5 Bayesian Estimation: Linear Regressions & Beyond

The previous two chapters dealt with frequentist/classical estimation methods of models under cross-sectional data, with a particular emphasis on the linear regression model. This chapter discusses Bayesian inference of linear regressions and more general models. Like ML estimation (Chapter 4), it is based on the likelihood $p(Y|\theta)$. In contrast to ML estimation, under the Bayesian paradigm, we treat data $Y = \{(y_i)'\}_{i=1}^n$ as fixed and the unknown parameter θ as random; given a distribution $p(\theta)$ that represents our knowledge on where θ lies prior to observing the data, we update this knowledge based on the likelihood – i.e. the data interpreted through the lens of a particular model, e.g. the linear regression model –, which results in the posterior distribution $p(\theta|Y)$.

This chapter is set out as follows. First, Section 5.1 discusses Bayesian estimation of the linear regression model, which includes Lasso and Ridge estimation. In turn, Section 5.2 explains how Bayesian inference can be implemented for more general models (likelihood and prior specifications), and Section 5.3 discusses model selection.

5.1 Bayesian Analysis of the Linear Regression Model

This section discusses Bayesian estimation of the linear regression model. Under appropriate (and commonly used) distributional families for the likelihood and prior, the posterior can be derived analytically. This illustrates some fundamental aspects of Bayesian estimation that apply for more general models (likelihood and prior specifications) – discussed in turn in Section 5.2 – as well.

5.1.1 Estimating $\beta|\sigma^2$

As laid out in Section 4.1 above, the linear regression model under conditional Normality of error terms leads to the likelihood function

$$p(Y|\beta,\sigma^2) = (2\pi\sigma^2)^{-n/2} exp\left\{-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right\} .$$

For now, assume we are interested only in estimating β and take σ^2 as given. Suppose our prior is Normal as well: $\beta|\sigma^2 \sim N(\underline{\beta}, \sigma^2\underline{V})$.[1] For simplicity, the conditioning on σ^2 is dropped from the following expressions. The posterior of β is then[2]

$$\begin{aligned}
p(\beta|Y) &\propto p(Y|\beta)p(\beta) \\
&\propto exp\left\{-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right\} exp\left\{-\frac{1}{2\sigma^2}(\beta-\underline{\beta})'\underline{V}^{-1}(\beta-\underline{\beta})\right\} \\
&\propto exp\left\{-\frac{1}{2\sigma^2}\left[-\beta'X'Y - Y'X\beta + \beta'X'X\beta + \beta'\underline{V}^{-1}\beta - \beta'\underline{V}^{-1}\underline{\beta} - \underline{\beta}'\underline{V}^{-1}\beta\right]\right\} \\
&\propto exp\left\{-\frac{1}{2\sigma^2}\left[\beta'[X'X+\underline{V}^{-1}]\beta + -2(Y'X+\underline{\beta}'\underline{V}^{-1})\beta\right]\right\} .
\end{aligned}$$

This lets us deduce that $\beta|Y \sim N(\bar{\beta}, \sigma^2\bar{V})$ with $\bar{V} = [\underline{V}^{-1} + X'X]^{-1}$ and

$$\bar{\beta} = \bar{V}[X'Y + \underline{V}^{-1}\underline{\beta}] = [\underline{V}^{-1} + X'X]^{-1}[X'X\hat{\beta}_{ML} + \underline{V}^{-1}\underline{\beta}] .^3$$

As in the simple example in Section 2.2.1, the posterior mean is a weighted average of the ML estimator and the prior mean, with weights given by their respective inverse covariance matrices $\sigma^{-2}X'X$ and $\sigma^{-2}\underline{V}^{-1}$, though σ^{-2} cancels out. The more information there is in the likelihood function – i.e. the smaller its variance $\sigma^2(X'X)^{-1}$ – the higher the weight on the ML estimator and the closer the posterior mean is to the ML estimator. Also, the posterior variance is no larger than the variance of the ML estimator, $\sigma^2(X'X)^{-1}$.

Loosely speaking, asymptotically, Bayesian estimation corresponds to ML estimation. In

[1] Conditioning the prior on σ^2 is natural because the meaning of size of coefficients is only obtained given σ^2 (it determines the scale of data).

[2] More precisely, this is the conditional posterior of $\beta|\sigma^2$.

[3] To see this more explicitly, write out the pdf under our "guess" $\beta|Y \sim N(\bar{\beta}, \sigma^2\bar{V})$:

$$p(\beta|Y,\sigma^2) \propto exp\left\{-\frac{1}{2\sigma^2}(\beta-\bar{\beta})'\bar{V}^{-1}(\beta-\bar{\beta})\right\} \propto exp\left\{-\frac{1}{2\sigma^2}\left[\beta'\bar{V}^{-1}\beta - 2\bar{\beta}'\bar{V}^{-1}\beta\right]\right\} .$$

By comparing terms with the above expression for $p(\beta|Y,\sigma^2)$, we can see that $\bar{V}^{-1} = \underline{V}^{-1} + X'X$ and $\bar{V}^{-1}\bar{\beta} = X'Y + \underline{V}^{-1}\underline{\beta}$, which leads to the expressions above.

other words, the influence of the prior vanishes as $n \to \infty$. To see this, first note that $X'X = n\hat{Q}$, with our previously defined $\hat{Q} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i'$. Second, recall that the finite sample distribution of $\hat{\beta}|\beta, X$ is $N(\beta, \frac{\sigma^2}{n}\hat{Q}^{-1})$ and, therefore, $\sqrt{n}(\hat{\beta}-\beta)|\beta, X \sim N(0, \sigma^2\hat{Q}^{-1})$. The posterior mean $\bar{\beta}$ converges to the ML estimator $\hat{\beta}$ as the sample size gets larger:

$$(\bar{\beta} - \hat{\beta}) = \left[\hat{Q} + n^{-1}\underline{V}^{-1}\right]^{-1}\left[\hat{Q}\hat{\beta} - n^{-1}\underline{V}^{-1}\underline{\beta}\right] - \hat{\beta} \to 0 \quad \text{as } n \to \infty \, .$$

