DERIVING PROPER UNIFORM PRIORS FOR REGRESSION COEFFICIENTS

N. van Erp and P. van Gelder

Structural Hydraulic and Probabilistic Design, TU Delft
Delft, The Netherlands

Abstract. In problems of model comparison between competing regression models, one must take care not to use improper priors. Improper priors introduce inverse infinities in the evidence factors, which do not cancel if one proceeds to compute the posterior probabilities of models which have different numbers of regression coefficients. We therefore derive a simple and parsimonious proper uniform prior for multiple regression models. We then look at the evidence values that result from using this prior.

Keywords: Bayesian Model Selection, Proper Uniform Priors, Regression Coefficients, Evidence Values, Invariance
PACS: 02.50.Cw

INTRODUCTION

There is a long tradition of the use of improper uniform priors for regression coefficients, that is, location parameters, in problems of parameter estimation [1]. However, in problems of model comparison between competing regression models one must generally avoid improper priors, whether uniform or not. This is because improper priors will introduce inverse infinities in the evidence values which may not cancel out if one computes the corresponding posterior probabilities [2]. We shall therefore derive a parsimonious proper uniform prior for univariate regression models. This univariate prior is then generalized to its multivariate equivalent. The resulting proper uniform prior has the nice property that it causes the scaled evidence values, e.g. the posterior probabilities of the competing regression models, to be invariant for changes in scale in both the dependent and independent variables.

DERIVING PROPER UNIFORM PRIORS FOR THE UNIVARIATE CASE

In what follows we derive the, trivial, limits of the univariate uniform prior for a single regression coefficient. The extension to the multivariate case (in the next paragraph) is based upon the basic idea introduced here.

Suppose we wish to regress an dependent vector $\mathbf{y}$ upon an independent vector $\mathbf{x}$. Then, using matrix algebra, the regression coefficient β may be computed as

$$\beta = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle} = \frac{\|\mathbf{x}\| \, \|\mathbf{y}\| \cos\theta}{\|\mathbf{x}\|^2} = \frac{\|\mathbf{y}\|}{\|\mathbf{x}\|} \cos\theta. \tag{1}$$

By examining (1) we see that β must lie in the interval

$$\frac{\|\mathbf{y}\|_{\max}}{\|\mathbf{x}\|_{\min}}(\cos\theta)_{\min} \leq \beta \leq \frac{\|\mathbf{y}\|_{\max}}{\|\mathbf{x}\|_{\min}}(\cos\theta)_{\max}. \tag{2}$$

Since $(\cos\theta)_{\min} = -1$ and $(\cos\theta)_{\max} = 1$, interval (2) reduces to:

$$-\frac{\|\mathbf{y}\|_{\max}}{\|\mathbf{x}\|_{\min}} \leq \beta \leq \frac{\|\mathbf{y}\|_{\max}}{\|\mathbf{x}\|_{\min}}. \tag{3}$$

So, having prior knowledge of the minimal length of the predictor, $\|\mathbf{x}\|_{\min}$, and the maximal length of the outcome variable, $\|\mathbf{y}\|_{\max}$, we may set limits to the possible values of β. It follows that the proper uniform prior of β must be the univariate uniform distribution with limits as given in (3):

$$p(\beta \mid I) = \frac{\|\mathbf{x}\|_{\min}}{2\|\mathbf{y}\|_{\max}}, \qquad -\frac{\|\mathbf{y}\|_{\max}}{\|\mathbf{x}\|_{\min}} \leq \beta \leq \frac{\|\mathbf{y}\|_{\max}}{\|\mathbf{x}\|_{\min}}, \tag{4}$$

where I stands for the prior background information that allows us to assign values to $\|\mathbf{x}\|_{\min}$ and $\|\mathbf{y}\|_{\max}$.

DERIVING PROPER UNIFORM PRIORS FOR THE MULTIVARIATE CASE

We now derive the limits of the multivariate uniform prior for m regression coefficients. The basic idea is a generalization of the very simple idea used to derive the limits for the univariate case. This generalization will involve a transition from univariate line pieces to multivariate ellipsoids.

