

1 Assigning labels in simulation studies

K-component Mixture problem where the labels 1,2,...,K are not identifiable without parameter restrictions.

Parameters $(\pi, \lambda_1, \lambda_2)$ give same distribution as $(1 - \pi, \lambda_2, \lambda_1)$. Notation: given the parameter vector

$$\theta = \left[\left(\begin{array}{c} \pi_1 \\ \lambda_1 \end{array} \right), \dots, \left(\begin{array}{c} \pi_K \\ \lambda_K \end{array} \right) \right]$$

let

$$\theta_\sigma = \left[\left(\begin{array}{c} \pi_{\sigma_1} \\ \lambda_{\sigma_1} \end{array} \right), \dots, \left(\begin{array}{c} \pi_{\sigma_K} \\ \lambda_{\sigma_K} \end{array} \right) \right]$$

be the parameter vector obtained under a permutation σ of the integers 1, ..., K. The point is that θ and θ_σ give the same density and so are indistinguishable unless we make further restrictions. In the following we will let θ^τ be the hypothetical true parameter, where for concreteness the labels 1, 2, ..., K on the parameters are given some fixed meaning, say by dictionary ordering of the λ_j . Here dictionary ordering will mean ordering based on the first coordinate of the vector, unless it is a tie, in which case we order on the second coordinate, and so forth. We do this ordering so that there is no ambiguity about what is meant by θ_σ .

We will let $\hat{\theta}$ be an estimator (or a Bayes simulation value $\tilde{\theta}$) in which the labels have no intrinsic meaning. These might be the slots 1,2,...,K in the computer program that provided the estimates. We will call these labels "meaningless", as any permutation of them gives the same likelihood (or Bayes posterior density). The labelling problem concerns how we might choose a permutation σ so that we can view $\hat{\theta}_\sigma$ as the estimator of θ^τ . We could use the same ordering scheme by which we defined θ^τ , but there are good reasons to be cautious of this that will be discussed later. That is, for now we leave the model redundant in the θ parameters and consider various resolutions of the labelling problem.

1.1 Label free standard errors?

Recall that we can obtain MLE's for the parameters, and that there is a method for constructing asymptotic standard errors for each π and λ that does not seem to involve permutation redundancy. That is, no matter how $\hat{\theta}$ is ordered in the computer output, we simply estimate the standard errors using the corresponding element of the inverse of the information matrix J, or the observed information, where the entry being used is determined using the same (meaningless) labels. We might then just interpret the standard error attached to $\hat{\lambda}_1$ as measuring the error that $\hat{\lambda}_1$ incurs in estimating one of the parameters $\lambda_1, \dots, \lambda_K$ of θ^τ without being specific about which one. That is, even though we have parameter redundancy, we seem to have the standard errors we want. And there must be some logic in which they are correct, in the sense of approximating the actual finite sample standard errors.

However, the meaning of "standard error" in a fixed finite sample, when the parameters are not identified, is not so easy. Consider a simulation study, for fixed n , designed to investigate accuracy of the purported asymptotic standard errors. Compute $\hat{\theta}^s$ for $s = 1, \dots, S$ from the model based on some true value θ^τ . To calculate the squared errors over the simulation one needs to put labels (i.e., pick a permutation) on the $\hat{\theta}^s$ so as to match components with θ^τ . In the underidentified θ context, all of the $\hat{\theta}_\sigma$ are equally likely, and so one cannot distinguish among them in that fashion. How can we say which one is estimating θ^τ ? To make this concrete, how are we going to estimate the mean squared error $E(\lambda_1 - \hat{\lambda}_1)^2$ when permutations of the labels on the simulated values would yield different $\hat{\lambda}_1^s$?

Thus the asymptotic and the finite sample points of view seem to be in conflict. We will soon argue that it might be reasonable to appeal to the asymptotics to resolve the labelling problem in the simulation study. That is, there is a sense in which the parameter estimates are asymptotically linked to true values by their consistency, and so can be labelled.