Also, the posterior variance $\sigma^2\bar{V}$, scaled up by n, converges to the finite-sample variance of the ML estimator, $\mathbb{V}[\hat{\beta}|X] = \frac{\sigma^2}{n}\hat{Q}^{-1}$, scaled up by n:

$$n\bar{V} - \hat{Q}^{-1} = \left[\hat{Q} + n^{-1}\underline{V}^{-1}\right]^{-1} - \hat{Q}^{-1} \to 0 \quad \text{as } n \to \infty \, .$$

Overall, this means that the posterior of β, standardized by $\sqrt{n}$, coincides asymptotically with the finite-sample distribution of $\hat{\beta}_{ML}$, standardized by $\sqrt{n}$:

$$\sqrt{n}(\beta - \bar{\beta})|Y \xrightarrow{d} N(0, \sigma^2\hat{Q}^{-1}) \sim \sqrt{n}(\hat{\beta} - \beta)|\beta, X \, .$$

In other words, for large n, we have that $\beta \overset{approx.}{\sim} N(\hat{\beta}, \frac{\sigma^2}{n}\hat{Q}^{-1})$, i.e. the posterior is a Normal distribution centered at $\hat{\beta}$, with variance $\mathbb{V}[\hat{\beta}] = \sigma^2\hat{Q}^{-1}$.[4] Note that under the Bayesian paradigm, β is a RV and data (Y, X) is fixed (for any given n), which means that $\hat{\beta}$ and $\hat{Q}$ are constants, even asymptotically.[5]

Note that, under appropriate choices of $\underline{V}$ (examples follow in Section 5.1.1.1 and Section 5.1.1.2), the posterior mean can be computed even if $X'X$ is singular. In contrast, $\hat{\beta}_{ML}$ is not defined in this case. In particular, the Bayesian approach can deal with the case of $k > n$, i.e. when we have more parameters to estimate than observations,[6] and it is often used when one has many parameters but little observations (big data), a setting in which $\hat{\beta}$ is noisy as $X'X$ is close to singular. Rather than explicitly referring to priors and posteriors, researchers more inclined to the frequentist paradigm would in these cases refer to the point estimators derived from Bayesian procedures as "regularization".

When deriving the posterior, we make use of its proportionality to the product of likelihood

[4] This can be shown even for quite general priors by expanding the posterior mean around $\hat{\beta}$.

[5] In contrast, under frequentist inference, as data is random and β is fixed, we would not let $\hat{Q}$ show up in the distribution that is obtained asymptotically, as $n \to \infty$. Rather, we would say that $N(0, \sigma^2\hat{Q}^{-1}) \xrightarrow{d} N(0, \sigma^2 Q^{-1})$. This is why we condition here on X.

[6] In fact, Bayesian methods work even without any observations. In that case, the prior simply does not get updated, i.e. the posterior equals the prior.

and prior, disregarding the denominator in Bayes formula:

$$p(\beta|Y) = \frac{p(Y|\beta)p(\beta)}{p(Y)} \propto p(Y|\beta)p(\beta) \ .$$

This denominator, $p(Y) = \int p(Y|\beta)p(\beta)d\beta$, is called the marginal likelihood or marginal data density (MDD). It is a weighted average of the likelihood evaluated at all (ex-ante) possible parameter values. As illustrated in Section 5.1.1.1, it is a measure of model fit. It can be obtained in three steps. First, invert Bayes' theorem to yield $p(Y) = \frac{p(Y|\beta)p(\beta)}{p(\beta|Y)}$. Next, insert the known expressions for the likelihood, prior and posterior on the RHS. Finally, simply cancel all β-terms because $p(Y)$ does not depend on β. We get

$$\begin{aligned} p(Y) &= \frac{p(Y|\beta)p(\beta)}{p(\beta|Y)} \\ &= \frac{(2\pi)^{-\frac{n}{2}} exp\left\{-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right\}(2\pi)^{-\frac{k}{2}}|\underline{V}|^{-\frac{1}{2}} exp\left\{-\frac{1}{2\sigma^2}(\beta-\underline{\beta})'\underline{V}^{-1}(\beta-\underline{\beta})\right\}}{(2\pi)^{-\frac{k}{2}}|\bar{V}|^{-\frac{1}{2}} exp\left\{-\frac{1}{2\sigma^2}(\beta-\bar{\beta})'\bar{V}^{-1}(\beta-\bar{\beta})\right\}} \\ &= \frac{(2\pi)^{-\frac{n}{2}} exp\left\{-\frac{1}{2\sigma^2}Y'Y\right\}|\underline{V}|^{-\frac{1}{2}} exp\left\{-\frac{1}{2\sigma^2}\underline{\beta}'\underline{V}^{-1}\underline{\beta}\right\}}{|\bar{V}|^{-\frac{1}{2}} exp\left\{-\frac{1}{2\sigma^2}\bar{\beta}'\bar{V}^{-1}\bar{\beta}\right\}} \ .^{7} \end{aligned}$$

5.1.1.1 Shrinkage (Ridge Regression)

Consider the Normal prior with $\underline{\beta} = 0$ and $\underline{V} = \frac{1}{\lambda}I$. For simplicity, set $\sigma^2 = 1$. Our posterior is then

$$\bar{V} = [X'X + \lambda I]^{-1} \ , \quad \bar{\beta} = [X'X + \lambda I]^{-1}X'X\hat{\beta} \ .$$

This shows that the posterior mean $\bar{\beta}$ shrinks the ML estimator $\hat{\beta}$ towards zero. The higher λ, the more shrinkage is applied. $\bar{\beta}$ is also called the Ridge estimator. It can be obtained under a penalized or "regularized" least squares procedure called the Ridge regression:

$$\bar{\beta} = \arg\min_{\beta} (Y-X\beta)'(Y-X\beta) + \lambda\beta'\beta \ ,$$

[7] Generally, as we conditioned our estimation of β on σ^2, the MDD derived in this section would be conditional on σ^2: $p(Y|\sigma^2) = \int p(Y|\beta,\sigma^2)p(\beta|\sigma^2)d\beta = \frac{p(Y|\beta,\sigma^2)p(\beta|\sigma^2)}{p(\beta|Y,\sigma^2)}$.