Say we have m independent vectors $\mathbf{x}_i$ that span some m-dimensional subspace S_m. Let $\hat{\mathbf{y}}$ be the part of $\mathbf{y}$ that lies in S_m and let $\mathbf{e}$ be the part that is orthogonal to S_m, then

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}. \tag{5}$$

Now, the projection $\hat{\mathbf{y}}$ is mapped on the orthogonal base spanned by the vectors $\mathbf{x}_i$ through the regression coefficients β_i, that is,

$$\hat{\mathbf{y}} = \sum_{i=1}^{m} \mathbf{x}_i \beta_i, \tag{6}$$

where

$$\beta_i = \frac{\langle \mathbf{x}_i, \mathbf{y} \rangle}{\langle \mathbf{x}_i, \mathbf{x}_i \rangle} = \frac{\langle \mathbf{x}_i, \hat{\mathbf{y}} + \mathbf{e} \rangle}{\langle \mathbf{x}_i, \mathbf{x}_i \rangle} = \frac{\langle \mathbf{x}_i, \hat{\mathbf{y}} \rangle}{\langle \mathbf{x}_i, \mathbf{x}_i \rangle} = \frac{\|\hat{\mathbf{y}}\|}{\|\mathbf{x}_i\|}\cos\theta_i. \tag{7}$$

Because of the independence of the $\mathbf{x}_i$ we have that $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = 0$ for $i \neq j$. So, if we take the norm of (6) we find

$$\|\hat{\mathbf{y}}\|^2 = \left\| \sum_{i=1}^{m} \mathbf{x}_i \beta_i \right\|^2 = \|\hat{\mathbf{y}}\|^2 \sum_{i=1}^{m} \cos^2\theta_i. \tag{8}$$

It follows from identity (8) that the angles θ_i in (7) must obey the constraint

$$\sum_{i=1}^{m} \cos^2 \theta_i = 1. \tag{9}$$

Combining (7) and (9), we see that all possible values of the coordinates β_i must lie on the surface of an m-variate ellipsoid centered at the origin and with respective axes

$$r_i = \frac{\|\hat{\mathbf{y}}\|}{\|\mathbf{x}_i\|}. \tag{10}$$

Since $\|\hat{\mathbf{y}}\| \leq \|\mathbf{y}\|$, the axes (10) may be maximized through our prior knowledge of the minimal lengths of the predictors and the maximal length of the outcome variable, that is,

$$\max(r_i) = \frac{\|\mathbf{y}\|_{\max}}{\|\mathbf{x}_i\|_{\min}}. \tag{11}$$

It follows that all possible values of the regression coefficients β_i must lie in the m-variate ellipsoid centered at the origin and with respective axes (11). If we substitute (11) into the identity for the volume of an m-variate ellipsoid

$$V = \frac{\pi^{m/2}}{\Gamma[(m+2)/2]} \prod_{i=1}^{m} r_i, \tag{12}$$

we find the volume of the parameter space of all possible values of the β_i, as

$$V = \frac{\pi^{m/2}}{\Gamma[(m+2)/2]} \frac{\|\mathbf{y}\|_{\max}^m}{\prod_{i=1}^m \|\mathbf{x}_i\|_{\min}}. \tag{13}$$

Let $X \equiv [\mathbf{x}_1 \cdots \mathbf{x}_m]$. Then for m independent variables $\mathbf{x}_i$ the product of the norms is equivalent to the square root of the determinant of $X^T X$, that is,

$$\prod_{i=1}^{m} \|\mathbf{x}_i\| = \left|X^T X\right|^{1/2}, \tag{14}$$

which is also the volume of the parallelepiped defined by the vectors $\mathbf{x}_i$. If the m predictors $\mathbf{x}_i$ are not independent we can transform them to an orthogonal basis $\tilde{\mathbf{x}}_i$ using the Gram-Schmidt orthogonalization process. Let $\tilde{X} \equiv [\tilde{\mathbf{x}}_1 \cdots \tilde{\mathbf{x}}_m]$. Then, since the volume of the parallelepiped is invariant under orthogonalization, we have

$$\left|\tilde{X}^T \tilde{X}\right|^{1/2} = \left|X^T X\right|^{1/2}. \tag{15}$$

Upon making the appropriate substitutions, we find the maximum volume of the prior accessible parameter space of the β_i, for both independent and dependent predictors, to be

$$V_{\max} = \frac{\pi^{m/2}}{\Gamma[(m+2)/2]} \frac{\|\mathbf{y}\|_{\max}^m}{|X^T X|_{\min}^{1/2}}. \tag{16}$$

That is, the parameter point $(\beta_1,\ldots,\beta_m)$ must lie in some ellipsoid with maximum volume (16).