A Bayesian faces an even more fundamental problem. We now imagine a Bayes simulation experiment in which we get a sequence of simulated values $\tilde{\theta}^s$, $s=1, \dots, S$ from the posterior distribution of θ given $Y = y$. In ordinary circumstances one would get Bayes estimates θ_B by calculation of the averages $\sum \tilde{\theta}^s / S$, and use the posterior variance as a measure of uncertainty. However, if there are no identifying restrictions into the

prior, all the posterior means $E[\tilde{\lambda}_j|Y = y]$ are identical to the the permutation invariance of the posterior distribution. This is because $\tilde{\theta}^s$ comes without meaningful labels. Thus in this case, we even have to make sense of the term "Bayes estimator."

There are multiple possible resolutions to this problem because there are a number of reasonable ways for one to assess experimental error. Perhaps the clearest way to see this linkage between error assessment and labelling is make explicit the role of loss functions in various proposed solutions. We start our discussion by considering explicit parameter constraints. Based on our critique of them, we will then offer as alternatives risk-based labels.

1.2 Explicit parameter constraints and their implied risk

This will discuss ordering restrictions on the parameters, and the desirability of using them to estimate parameters and evaluate risk. **(to be completed)**

Strategy. Assign meaningful labels to θ by using the values of λ 's or π to assign meaningful labels to each component. The simplest example of this might be left-right ordering. If λ is a scalar parameter, we can set $\lambda_1 < \lambda_2 < \dots < \lambda_K$. If we have done so, we could also take our unlabelled estimator $\hat{\theta}^s$ and label it based on the values of $\hat{\lambda}$, so that $\hat{\lambda}_{(1)}$, the first order statistic, estimates λ_1 , and so forth. In the process, the π 's become labelled, and we use $\hat{\pi}_{(j)}$ to estimate π_j . Returning to the original estimators $\hat{\theta}^s$, we see that we have chosen a permutation (σ) based on the data, and are using $\hat{\theta}_{(\sigma)}^s$ to estimate θ^r . Corresponding to this estimation there is a loss function $\|\theta^r - \hat{\theta}_{(\sigma)}^s\|^2$ that describes our overall error in estimation.

1.2.1 The latent experiment as idea generator

However, is this loss appropriate? Suppose that there are $K=2$ components in which case we would always use $\hat{\lambda}_{(1)}$ to estimate λ_1 , the smaller λ . To make the issues concrete, let us suppose that the smaller λ is biologically meaningful in that we have two known groups, and one is virtually certain to have a larger parameter value. For example we might have a sample of heights from the Naval Academy, but lack gender identity. Suppose the model is $\pi N(\lambda_1, \sigma^2) + (1 - \pi)N(\lambda_2, \sigma^2)$, $\lambda_1 < \lambda_2$. Since females should be shorter and so correspond to λ_1 , we might assume that $\lambda_{(1)}$ is estimating the female heights. Such a scheme would certainly be consistent.

However such a reporting of risk seems optimistic if we compare it with what we would do if we saw the male/female labels \tilde{J} . To illustrate, suppose that we could observe the Y, \tilde{J} data, where we see the labels of the observations from \tilde{J} , and so we could construct by maximum likelihood estimates $\pi^*, \lambda_1^*, \lambda_2^*$. These estimates now have meaningful labels male/female that we can match with the true values because in this experiment we saw the two groups. Suppose one went ahead with a simulation and applied the ordering based labelling scheme to these estimates as if the labels were meaningless. If one did so, whenever the original ordering of $\lambda_1^* < \lambda_2^*$ fails (the males in the sample are shorter on average), the relabelled estimates are pushed closer to the true values, and so have less squared error than the original "correct" ordering. We should also note that if π_1 is small, as in our example, λ_1^* will be considerably more variable than λ_2^* due to its small sample size.

Moreover, since we also know that at the Naval Academy the males should be more frequent than females, so $\pi_2 > \pi_1$, maybe we should call the group with the larger $\hat{\pi}$ value the male group. This would seem to make as much sense, but would give a different version of risk in small samples.

TBC: Dictionary ordering. For Vector λ , as left right impossible. Gives identifiable restrictions for the parameter space. Problem: in continuous data, labels determined by first coordinate only because there are no ties. And if we switch the coordinates, only some other coordinate matters.

