something from the experimenter's hand, were systematically selecting the side opposite where the experimenter's hand was.

9.2.5 Number of Coins and Flips per Coin

When we collect more data, our estimate of the model parameters becomes more certain. For example, in the previous section's investigation of therapeutic touch (TT), we know that the sensitivity of the experiment could have been larger if more data were collected. Indeed, if enough data were collected, maybe we could conclude that TT practitioners can, on average, sense the experimenter's hand, albeit only weakly.

This shows that we have two ways to get more data: We could include more flips per coin or include more coins. If we have a choice of including more coins or including more flips per coin, which should we choose? If our goal is to estimate the hyperparameters, then the answer is: More coins.

the larger number of coins puts more constraint on the posterior estimate of μ and κ than the fewer number of coins. The individual coins will have estimates of θ_j that are less specific and more influenced by the other coins, but this is appropriate for the premise of our model: We are presuming that each coin is an independent representative of the same mint.