To assign the proper uniform prior that spans the largest possible parameter space, we let this proper prior be equal in value to the inverse of (16). It then follows that the multivariate equivalent of (4) may be written as

$$p(\beta_1,\ldots,\beta_m \mid I) = \frac{\Gamma[(m+2)/2]}{\pi^{m/2}} \frac{\left|X^T X\right|_{\min}^{1/2}}{\|\mathbf{y}\|_{\max}^{m}}, \tag{17}$$

where $(\beta_1,\ldots,\beta_m) \in ellipsoid$ and

$$\left|X^T X\right|_{\min}^{1/2} = \left(\prod_{i=1}^{m} \|\mathbf{x}_i\|_{\min}\right) \begin{vmatrix} 1 & |\cos\phi_{12}|_{\max} & \cdots & |\cos\phi_{1m}|_{\max} \\ |\cos\phi_{12}|_{\max} & 1 & \cdots & |\cos\phi_{2m}|_{\max} \\ \vdots & \vdots & \ddots & \vdots \\ |\cos\phi_{1m}|_{\max} & |\cos\phi_{2m}|_{\max} & \cdots & 1 \end{vmatrix}^{1/2} \tag{18}$$

where $\left|\cos\phi_{ij}\right|$ is the absolute value of the correlation between the predictors $\mathbf{x}_i$ and $\mathbf{x}_j$ and I stands for the prior background information that allows us to assign values to $\|\mathbf{y}\|_{\max}$ and $\left|X^T X\right|_{\min}^{1/2}$.

From (17) and (18), we see that maximizing the volume of the prior parameter space is accomplished by maximizing the length of the dependent variable $\mathbf{y}$ and minimizing the determinant of the inner product of the matrix X. The latter is accomplished by minimizing the lengths of the independent variables $\mathbf{x}_1,\ldots,\mathbf{x}_m$ and maximizing the absolute values of their correlations, $\cos\phi_{12},\ldots,\cos\phi_{m-1,m}$.

DECONSTRUCTING THE EVIDENCE VALUES

Having derived a suitable parsimonious proper uniform prior for the multivariate case, we now look more closely at the evidence values which result from using this prior.

Suppose we wish the compute the evidence of a specific model M

$$M : \mathbf{y} = X\beta + \mathbf{e}, \tag{19}$$

where $\beta = (\beta_1,\ldots,\beta_m)$, $\mathbf{e} = (e_1,\ldots,e_N)$ and $e_i \sim N(0,\sigma)$ for $i = 1,\ldots,N$ and some value of σ. Then the corresponding likelihood is

$$p(\mathbf{y} \mid X,\beta,\sigma,M) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{(\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta)}{2\sigma^2}\right]. \tag{20}$$

For the unknown beta coefficients β of model (19) we will use the derived proper uniform prior (17), and for the unknown error σ we use a proper Jeffreys prior with normalizing constant A. The proper prior for the unknown parameters then becomes

$$p(\beta,\sigma \mid I) = \frac{\Gamma[(m+2)/2]}{\pi^{m/2}} \frac{\left|X^T X\right|_{\min}^{1/2}}{\|\mathbf{y}\|_{\max}^{m}} \frac{A}{\sigma}, \tag{21}$$

where $(\beta_1,\ldots,\beta_m) \in ellipsoid$ and $a \leq \sigma \leq b$. If we combine the prior (21) with the likelihood (20), we get the following multivariate distribution

$$p(\mathbf{y},\beta,\sigma \mid X,M,I) = \frac{\Gamma[(m+2)/2]}{\pi^{m/2}} \frac{\left|X^T X\right|_{\min}^{1/2}}{\|\mathbf{y}\|_{\max}^{m}} \frac{A}{\sigma} \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left[-\frac{(\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta)}{2\sigma^2}\right]. \tag{22}$$

Upon integrating out the $m+1$ unknown parameters β and σ we are left with the following evidence value, [1]:

$$p(\mathbf{y} \mid X,M,I) = \frac{\Gamma[(m+2)/2]}{\pi^{m/2}} \frac{\left|X^T X\right|_{\min}^{1/2}}{\|\mathbf{y}\|_{\max}^{m}} A \frac{(2\pi)^{(m-N)/2}}{\left|X^T X\right|^{1/2}} \frac{\Gamma[(N-m)/2]}{2} \frac{2^{(N-m)/2}}{\|\mathbf{y}-\hat{\mathbf{y}}\|^{N-m}} \tag{23}$$