Risk: Corresponding estimators would flip labels greatly depending on coordinate choice, and if true mixture has λ_1, λ_2 equal in first coordinate and not in others, asymptotic labelling error never goes away, as the first coordinate just represents noise around the common value (If the first coordinates are close, this would happen until sample size is very large.)

pi-based ordering. Group 1 is one with largest π , Group 2 has second largest, and so forth. Group label not identifiable if any π 's are equal. Could be physically meaningful as above example. Risk on π is order statistic based so optimistic compared to latent experiment. Bigger risk problem: if true π is .5, then labels of the two groups will oscillate meaninglessly due to sampling error, consistency fails. When true π is near .5, we can expect large overlap in λ_1 and λ_2 distributions due to sampling error (crossing .5) in the π estimator.

1.3 Risk in other identifiable parameterizations

Now suppose that we have made the model's parameters identifiable in some other way, by using the distribution function Q_K itself, or some moment parameterization. Now risk should only be defined on the new identifiable parameters. Ie, in a label free manner. One might find this desirable alternative to using identifying restrictions for the reasons discussed above.

However, even in this case, asymptotic results suggest that it still might be meaningful to talk about the errors made in the redundant θ parameter. Moreover, we think that in many practical situations, one is directly interested in the component distributions, and so might want to create a reasonable method for describing the component parameter errors. We will link our methods with loss functions $L(\theta^\tau, \hat{\theta})$ which are label free, where by label-free we mean invariant under permutations on the two arguments, as a way of interpreting the results.

1.3.1 Consistent labelling schemes

Logic: Suppose that $\hat{Q}_{K,n}$ converges to Q_K^τ as n goes to infinity. Need distance measure here. For the right distances, we can also conclude that there must exist "consistent labelling" schemes. That is, there exist schemes (methods) for creating a sequence σ_n of permutations, possibly depending on both $\hat{\theta}_n$ and θ^τ , such that $\hat{\theta}_{n,\sigma_n}$ converges to θ^τ in probability. We will call such a labelling scheme *consistent*. If the labelling scheme also satisfies

$$n^{1/2} J_\tau^{1/2}(\hat{\theta}_{\sigma_n} - \theta^\tau) \rightarrow N(0, I),$$

where J_τ is the unit information matrix corresponding to θ^τ , then we will say that the scheme is *efficient*. To date we have discussed schemes in which σ_n is based only on $\hat{\theta}$, and is determined by some ordering constraints on the parameters. In the examples we discussed, there exist parameter values for which the efficiency result above would fail, as when there are ties in the pi's or equalities in the first coordinate in the dictionary ordering. Thus they are not globally efficient.

How would one determine whether a particular labelling scheme is efficient asymptotically? If we used the wrong labels, and so were converging to the wrong ordering of θ^τ , then we should expect to find that the mislabelled versions will eventually be very unlikely in the above normal limiting distribution, but the right permutation should have a "typical" density.

This suggests a labelling strategy. Let us choose the permutation $\hat{\sigma}$ by picking the one with the highest normal density. To operationalize this, note that for a given permutation σ the exponent in the normal density is a decreasing function of:

$$SS(\sigma) = (\hat{\theta}_\sigma - \theta^\tau)' J_\tau (\hat{\theta}_\sigma - \theta^\tau).$$

Therefore the $\hat{\sigma}$ that maximizes over the normal densities must minimize $SS(\sigma)$. Now this assignment of labels depends on the true values, so we should write $\hat{\sigma}(\hat{\theta}, \theta^\tau)$. However θ^τ would be known in the simulation study, as would J_τ .

1.3.2 Risk based labels

Note that we can think of the selected $\hat{\sigma} = \hat{\sigma}(\hat{\theta}, \theta^\tau)$ as the minimizing argument of an implicit loss function

$$L(\theta^\tau, \hat{\theta}) = \min_{\sigma} SS(\sigma)$$

that does not itself depend on defined labels—that is, it is permutation invariant. Note also that for each fixed n , using this labelling scheme will minimize

$$tr[Var_{\tau}(J^{1/2}(\hat{\theta}_{\sigma} - \theta^{\tau}))]. \tag{1}$$

over all possible choices of labelling schemes σ , and the minimum risk will be

$$E[L(\theta^{\tau}, \hat{\theta})].$$

We will therefore say that such a labelling scheme is *minimum risk based*. To generalize, one could replace $SS(\sigma)$ above with another permutation dependent loss function, consider the minimum over the permutations, and so obtain other *risk based labels*.