where $\beta'\beta = \sum_{j=1}^{k} \beta_j^2$.[8] Under Bayesian estimation, λ is called a hyperparameter (i.e. a parameter that indexes the prior of the parameters β), while under frequentist interpretation of this estimation method, λ is called a tuning- or regularization-parameter.[9]

With this prior and $\sigma^2 = 1$, the MDD expression from above simplifies to

$$p(Y) = \frac{(2\pi)^{-\frac{n}{2}} \exp\{-\frac{1}{2}Y'Y\}|\lambda^{-1}I_k|^{-\frac{1}{2}}}{|X'X + \lambda I|^{\frac{1}{2}} \exp\{-\frac{1}{2}\bar{\beta}'\bar{V}^{-1}\bar{\beta}\}} \quad .$$

In turn, we get the log MDD

$$\begin{aligned}
\log p(Y) &= c - \frac{1}{2}Y'Y + \frac{1}{2}\bar{\beta}'\bar{V}^{-1}\bar{\beta} - \frac{1}{2}\log\left(|\lambda^{-1}I|\right) - \frac{1}{2}\log\left(|X'X + \lambda I|\right) \\
&= c - \frac{1}{2}\left[Y'Y - Y'X\bar{V}X'Y\right] - \frac{1}{2}\log\left(\lambda^{-k}|X'X + \lambda I|\right) \\
&= c - \frac{1}{2}\left[Y'Y - Y'X[X'X + \lambda I]^{-1}X'Y\right] - \frac{1}{2}\log|\lambda^{-1}X'X + I| \ ,^{10}
\end{aligned}$$

where $c = -\frac{n}{2}\log(2\pi)$ is a constant that does not depend on Y or λ.

The (log) MDD shows that selecting the optimal λ (i.e. select the optimal model, estimation approach) implies a trade-off between in-sample fit and model complexity. The first term measures the in-sample fit. As $\lambda \to 0$, our prior becomes completely non-informative and this term goes to $-\frac{1}{2}Y'M_XY = \frac{1}{2}\hat{U}'\hat{U} = \frac{1}{2}SSR$, where $\hat{U}$ are fitted residuals from an OLS or ML (i.e. non-penalized) regression. As $\lambda \to \infty$, our prior becomes a point-mass at $\underline{\beta} = 0$ and this term goes to $-\frac{1}{2}Y'Y$; as our posterior is also a point-mass at $\bar{\beta} = 0$, we just predict values of zero for Y and we get a very large $SSR = SST = Y'Y$.

The second term is the penalty for model complexity (number of parameters to estimate). As $\lambda \to 0$, it goes to $-\infty$, whereas if $\lambda \to \infty$, it goes to $-\frac{1}{2}\log 1 = 0$. The latter signifies zero model complexity, as we do not in fact estimate anything but simply impose our prior belief. The former signifies very high model complexity, as we use a completely uninormative prior.

Ridge regression is a popular tool for predicting y_i based on x_i in a setting where there are

[8] To see this, note that maximizing the posterior – which leads to the posterior mode, equal to the posterior mean $\bar{\beta}$ under a Normal distribution – is equivalent to minimizing this objective function. Also, alternatively, we can write $\bar{\beta} = \arg\min_\beta (Y - X\beta)'(Y - X\beta)$ s.t. $\sum_{j=1}^{k} \beta_j^2 \leqslant C$, whereby there is a one-to-one relationship between λ and C, i.e. there is always a unique C that gives the same $\bar{\beta}$ above, which is computed for a given λ, and vice versa.

[9] It is a good idea to standardize the regressors before applying this estimation, i.e. for each variable in X, subtract the mean and divide by the standard deviation.

[10] The second equality inserts for $\bar{\beta} = \bar{V}X'Y$ and uses the result $\log(a) + \log(b) = \log(ab)$. The third equality inserts for $\bar{V} = [X'X + \lambda I]^{-1}$ and uses the result that $|jA| = j^m|A|$ for a scalar j and an $m \times m$ matrix A.

supposedly many covariates in the true model, i.e. when the true model is "dense". Both under the Bayesian and frequentist interpretations of this estimation procedure, a common approach of choosing λ is to maximize $p(Y)$ or an approximation of it (see below). This suggests that the optimal λ and hence optimal model (in terms of out-of-sample fit) is not the one which fits the sample data best (leads to the lowest SSR), but the one which exploits the trade-off between in-sample fit and model complexity best. This is related to the bias-variance-trade-off; for any $\lambda > 0$, the Ridge estimator $\bar{\beta}$ is biased (even if $\mathbb{E}[u_i|x_i] = 0$), but its variance is lower compared to the OLS case of $\lambda = 0$, and this possibly leads to lower frequentist risk. To maximize out-of-sample fit, it is usually desirable to sacrifice a bit on the in-sample fit but reduce model complexity, or, put differently, incur a bias but reduce the variance of the estimator.

To illustrate the trade-off inherent in $p(Y)$, suppose there are k potential regressors and we would like to know which ones we should use. As we add regressors, the goodness of fit-term increases (weakly) since SSR is non-decreasing; we could always set the coefficient of the additional regressor to zero and keep SSR constant. As we do not change λ To see what happens to the penalty term, write it as

$$-\frac{1}{2}\log|X'X+\lambda I| = -\frac{1}{2}\log\left|n\left(Q_n+\frac{\lambda}{n}I\right)\right| = -\frac{k}{2}\log n - \frac{1}{2}\log\left|Q_n+\frac{\lambda}{n}I\right| ,$$

where k is the number of regressors. This suggests that every additional regressor increases the penalty term (makes it more negative), and the larger the sample size, the larger this penalty increase.