where $\hat{\mathbf{y}} \equiv X\left(X^T X\right)^{-1} X^T \mathbf{y}$ is the maximum likelihood estimate of $\mathbf{y}$. We may rearrange these terms to get

$$p(\mathbf{y} \mid X,M,I) = \frac{\|\mathbf{y}-\hat{\mathbf{y}}\|^m}{\left|X^T X\right|^{1/2}} \frac{\left|X^T X\right|_{\min}^{1/2}}{\|\mathbf{y}\|_{\max}^{m}} \frac{\Gamma[(m+2)/2]\,\Gamma[(N-m)/2]}{\|\mathbf{y}-\hat{\mathbf{y}}\|^N} \frac{A}{2\pi^{N/2}}. \tag{24}$$

In the evidence (24), the factor

$$\text{shrinkage} = \frac{\|\mathbf{y}-\hat{\mathbf{y}}\|^m}{\left|X^T X\right|^{1/2}} \frac{\left|X^T X\right|_{\min}^{1/2}}{\|\mathbf{y}\|_{\max}^{m}} \tag{25}$$

is the shrinkage of the posterior accessible parameter space of β relative to the prior accessible space. The greater this shrinkage, the smaller the evidence. Shrinkage may occur if $\hat{\mathbf{y}} \to \mathbf{y}$. If this is the case, then the factor that rewards goodness of fit

$$\text{goodness of fit} = \frac{1}{\|\mathbf{y}-\hat{\mathbf{y}}\|^N} \tag{26}$$

will compensate for the shrinkage penalty. So there is a trade-off between the shrinkage (25) and the goodness of fit (26). Furthermore, there is a factor

$$\text{penalty} = \Gamma[(m+2)/2]\,\Gamma[(N-m)/2] \tag{27}$$

that penalizes regression models with a greater number of regression coefficients m. The last factor $A/\left(2\pi^{N/2}\right)$ is shared by the evidence values of all the competing regression models and eventually cancels out if one computes the corresponding posterior probabilities.

INVARIANCE FOR CHANGE OF SCALE OF THE EVIDENCE

The evidence value (24) is invariant under a change in scale of the dependent variables, that is, for the transformation $\tilde{X} \mapsto cX$, where c is some constant. The normalized

evidence value, e.g. posterior probability of the model, is also invariant for a change in scale in the dependent variable, that is, under the transformation $\tilde{\mathbf{y}} \mapsto c\mathbf{y}$.

If one uses instead of prior (21) the more generic prior

$$p(\beta, \sigma \mid I) = \frac{AB}{\sigma}, \tag{28}$$

then the corresponding evidence term becomes

$$p(\mathbf{y} \mid X, M, I) = \frac{B}{|X^T X|^{1/2}} \frac{\Gamma[(N-m)/2]}{\|\mathbf{y} - \hat{\mathbf{y}}\|^{N-m}} \frac{A}{2\pi^{(N-m)/2}} \tag{29}$$

and there seems to be a lack of invariance of the evidence values of regression models under changes of scale. But if we replace the generic term B with the proper uniform prior (17), we see that invariance was never lacking.

DISCUSSION

Using informational consistency requirements, Jaynes [3] derived the form of maximal non-informative priors for location parameters, that is, regression coefficients, to be uniform. However, this does not tell us what the limits of this this uniform distribution should be, that is, what particular uniform distribution to use. If we are faced with a parameter estimation problem these limits of the uniform prior are irrelevant, since we may scale the product of the improper uniform prior and the likelihood to one, thus obtaining a properly normalized posterior. However, if we are faced with a problem of model selection then the value of the uniform prior is an integral part of the evidence, which is used to rank the various competing models. We have given here some guidelines for choosing a parsimonious proper uniform prior. To construct such a parsimonious prior one needs to assign a maximal length to the dependent variable $\mathbf{y}$, minimal lengths to the independent variables $\mathbf{x}_1, \ldots, \mathbf{x}_m$ and maximal absolute values to the correlations of the independent variables, $\cos\phi_{12}, \ldots, \cos\phi_{m-1,m}$.

REFERENCES

1. Arnold Zellner, *An Introduction to Bayesian Inference in Econometrics*, J. Wiley & Sons, Inc., New York (1971); Wiley Classics Library Edition (1996).
2. Larry G. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation*, Springer-Verlag, New-York, 1988.
3. Edwin T. Jaynes, Prior probabilities, *IEEE Trans. Systems Sci. Cybernetics* **SSC-4** (3), 227–241 (1968).