Is this a reasonable thing to do? One aspect of this system that one might find troubling is that we are using the asymptotic true values in the risk-based labels, and that might make the estimated errors too optimistic. (In fact, they are more optimistic than the data based labels by (1).)

Part of the solution lies in giving the right interpretation to reported results: When we do the simulation study, and report the errors in this manner, we must report them as the errors that we would make if we were able to make risk optimal assignments of the components.

It is some consolation that asymptotically, the effect of this strategy should be negligible. That is, for large n this strategy should be insensitive to the choice of θ^{τ} in the sense that it could be replaced with any θ' in a neighborhood of the true values without changing most of the labels in the simulation. So we use θ^{τ} only "weakly". It should also not improve the asymptotic standard errors in the sense of making them any smaller than any other efficient scheme.

1.3.3 Latent risk based labels

However, we still might wish to reduce our unwarranted gains from using truth based labels. Let us do so based on the uncertainty in the latent experiment. Again imagine that we could observe the Y, \tilde{J} data, where we see the labels of the observations from \tilde{J} , and so we could construct by maximum likelihood estimates $\pi^*, \lambda_1^*, \lambda_2^*$. As before, these estimates have meaningful labels that match the true values because in this experiment we saw the two groups. Suppose one went ahead with a simulation and our risk based labelling scheme to these estimates as if the labels were meaningless. Once again, whenever the original ordering of $\pi^*, \lambda_1^*, \lambda_2^*$ is altered, the altered estimates are pushed towards the true values, and so have less squared error than the "correct" ordering. Most people would agree that the reported error is optimistic. We therefore consider a second proposal in which we do not use the true values for labelling.

Suppose one were to replace the true values θ^{τ} in the above $SS(\sigma)$ calculations by θ^* , the latent maximum likelihood estimators. In a simulation we could do this if we generate the data by first generating \tilde{J}^s and then $Y^s | \tilde{J}^s$, generating θ_s^* in the process. In this complete data context, the asymptotic distribution of $\theta^* - \hat{\theta}_{\sigma_n}$ would be normal with mean zero and inverse variance matrix $J_{aug} - J$, where J_{aug} is the information in the augmented data. Using this in the SS criterion would give an asymptotically optimal way of selecting the labels based on the latent experiment.

Of course, this means that in one's simulation study it would be quite possible to get nearly the same Y , and therefore similar point estimates, but have the labels switched due to the differences that occurred in the J data. In fact, for any fixed $Y=y$ there is a probability $\alpha(y)$ that the label agrees with the truth-based label, where $\alpha(y)$ is determined by the distribution under θ^{τ} of \tilde{J} given $Y = y$.

Of course, we are still cheating a bit, in that in our real data we do not have θ^* any more than we have θ^{τ} . But this mechanism would eliminate a source of optimism in that the gain in labelling does not reduce the underlying errors of θ^* makes as an estimate of θ . TBC:

- This should correspond to some sort of optimal risk based labelling where now the labels are not allowed to depend on θ^{τ} , but are allowed to depend on θ^* .
- Implementation difficulties? Computing J could be done numerically at the beginning of the study.
- Implication for bootstrap studies.

2 Bayes version

Consider an unlabelled Bayes prior B on θ , by which we mean one which gives same likelihood to all values of θ_σ as σ varies. Then, in the posterior, the parameters have a distribution that is symmetric over permutations of labels as well.

In such an experiment, the variable labels are not distinguishable. That is, there is nothing in the data that can tell us which labeling is more likely. And the values of $E\{\lambda_1|y\}$ and $E\{\lambda_2|y\}$ are identical. In such a setting, we might try to estimate, without reference to labels, the (unlabelled) set of true values $\{\lambda_1, \lambda_2\}$ by minimizing a loss function that does not depend on labels. The above arguments suggest the loss function

$$L(\tilde{\theta}^\tau, \theta_b) = \min_{\sigma} SS(\sigma)$$

where θ_b is the estimator, $\tilde{\theta}^\tau$ is the value of θ realized in the current experiment, and

$$SS(\sigma) = (\tilde{\theta}_\sigma - \theta_b)' J_b(\tilde{\theta}_\sigma - \theta_b).$$

This $L(\tilde{\theta}^\tau, \theta_b)$ defines a label free loss function for estimators θ_b . That is, the loss function is constant over permutations of θ_b elements. From this one can define the Bayes risk: $E_B[L(\tilde{\theta}^\tau, \theta_b)]$ and hence the Bayes estimator as the minimizer θ_B of $E_B[L(\tilde{\theta}^\tau, \theta_b)|Y = y]$ over θ_b . These are label free Bayes estimators, in that every permutation of the elements results in the same value of the objective function.