The log MDD is often approximated by the likelihood evaluated at the ML estimator (a measure of fit) plus (the relevant part of) the above penalty for model complexity:

$$\log p(Y) \approx \log p(Y|\hat{\theta}_{ML}) - \frac{k}{2}\log n .$$

This is the so-called Bayesian or Schwarz information criterion (BIC/SIC). While under Bayesian inference, one would evaluate models and choose hyperparameters based on $p(Y)$ (see Section 5.3 below), the BIC is popular in frequentist inference as it can be computed without explicitly relying on Bayesian techniques and without assuming a prior in particular. Alternative information criteria, like the Akaike information criterion (AIC) model qualitatively the same trade-off between in-sample fit and model complexity.

5.1.1.2 Shrinkage & Selection (Lasso Regression)

Above, we assumed the Normal prior $\beta \sim N(0, \lambda^{-1}I)$. This means that, under the prior, the individual $\{\beta_j\}_{j=1:k}$ are independent with (marginal) means of zero and variances λ^{-1}, and it leads to the pdf $p(\beta) = \prod_{j=1}^{k}(2\pi\lambda^{-1})^{-\frac{1}{2}} exp\left\{-\frac{\lambda}{2}\beta_j^2\right\}$. Suppose we replace it this with the pdf $p(\beta) = \prod_{j=1}^{k}\frac{1}{2}\lambda\ exp\left\{-\lambda|\beta_j|\right\}$, which is a Laplace distribution with marginal means of zero and variances λ^{-1}. This prior has fatter tails and a higher peak at its mode, along with a kink at the mean of zero. The implied posterior mode is the LASSO estimator and solves

$$\min_{\beta} (Y - X\beta)'(Y - X\beta) + \tilde{\lambda}\|\beta\| ,$$

where $\|\beta\| = \sum_{j=1}^{k}|\beta_j|$ and $\tilde{\lambda} = 2\sigma^2\lambda$ is a one-to-one transformation of λ.[11]

With absolute values instead of squares, a low β_j contributes relatively more to the Lasso than the Ridge penalty, and vice versa for high values of β_j. Thus, as λ increases, Lasso not only shrinks all coefficients towards zero, but it also reduces some coefficents all the way to zero. As a result, Lasso estimation is used to select the most important regressors out of many potential ones, preferably in a setting where the true model is "sparse", i.e. when not all available covariates actually belong into the model. As opposed to that, Ridge regression only shrinks coefficients towards zero, but does not set any of them exactly equal to zero (it does so only in the limit as $\lambda \to \infty$). It can be used in high-dimensional models, where there are a lot of regressors, all of which could be important (non-zero), i.e. when the true model is "dense". As discussed above, by shrinking them towards zero, Ridge estimation can reduce the variance and improve the out-of-sample fit. Both Lasso and Ridge are of less use if the interest lies in estimating some β_j, i.e. the (causal) effect of x_{ij} on y_i, because then an unbiased estimator is preferred and it is important to include all variables correlated with x_{ij}, and not shrink them to zero, as would be done under Lasso.[12]

The Lasso estimator can only be obtained numerically. Under the frequentist approach, one would numerically optimize the objective function above to find the point estimator (see Section 8.1). Its finite-sample variance can be approximated by bootstrapping (see Section 7.1), while its asymptotic distribution is derived using extremum estimation theory (see Chapter 6).[13] Under the Bayesian approach, one would rely on posterior sampling

[11] Again, this can be seen by noting that maximizing the posterior is equivalent to minimizing this objective function. Also analogously to before, we can write $\tilde{\beta} = \arg\min_{\beta}(Y - X\beta)'(Y - X\beta)$ s.t. $\sum_{j=1}^{k}|\beta_j| \leqslant C$, with a one-to-one relationship between λ and C.

[12] Out of several, highly correlated variables, Lasso would shrink all but one to zero.

[13] The Ridge estimator is available analytically, and therefore its variance can be derived as well. However, its asymptotic distribution is derived in the same way. By the intuition provided above, the asymptotic distributions of the Lasso and Ridge estimators coincide with the asymptotic distribution of the OLS/ML estimator, as the prior influence vanishes asymptotically.

techniques (see Section 5.2 and Section 8.2) to obtain the posterior numerically, based on which one can compute various statistics like the mode (i.e. the Lasso point estimator), mean, variance, etc.

Both the Lasso and Ridge priors can be used in other models than linear regressions, e.g. the probit model. It is easy to see that this leads to a log-posterior that is equal to the log-likelihood plus the corresponding Lasso or Ridge penalty term. In frequentist analysis, this penalized objective function is formed without reference to the prior or posterior, just like in the penalized OLS regressions above.

5.1.2 Estimating (β, σ^2)

Above, we used the likelihood $p(Y|\beta, \sigma^2)$ and the conditional prior $\beta|\sigma^2 \sim N(\underline{\beta}, \sigma^2\underline{V})$ to derive the conditional posterior $\beta|Y, \sigma^2 \sim N(\bar{\beta}, \sigma^2\bar{V})$, though we omitted the conditioning on σ^2 for notational simplicity. If we are interested in the joint estimation of (β, σ^2), we specify a (joint) prior $p(\beta, \sigma^2)$ and derive the joint posterior $p(\beta, \sigma^2|Y)$.

While in principle any prior could be used, typically a Normal-Inverse Gamma prior is chosen because it leads to an analytical expression for the posterior. This prior breaks up the joint prior for (β, σ^2) into a conditional prior for $\beta|\sigma^2$ and a marginal prior for σ^2:

$$p(\beta, \sigma^2) = p(\beta|\sigma^2)p(\sigma^2) \ , \quad \beta|\sigma^2 \sim N(\underline{\beta}, \sigma^2\underline{V}) \ , \quad \sigma^2 \sim IG(\underline{\nu}, \underline{s}^2) \ .$$

This leads to a posterior that is split up analogously:

$$p(\beta, \sigma^2|Y) = p(\beta|Y, \sigma^2)p(\sigma^2|Y) \ ,$$

whereby the densities $p(\beta|Y, \sigma^2)$ and $p(\sigma^2|Y)$ are available analytically. They are found in two steps.

