10 Fast Estimation without Markov Chains

In chapter 9, we focused on the Gibbs sampling algorithm, because our goal was to compute the mean of the posterior for $\boldsymbol{\beta}$ (equation 9.2, page 161). However, given the posterior distribution $\mathcal{P}(\boldsymbol{\beta}, \sigma, \theta \mid m)$ in equation 9.11, we could alternatively use its maximum as a point estimate, the so-called Maximum A Posteriori (MAP) estimator.

Thus, from the (marginal) posterior distribution of the coefficients, $\mathcal{P}(\boldsymbol{\beta}|y)$, we could compute:

$$\beta_{\text{MAP}} \equiv \arg\max_{\boldsymbol{\beta}} \mathcal{P}(\boldsymbol{\beta}|y) = \arg\max_{\boldsymbol{\beta}} \int \mathcal{P}(\boldsymbol{\beta}, \sigma, \theta|y)\, d\sigma\, d\theta. \tag{10.1}$$

We show here that we can compute this quantity or approximations to it without Gibbs sampling, as the solution to a maximization problem. Whether this MAP estimator is better than the mean of the posterior, and whether it is easier to implement it may depend on the problem and on the kind of software and expertise available to the user.

In the three main sections of this chapter, we offer three alternative estimators—a maximum a posteriori (MAP) estimator, a marginal MAP estimator, and a conditional MAP estimator. In applications, we have used only the conditional MAP estimator to any large extent, but we present all three so that a user interested in implementing any of them will have many of the necessary calculations. All three estimators also provide additional insights about the model. For further information and alternative refinements, see Leonard and Hsu (1999), Greenland (2001), and Gelman et al. (2003).

10.1 Maximum A Posteriori Estimator

Instead of maximizing the marginal posterior for $\boldsymbol{\beta}$, we could alternatively maximize the whole posterior, obtaining estimates for σ and θ as well. This has the advantage of enabling one to compute forecast standard errors easily via simulation. In order to compute this estimator, we have to solve the following maximization problem:

$$\max_{\boldsymbol{\beta},\sigma,\theta} \mathcal{P}(\boldsymbol{\beta}, \sigma, \theta|y). \tag{10.2}$$

It is now straightforward to write the first order conditions for this problem:

$$\boldsymbol{\beta}_k = \left(\frac{\mathbf{X}_k' \mathbf{X}_k}{\sigma_k^2} + \theta s_k^+ \mathbf{C}_{kk} \right)^{-1} \left(\frac{\mathbf{X}_k' y_k}{\sigma_k^2} + \theta \sum_j s_{jk} \mathbf{C}_{kj} \boldsymbol{\beta}_j \right),$$

$$\sigma_k^{-2} = \frac{\mathbb{d} + T_k - 2}{\mathbb{e} + \mathrm{SSE}_k},$$

$$\theta = \frac{\mathbb{f} + r - 2}{\mathbb{g} + \sum_{ij} W_{ij} \boldsymbol{\beta}_i' \mathbf{C}_{ij} \boldsymbol{\beta}_j}.$$

Thus, we have written the equation for the coefficient $\boldsymbol{\beta}_k$ in such a way that $\boldsymbol{\beta}_k$ is not on the right side (because $s_{kk} = 0$). This suggests a simple iterative algorithm in which one starts, for example, with $\boldsymbol{\beta}$ and σ obtained via equation-by-equation least squares and then updates the estimates for all three parameters according to the preceding expressions until convergence.

10.2 Marginal Maximum A Posteriori Estimator

We begin by factoring the marginal posterior into the product of two terms:

$$\mathcal{P}(\boldsymbol{\beta}|y) \propto \left[\int d\sigma \mathcal{P}(m|\boldsymbol{\beta}, \sigma) \mathcal{P}(\sigma) \right] \left[\int d\theta \mathcal{P}(\boldsymbol{\beta}|\theta) \mathcal{P}(\theta) \right].$$

The first term is often called an "effective" likelihood, that is, a likelihood in which we already have incorporated the effect of the uncertainty about σ by integrating it out. Similarly, the second term can be thought of as an "effective" or marginal prior, which also takes into account the fact that the parameters of the prior are known with uncertainty. The integrals in the preceding expressions can be readily computed:

$$\int d\sigma \mathcal{P}(m|\boldsymbol{\beta}, \sigma) \mathcal{P}(\sigma) \propto \prod_i \left(\frac{1}{\mathbb{e} + \mathrm{SSE}_i(\boldsymbol{\beta})} \right)^{\frac{\mathbb{d}+T_i}{2}},$$

where we have defined SSE_i in equation 9.12 (page 165). Similarly, we have

$$\int d\theta \mathcal{P}(\boldsymbol{\beta}|\theta) \mathcal{P}(\theta) \propto \left(\frac{1}{\mathbb{g} + H[\boldsymbol{\beta}]} \right)^{\frac{\mathbb{f}+r}{2}},$$

where

$$H[\boldsymbol{\beta}] \equiv \sum_{ij} W_{ij} \boldsymbol{\beta}_i' \mathbf{C}_{ij} \boldsymbol{\beta}_j.$$