Assume now we wish to carry this out in a Bayes simulation. Algorithm? Let us write the permutation that achieves the minimum in $SS(\sigma) = (\tilde{\theta}_\sigma^s - \theta_b)' J_b(\tilde{\theta}_\sigma^s - \theta_b)$ as $\sigma_s^* = \sigma(\theta_b, s)$. We should then minimize

$$\sum_s (\hat{\theta}_{\sigma_s^*}^s - \theta_b)' J(\hat{\theta}_{\sigma_s^*}^s - \theta_b)$$

to find the Bayes estimates. This would be an ordinary weighted least squares problem except for the dependence of σ_s^* on θ_b . Thus there might be some hope of using iteratively recentered least squares, in which we start with some initial θ_1 , determine $\sigma_{s,1}^* = \sigma(\theta_1, s)$, and then solve

$$\sum_s (\hat{\theta}_{\sigma_{s,1}^*}^s - \theta_b) = 0$$

for θ_2 . (Note that J_b drops out in this step because it is the same for all s , but that J_b still plays a role in determining the labels.) The new parameter estimate θ_2 gives new best permutations $\sigma_{s,2}^*$, and we can repeat the least squares to get a new θ_3 .

TBC. If the algorithm is unstable, one could shorten the step length to ensure monotonicity of the objective function.

TBC. This problem has the potential to have local minima, as it is certainly not convex, in which case the above algorithm is not sufficient to reach the global minimum.

Note that when one is done, one also has the optimal sigmas σ^* that give implied labels to the simulated θ , and that the Bayes estimators are then averages based those labels. It might seem at first as if the Bayes approach might have eliminated a need to worry about labelling optimism. However, the problem still exists if we try to interpret the posterior variance of the newly labelled simulation elements as representing our true uncertainty.

TBC. The question is: does the variability of $\tilde{\theta}_\sigma^s$ reflect what we mean by uncertainty? Once again, let us refer to the latent experiment (\tilde{J}, Y) . Relative to it, there is optimistic labelling occurring, and so we might want to remove some of this optimism by creating a modified loss function.

Suppose we could observe \tilde{J} and Y together. We might achieve this in a simulation by simulating \tilde{J}^s together with $\tilde{\theta}^s$. Treating \tilde{J}^s, Y as the realized data, one could construct a labelled posterior mean $E[\tilde{\theta}|J^s, Y]$ for the parameters. Notice that in an identified model the Bayes risk would have the decomposition

$$E[\theta - E[\theta|Y]]^2 = E[\theta - E[\theta|Y, J]]^2 + E[E[\theta|Y] - E[\theta|Y, J]]^2$$

In our problem the first term on the right side can be calculated as we have the labels for θ , but the second not, as Y contains no label information.

This suggests a latent labelling scheme where one set the labels based on minimizing the second term, over permutations, where $E[\theta|y]$ is replaced with θ_b , and we now minimize over labels on the $\tilde{\theta}_s^* = E[\tilde{\theta}|J^s, Y]$. That is, we would minimize the risk function:

$$\sum_s \min_{\sigma} (\tilde{\theta}_{s,\sigma}^* - \theta_b)' W (\tilde{\theta}_{s,\sigma}^* - \theta_b)$$

We need a weight matrix W in the SS criterion, and I suspect the correct one is again $J_{aug} - J$. As in the frequentist case, if the latent means are "out of order" compared with the realization $\tilde{\theta}^s$, it will be reflected in our Bayes estimator through its greater uncertainty. To put this in terms of risk, in this way we can keep relabelling from reducing the risk inherent in the first term $E[\theta - E[\theta|Y, J]]^2$.