The first step simply involves repeating the calculations done for $\beta|\sigma^2$. Specifically, we derive the conditional posterior $p(\beta|Y, \sigma^2)$ and the conditional MDD $p(Y|\sigma^2)$ ("marginal likelihood of σ^2"). Recall that the first is obtained by using the proportionality of posterior to the product of likelihood and prior,

$$p(\beta|Y, \sigma^2) = \frac{p(Y|\beta, \sigma^2)p(\beta|\sigma^2)}{p(Y|\sigma^2)} \propto p(Y|\beta, \sigma^2)p(\beta|\sigma^2) \ ,$$

and by guessing and verifying the distributional family of $\beta|Y, \sigma^2$, while the second is obtained

by inverting Bayes' formula,

$$p(Y|\sigma^2) = \frac{p(Y|\beta,\sigma^2)p(\beta|\sigma^2)}{p(\beta|Y,\sigma^2)} \ ,$$

inserting the known expressions on the RHS and cancelling all terms that involve β.

The second step is analogous; we use $p(Y|\sigma^2)$ – obtained in the first step – and $p(\sigma^2)$ to derive

$$p(\sigma^2|Y) = \frac{p(Y|\sigma^2)p(\sigma^2)}{p(Y)} \propto p(Y|\sigma^2)p(\sigma^2) \ ,$$

and we then invert Bayes' formula to obtain

$$p(Y) = \frac{p(Y|\sigma^2)p(\sigma^2)}{p(\sigma^2|Y)} \ ,$$

the actual, unconditional MDD with both β and σ^2 integrated out. This is again done by inserting all the known formulas on the RHS and cancelling all terms that involve σ^2.

More specifically, the first step yields $\beta|Y,\sigma^2 \sim N(\bar{\beta},\sigma^2\bar{V})$ as well as

$$p(Y|\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}}|\underline{V}|^{-\frac{1}{2}}|\bar{V}|^{\frac{1}{2}}exp\left\{-\frac{1}{2\sigma^2}\left[Y'Y + \underline{\beta}'\underline{V}^{-1}\underline{\beta} - \bar{\beta}'\bar{V}^{-1}\bar{\beta}\right]\right\} \ ,$$

with $\bar{\beta}$ and $\bar{V}$ as defined previously. Now for the second step. Under the prior $\sigma^2 \sim IG(\underline{\nu},\underline{s}^2)$ from above, the prior-pdf of σ^2 is

$$p(\sigma^2) \propto (\sigma^2)^{-(\underline{\nu}+2)/2}exp\left\{-\frac{\underline{s}^2}{2\sigma^2}\right\} \ .$$

For the (marginal) posterior of σ^2, we then get

$$\begin{aligned}
p(\sigma^2|Y) &\propto p(Y|\sigma^2)p(\sigma^2) \\
&\propto (\sigma^2)^{-\frac{n}{2}}exp\left\{-\frac{1}{2\sigma^2}\left[Y'Y + \underline{\beta}'\underline{V}^{-1}\underline{\beta} - \bar{\beta}'\bar{V}^{-1}\bar{\beta}\right]\right\}|\sigma^2|^{-(\underline{\nu}+2)/2}exp\left\{-\frac{\underline{s}^2}{2\sigma^2}\right\} \\
&\propto (\sigma^2)^{-\frac{n+\underline{\nu}+2}{2}}exp\left\{-\frac{1}{2\sigma^2}\left[\underline{s}^2 + Y'Y + \underline{\beta}'\underline{V}^{-1}\underline{\beta} - \bar{\beta}'\bar{V}^{-1}\bar{\beta}\right]\right\} \ ,
\end{aligned}$$

which lets us deduce that

$$\sigma^2|Y \sim IG(\bar{\nu},\bar{s}^2) \ , \quad \text{with} \quad \bar{\nu} = \underline{\nu} + n \ , \quad \bar{s}^2 = \underline{s}^2 + Y'Y + \underline{\beta}'\underline{V}^{-1}\underline{\beta} - \bar{\beta}'\bar{V}^{-1}\bar{\beta} \ ,$$

i.e. the (marginal) posterior of σ^2 is also Inverse-Gamma-distributed. Finally, for the MDD

we get

$$\begin{aligned} p(Y) &= \frac{p(Y|\sigma^2)p(\sigma^2)}{p(\sigma^2|Y)} \\ &= \frac{(2\pi\sigma^2)^{-\frac{n}{2}}|\underline{V}|^{-\frac{1}{2}}|\bar{V}|^{\frac{1}{2}}exp\left\{-\frac{1}{2\sigma^2}\left[Y'Y+\underline{\beta}'\underline{V}^{-1}\underline{\beta}-\bar{\beta}'\bar{V}^{-1}\bar{\beta}\right]\right\}(\underline{s}^2)^{\underline{\nu}/2}(\sigma^2)^{-(\underline{\nu}+2)/2}exp\left\{-\frac{\underline{s}^2}{2\sigma^2}\right\}}{(\bar{s}^2)^{\bar{\nu}/2}(\sigma^2)^{-(\bar{\nu}+2)/2}exp\left\{-\frac{\bar{s}^2}{2\sigma^2}\right\}} \\ &= (2\pi)^{-\frac{n}{2}}|\underline{V}|^{-\frac{1}{2}}|\bar{V}|^{\frac{1}{2}}(\underline{s}^2)^{\underline{\nu}/2}(\bar{s}^2)^{-\bar{\nu}/2} \ . \end{aligned}$$

To summarize: starting from the Normal-Inverse Gamma prior

$$p(\beta,\sigma^2)=p(\beta|\sigma^2)p(\sigma^2) \ , \quad \beta|\sigma^2\sim N(\underline{\beta},\sigma^2\underline{V}) \ , \quad \sigma^2\sim IG(\underline{\nu},\underline{s}^2) \ ,$$

we get the Normal-Inverse Gamma posterior

$$p(\beta,\sigma^2|Y)=p(\beta|Y,\sigma^2)p(\sigma^2|Y) \ , \quad \beta|Y,\sigma^2\sim N(\bar{\beta},\sigma^2\bar{V}) \ , \quad \sigma^2|Y\sim IG(\bar{\nu},\bar{s}^2) \ ,$$

with

$$\begin{aligned} \bar{V} &= [\underline{V}^{-1}+X'X]^{-1} \ , & \bar{\beta} &= \bar{V}[X'Y+\underline{V}^{-1}\underline{\beta}] \ , \\ \bar{\nu} &= \underline{\nu}+n \ , & \bar{s}^2 &= \underline{s}^2+Y'Y+\underline{\beta}'\underline{V}^{-1}\underline{\beta}-\bar{\beta}'\bar{V}^{-1}\bar{\beta} \ . \end{aligned}$$

A draw $(\beta^m,(\sigma^2)^m)$ from this posterior is obtained by first drawing $(\sigma^2)^m\sim p(\sigma^2|Y)$ and then drawing $\beta^m\sim p(\beta|Y,(\sigma^2)^m)$. Most statistical software contain commands to draw from Normal and Inverse Gamma distributions.