Finally, we plug the preceding expressions in equation 10.1 and take its log, which gives the marginal MAP estimator:

$$\beta_{\text{mMAP}} \equiv \arg\min_{\boldsymbol{\beta}} \sum_i (\mathbb{d} + T_i) \ln\left(1 + \frac{\text{SSE}_i(\boldsymbol{\beta})}{\mathbb{e}}\right) + (\mathbb{f} + r) \ln\left(1 + \frac{H[\boldsymbol{\beta}]}{\mathbb{g}}\right). \quad (10.3)$$

10.3 Conditional Maximum A Posteriori Estimator

We now compare equation 10.3 with the one we would obtain if we had conditioned on the fixed values of the parameters σ_i and θ (corresponding to degenerate, point mass priors):

$$\beta_{cMAP} \equiv \arg\min_{\boldsymbol{\beta}} \sum_i \frac{1}{\sigma_i^2} \text{SSE}_i(\boldsymbol{\beta}) + \theta H(\boldsymbol{\beta}). \quad (10.4)$$

By comparing equations 10.4 and 10.3, we first observe that the effect of uncertainty on σ_i and θ is to replace the quadratic cost functions in equation 10.4 with the concave-shaped cost functions of equation 10.3. Functions other than the logarithm could be obtained (e.g., the absolute value) by choosing different densities for σ and θ (see Girosi, 1991, for a characterization of the class of functions that can be obtained in this way).

The effect of having concave functions in the likelihood is to provide some robustness against variation in squared error over the cross sections. If one cross section has large squared error, and we use the estimator of equation 10.4, we may end up with an estimator too focused on making the error in that particular cross section small, at the expense of the squared error in the other cross sections. Using equation 10.3 instead will prevent such outliers from having undue effects.

When $\mathbb{e}$ is "large" with respect to SSE_k, we have $\ln(1 + \frac{\text{SSE}_i(\boldsymbol{\beta})}{\mathbb{e}}) \approx \frac{\text{SSE}_i(\boldsymbol{\beta})}{\mathbb{e}}$. This makes sense, because we can make $\mathbb{e}$ large, for example, by choosing a prior distribution for σ_i^{-2} with a small variance, which is closer to the case considered in equation 10.4 (with 0 variance). Similarly, if our prior for θ has a small variance, so that we are fairly certain about the value of θ, the prior term in equation 10.3 will tend to the corresponding term in equation 10.4.

The role of the quantities in equation 10.3—$\mathbb{d}$, $\mathbb{f}$, r, and T_i—is also clear. The parameters $\mathbb{d}$ and $\mathbb{f}$ control the concentration of the densities for σ_i^{-2} and θ around their means. Therefore, we expect them to appear as weights in equation 10.3. The number T_i is the number of observations in cross section i, and it appears as a weight in the likelihood so that cross sections with more observations are smoothed relatively less. The number r is the rank of the inverse covariance matrix in the prior for $\boldsymbol{\beta}$, and it reflects the amount of information carried by the prior (if r is small, then the prior is constant on most of its domain). Notice that the counterpart of r in the likelihood is T_i, which measures how much information is in the likelihood.

Equation 10.3 provides an alternative estimator to the posterior mean. It does not require Gibbs sampling, but it requires minimizing a nonconvex cost function. When σ_i^{-2} and θ have small variance, and therefore equation 10.3 approaches equation 10.4, it might be simpler to solve equation 10.3 than to use the Gibbs sampler. However, because the

objective function is not convex, it is difficult to exclude the possibility of local minima, which can lead to suboptimal solutions. On the other hand, Gibbs sampling algorithms are guaranteed only to converge asymptotically and so, as is usual with MCMC techniques, it is never entirely clear when they have converged. Another disadvantage of this approach is that, unlike the Gibbs sampling, it does not provide model-based standard errors, because it does not provide estimates for σ and θ.

10.4 Summary

This chapter offers three fast algorithms for estimating the mode of the posterior. Whether these point estimators or those in chapter 9 are preferable in any one application is an empirical question. The estimators offered here are faster and also require less expertise of users than the MCMC algorithm in chapter 9.