5.1.3 Uniform and Improper Priors

Suppose that instead of assuming a Normal-Inverse Gamma prior for (β,σ^2), we take a Uniform prior $p(\beta,\sigma^2)\propto c$. If this prior density is not fully fledged out by specifying the bounds of this Uniform distribution, we speak of an improper prior. Under such a flat prior, the posterior of (β,σ^2) is proportional to the likelihood:

$$p(\beta,\sigma^2|Y)\propto p(Y|\beta,\sigma^2)p(\beta,\sigma^2)\propto p(Y|\beta,\sigma^2) \ .$$

Hence, the posterior mode (which is usually close to the posterior mean) is equal to the ML estimator, and, loosely speaking, Bayesians and frequentists agree not only asymptotically, but even in any finite sample.

Specifically, following the same steps as we did above under the Normal-Inverse Gamma prior, we can see that

$$p(\beta|Y,\sigma^2) \propto p(Y|\beta,\sigma^2)$$

leads to $\beta|Y,\sigma^2 \sim N(\hat{\beta},\sigma^2(X'X)^{-1})$; the conditional posterior of $\beta|\sigma^2$ has ML estimator $\hat{\beta}$ as its mean (and mode) and the variance of the ML estimator, $\mathbb{V}[\hat{\beta}|\beta] = \sigma^2(X'X)^{-1}$, as its variance. Moreover,

$$p(\sigma^2|Y) \propto p(Y|\sigma^2)$$

leads to $\sigma^2|Y \sim IG\left(n-k-2, n\hat{\sigma}^2\right)$,[14] i.e. the (marginal) posterior mean of σ^2 is $\frac{n}{n-k-2}\hat{\sigma}^2$ and the mode is $\frac{n}{n-k+2}\hat{\sigma}^2$.

5.2 Bayesian Inference for General Models

In the above estimation of the linear regression model, the likelihood is available analytically: it is a Normal distribution. Taking the prior to be Normal as well, we get an analytical expression for the posterior. In fact, we call this a "conjugate" prior, because it leads to a posterior in the same distributional family. Likewise, the Inverse-Gamma prior for σ^2 and the joint, Normal-Inverse-Gamma prior for (β,σ^2) are conjugate as well.

In many other cases, like the Laplace/Lasso prior above, the posterior cannot be derived analytically. In fact, it may be that not even the likelihood is available analytically, as is the case for dynamic macroeconomic (DSGE) models. In those cases, we derive the posterior numerically using posterior sampling methods. These are discussed in Section 8.2. The following is a high-level overview of how to conduct Bayesian estimation numerically using these methods.

The only requirement to apply posterior sampling methods is that one needs to be able to evaluate the likelihood and prior.[15] The output of these sampling techniques is a set of draws from the posterior, $\{\theta^m\}_{m=1}^M$, or, more generally, a so-called particle approximation of the posterior, $\{\theta^m, W^m\}_{m=1}^M$, i.e. a set of parameter values with assigned weights (which could be all the same). They approximate the posterior in the sense that

$$\mathbb{E}[g(\theta)|Y] = \int g(\theta)p(\theta|Y)d\theta \approx \frac{1}{M}\sum_{m=1}^{M} W^m g(\theta^m) \ ,$$

i.e. we can approximate any moment of the posterior distribution (like its mean, variance,

[14] Note that $Y'Y - \hat{\beta}'X'X\hat{\beta} = (Y-X\hat{\beta})'(Y-X\hat{\beta}) = n\hat{\sigma}^2$.

[15] Sometimes, we also need to be able to sample from the prior, but this is usually no shortcoming, as the prior is specified by the researcher and available analytically.

5th percentile, etc.) using the particles that come out of the posterior sampler.[16] An approximation of the marginal posterior of some $\theta_1 \subset \theta$ is obtained simply by taking the corresponding parts of the particle-values: $\{\theta_1^i, W^m\}_{m=1}^M$.

5.3 Model Selection & Model Averaging

An important part of empirical work is the search for the right model (specification). The above discussion on selecting λ or the number of regressors are examples of model selection. They illustrate that model uncertainty can be thought of as (hyper)parameter uncertainty, as elaborated on in the following. Model selection is an integral part of Bayesian inference, whereas under frequentist inference, it is problematic, as parameters are regarded as fixed and inference conditions on a particular model.[17] [18]

Model Selection Suppose we have a collection of models $(M_j)_{j=1}^J$ with parameter vectors θ^j, likelihood functions $p(Y|\theta^j, M_j)$ and prior distributions $p(\theta^j|M_j)$. Starting from some prior model probabilities $\pi_{j,0} = \mathbb{P}[M_j \text{ is true}]$, we can compute posterior model probabilities as

$$\pi_{j,n} = \mathbb{P}[M_j \text{ is true } |Y] = \frac{\pi_{j,0} p(Y|M_j)}{\sum_{j=1}^J \pi_{j,0} p(Y|M_j)} \propto \pi_{j,0} p(Y|M_j) \ ,$$

where $p(Y|M_j)$ is the marginal likelihood for model M_j. Given posterior model probabilities, one could select the model with the highest posterior probability or compare two models M_j and M_i using posterior odds $\pi_{j,n}/\pi_{i,n}$. If the models are regarded as equally likely a-priori – $\pi_{j,0} = \pi_{i,0}$ –, the posterior odds are equal to the ratio of MDDs: $\pi_{j,n}/\pi_{i,n} = p(Y|M_j)/p(Y|M_i)$, i.e. the relative probability of M_j being true equals its relative out-of-sample fit.

For example, say we have $J = 2$ possible regression specifications:

$$M_1: \ \ y_i = \beta_1^1 x_{1i} + \beta_2^1 x_{2i} + u_i \ , \qquad \text{and} \qquad M_2: \ \ y_i = \beta_1^2 x_{1i} + v_i \ .$$

To select among the two, we could estimate each model and compute their respective MDDs $p(Y|M_j)$, $j = 1, 2$. Together with prior model probabilities $\{\pi_{j,0}\}_{j=1,2}$, they lead to posterior model probabilities $\{\pi_{j,n}\}_{j=1,2}$.

[16] It is important to store the particles after running the posterior sampling algorithm on the hard drive in order not to have to run the algorithm anew every time a new posterior statistic has to be computed for the same model.

[17] Nevertheless, even in the frequentist paradigm, researchers do perform model selection by relying on heuristic statistics such as the BIC.

[18] For more details, see the lecture delivered by Chris Sims at the 7th Lindau meeting of economic sciences: https://mediatheque.lindau-nobel.org/recordings/40282/christopher-sims.

This model selection problem can also be thought of as a hyperparameter selection problem. To illustrate, suppose for simplicity that the (marginal) prior for β_1 is the same under both models. Let

$$p(\beta_1^1, \beta_2^1|M_1) = N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right) , \qquad \text{and} \qquad p(\beta_1^2|M_1) = N(0,1) .$$

Now, say that we conduct estimation just for a single model, the more general of the two above, which nests both specifications:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i , \quad \text{with prior} \quad \beta \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & \lambda\end{bmatrix}\right) .$$

As $\lambda \to 0$, we get model M_2. Therefore, the above model selection problem is equivalent to estimating this single model and selecting between $\lambda = 0$ and $\lambda = 1$. The crucial point behind this result is that the restricted model M_2 can be thought of as estimating this general model and using the prior

$$\beta \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 0\end{bmatrix}\right) ,$$

i.e. the marginal prior for β_1 is $N(0,1)$ and the marginal prior for β_2 is a point-mass at zero.[19]

Model Averaging Rather than selecting the most likely model, one could also use the posterior model probabilities to construct an overall, "model-averaged" posterior:

$$p(\theta|Y) = \sum_{j=1}^{J} \pi_{j,n} p(\theta|Y, M_j) .$$

The resulting posterior mean is a weighted average of the posterior means under the individual models. Model averaging is particularly common in forecasting time series variables, as the forecast averaged over many models tends to outperform any single model.

Just as model selection can be thought of as selecting hyperparameter-values, model averaging can also be thought of as averaging across hyperparameter-values. Consider the regression example from above, and let $p(Y|\beta)$ be the likelihood associated with our single, general specification. Take the prior to be a combination of the priors $p(\beta^1|M_1)$ and $p(\beta^2|M_2)$

[19] Of course, strictly speaking, the Normal distribution is not defined for a variance of zero. However, $N(0,0)$ should be thought of as the limit of $N(0,\lambda)$ as $\lambda \to 0$, which is a point-mass at zero.

with respective weights $\pi_{1,0}$ and $\pi_{2,0}$:

$$p(\beta) = \pi_{1,0}p(\beta|M_1) + \pi_{2,0}p(\beta|M_2) = \pi_{1,0}N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right) + \pi_{2,0}N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 0\end{bmatrix}\right) .$$

In other words, our prior averages across values $\lambda = 0$ and $\lambda = 1$. The resulting posterior is the model-averaged posterior:

$$\begin{aligned}
p(\beta|Y) &= \frac{p(Y|\beta)p(\beta)}{p(Y)} \\
&= \pi_{1,0}\frac{p(Y|\beta)p(\beta|M_1)}{p(Y)} + \pi_{2,0}\frac{p(Y|\beta)p(\beta|M_2)}{p(Y)} \\
&= \pi_{1,0}\frac{p(\beta|Y,M_1)p(Y|M_1)}{p(Y)} + \pi_{2,0}\frac{p(\beta|Y,M_2)p(Y|M_2)}{p(Y)} \\
&= \pi_{1,n}p(\beta|Y,M_1) + \pi_{2,n}p(\beta|Y,M_2) \ .^{20}
\end{aligned}$$

Hierarchical Bayes Hierarchical Bayes modeling generalizes model-selection and -averaging. Under this approach, we treat hyperparameters λ just like the other parameters, θ, despite the fact that θ appears in the likelihood, whereas λ does not. Given a prior distribution $p(\theta, \lambda) = p(\theta|\lambda)p(\lambda)$, one forms the posterior $p(\theta, \lambda|Y)$.

To illustrate, consider the example from above. Note that the above prior, obtained by averaging two priors with different values of the hyperparameter λ,

$$p(\beta) = \pi_{1,0}p(\beta|M_1) + \pi_{2,0}p(\beta|M_2) = \pi_{1,0}p(\beta|\lambda = 1) + \pi_{2,0}p(\beta|\lambda = 0) \ ,$$

is the marginal prior for β under the following, joint prior for (β, λ):

$$p(\beta, \lambda) = p(\beta|\lambda)p(\lambda) \ , \quad \text{with } \ \beta|\lambda \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & \lambda\end{bmatrix}\right) \ , \quad \lambda = \begin{cases} 0 & \text{w.p. } \pi_{1,0} \\ 1 & \text{w.p. } \pi_{2,0} \end{cases} ,$$

i.e. the joint prior for $p(\beta, \lambda)$ is the product of a conditional prior for $\beta|\lambda$ and a marginal prior – a so-called hyperprior – for λ. By the above discussion, we know that this prior leads

[20]The third line uses the Bayes formula, $p(\beta|M_j) = p(Y|\beta)p(\beta|M_j)/p(Y|M_j)$, and rearranges it to insert for $p(Y|\beta)p(\beta|M_j)$. The fourth line reasons as follows. We know $\pi_{j,n} \propto \pi_{j,0}p(Y|M_j)$. We also know that, because $p(\beta|Y, M_1)$ and $p(\beta|Y, M_2)$ integrate to one, for $p(\beta|Y)$ to integrate to one, $\pi_{1,0}p(Y|M_1)/p(Y) \propto \pi_{1,n}$ and $\pi_{2,0}p(Y|M_2)/p(Y) \propto \pi_{2,n}$ need to sum to one. Together, this implies $\pi_{j,0}p(Y|M_j)/p(Y) = \pi_{j,n}$, $j = 1, 2$.

to the posterior

$$p(\beta|Y) = \pi_{1,n}p(\beta|Y, M_1) + \pi_{2,n}p(\beta|Y, M_2) = \pi_{1,n}p(\beta|Y, \lambda = 1) + \pi_{2,n}p(\beta|Y, \lambda = 0) \ .$$

This posterior is the marginal posterior of β from the following, joint posterior for (β, λ):

$$p(\beta, \lambda|Y) = p(\beta|Y, \lambda)p(\lambda|Y) \ , \quad \text{with } \ \lambda|Y = \begin{cases} 0 & \text{w.p. } \pi_{1,n} \\ 1 & \text{w.p. } \pi_{2,n} \end{cases} .$$

This illustrates that a model-averaged posterior is equivalent to the marginal posterior for the parameters β under a hierarchical Bayes model with a suitable hyperprior for λ. Moreover, selecting the model with higher posterior probability is equivalent to conducting a hierarchical Bayes estimation, finding the (marginal) mode λ^* of the hyperparameters λ (i.e., here, $\lambda = 1$ if $\pi_{2,n} > \pi_{1,n}$, and $\lambda = 0$ otherwise) and taking the conditional posterior $p(\beta|Y, \lambda^*)$ with the value of hyperparameters set to this mode.

Priors are (almost) always indexed by some hyperparameters, which determine their exact location and shape. For example, in Section 5.1, the Normal-Inverse Gamma prior for $\theta = (\beta, \sigma^2)$ in the linear regression model uses the hyperparameters $\lambda = (\underline{\beta}, \underline{V}, \underline{\nu}, \underline{s}^2)$. The Ridge/shrinkage prior sets $\underline{\beta} = 0$ and $\underline{V} = \lambda^{-1}I$, limiting the set of hyperparameters when $\beta|\sigma^2$ is estimated to a single scalar λ. Given a prior for θ, $p(\theta|\lambda)$, one derives the posterior $p(\theta|Y, \lambda)$ and log MDD $p(Y|\lambda) = \int p(Y|\theta)p(\theta|\lambda)d\theta = \frac{p(Y|\theta)p(\theta|\lambda)}{p(\theta|Y,\lambda)}$, all conditional on these hyperparameters λ.

While researchers usually have a rough idea which values for the parameters θ are a-priori sensible, they often do not have a good way to specify the values of all hyperparameters in λ. Model selection and model averaging are two ways to deal with this.[21] Under the former we select the λ that maximizes the MDD $p(Y|\lambda)$ (recall that in Section 5.1.1.1, we argued that the λ^* that maximizes the MDD $p(Y|\lambda)$ leads to the model that optimizes the trade-off between in-sample fit and model complexity, leading to the best out-of-sample fit), while under the latter we consider many values for λ and construct a model-averaged posterior as the weighted average of the posteriors under different values for λ.

Both model selection and model averaging are facilitated by hierarchical Bayes modeling. Usually, a Uniform hyperprior is assumed: $p(\lambda) \propto c$. This means that we regard each value of λ (each model) as equally likely a-priori. We get the prior $p(\theta, \lambda) = p(\theta|\lambda)p(\lambda) \propto p(\theta|\lambda)$. Under this prior, the marginal posterior of λ has the same shape as the log MDD conditional

[21] The third is to conduct a prior sensitivity analysis, i.e. to explore the robustness of results to changes in the prior (in the values of (some) hyperparameters).

on λ, $p(Y|\lambda)$:

$$p(\lambda|Y) \propto p(Y|\lambda)p(\lambda) \propto p(Y|\lambda) \ .$$

Therefore, the λ^* that maximizes $p(Y|\lambda)$ is equal to the mode of $p(\lambda|Y)$ in the hierarchical Bayes model, just as in the simple example above.[22] Moreover, averaging $p(\theta|Y,\lambda)$ for many values λ using weights proportional to $p(Y|\lambda_j)p(\lambda_j)$ is equivalent to taking the marginal posterior

$$p(\theta|Y) = \int p(\theta,\lambda|Y)p(\lambda|Y)d\lambda \ ,$$

analogous to the simple example above.[23] [24]

Sometimes, we can find λ^* and $p(\theta|Y)$ analytically. In other cases, we find them numerically. Given a numerical approximation of the posterior $p(\theta,\lambda|Y)$, $\{(\theta^m,\lambda^m),W^m\}_{m=1}^M$, we get a numerical approximation of the marginal posterior $p(\lambda|Y)$ by simply taking $\{\lambda^m,W^m\}_{m=1}^M$. The mode of this distribution is λ^*. Similarly, we get a numerical approximation of the marginal posterior $p(\theta|Y)$ by simply taking $\{\theta^m,W^m\}_{m=1}^M$.

Appendix

[22] Note that this procedure works always, whereas maximizing $p(Y|\lambda)$ w.r.t. λ is cumbersome and inefficient if $p(Y|\lambda)$ is not available analytically. One would first need to derive $p(\theta|Y,\lambda)$ and $p(Y|\lambda)$ many times, each time for a different value of λ. Then, among all those values, one would need to choose the value of λ that leads to the highest $p(Y|\lambda)$.

[23] In the example, we average $p(\theta|Y,\lambda_j)$ for two discrete values $\lambda_j \in \{0,1\}$ using weights $\pi_{j,n}$ proportional to $p(Y|\lambda_j)p(\lambda_j)$, whereas here we perform an integration as we average over a continuous λ.

[24] See Giannone et al. (2015) for a more detailed discussion on hierarchical Bayes modeling.