A Symmetric Prior for the Regression Coefficients of a Nominal Input Variable

Kevin S. Van Horn

July 5, 2015

Note: I first derived this prior circa 2005, but did not publish it. Lenk and Orme [4] independently propose an "effects prior" using the same covariance matrix Σ described here, in the context of a hierarchical regression model. Their derivation assumes an effects coding and proceeds from different premises than those used herein.

1 Effects

Suppose that we have a K-level nominal input variable used in a Bayesian regression analysis, with each level k encoded as a row vector

$$X_k = (x_{k1}, \ldots, x_{kp}) .$$

Let β_i be the regression coefficient corresponding to the i-th element of the encoding, so that a level of k contributes the term

$$\alpha_k = X_k \beta = \sum_i \beta_i x_{ki}$$

to the overall regression sum. We call α_k the *effect* of level k.

Any prior on β defines a corresponding joint prior on the effects α_k via the above equation. Our goal is to construct an appropriate prior distribution for β using as our only prior information some notion of how large any of the effects may plausibly be: we want the prior mean for each α_k to be 0, and the prior variance to be some given value σ^2. Since this information makes no distinction between the levels, the joint prior for the effects should be symmetric: reordering the levels should leave this joint prior unchanged.

We would like the effects to indicate the differences between levels, and not include any constant (independent of level) contribution to the overall

regression sum; thus we require that

$$\sum_{k=1}^{K} \alpha_k = 0.$$

This implies that the joint distribution for the effects is degenerate. In the remainder of this note we therefore define the vector α to be the first $K-1$ effects,

$$\alpha' = (\alpha_1, \ldots, \alpha_{K-1}),$$

and use

$$\alpha_K = -\sum_{k=1}^{K-1} \alpha_k.$$

2 Encodings

Using $\alpha_k = X_k \beta$ we have

$$0 = \sum_{k=1}^{K} \alpha_k = \left(\sum_{k=1}^{K} X_k \right) \beta$$

and so, assuming a non-degenerate (full-dimensional) prior over β, the level encodings must satisfy

$$\sum_{k=1}^{K} X_k = \mathbf{0}.$$

We therefore define the matrix X to be the first $K-1$ row vectors X_k:

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_{K-1} \end{pmatrix}$$

and use

$$X_K = -\sum_{k=1}^{K-1} X_k.$$

Our equation defining the effects then becomes

$$\alpha = X\beta$$

and so, to have a one-to-one correspondence between effects vectors α and regression-coefficient vectors β, we require that X be invertible. That is,

- X must be a square matrix (we require $p = K - 1$);
- no level encoding X_k, $k \neq K$, may be expressible as a linear combination of the remaining level encodings (excluding X_K).

One example of an encoding satisfying these requirements is effects coding:

$$\begin{aligned} x_{ki} &= \begin{cases} 1 & \text{if } i = k \\ 0 & \text{if } i \neq k \end{cases} \quad \text{for } k \neq K \\ x_{Ki} &= -1 \quad \text{for all } i. \end{aligned}$$

3 An obvious prior that doesn't work

With effects coding the obvious symmetric prior for β,

$$\beta_k \sim \text{Normal}\,(0, \sigma)\,,$$

leads to a very *asymmetric* prior for the effects α_k: for $k \neq K$ we have

$$\alpha_k \sim \text{Normal}\,(0, \sigma)$$

independently ($\text{Cov}\,(\alpha_j, \alpha_k) = 0$ if $j, k \neq K$), but for $k = K$ we have

$$\alpha_K \sim \text{Normal}\left(0,\, \sqrt{K-1}\sigma\right)$$

and for $k \neq K$ the covariance between α_k and α_K is

$$\begin{aligned} \text{Cov}\,(\alpha_k, \alpha_K) &= \sum_{j=1}^{K-1} \text{Cov}\,(\alpha_k, \alpha_j) \\ &= \sigma^2. \end{aligned}$$

4 Solution strategy

We find an appropriate prior for β by first constructing a symmetric prior for the effects themselves, then solving for the corresponding prior on β. The prior we derive for α turns out to be a multivariate normal with mean vector $\mathbf{0}$ and a covariance matrix Σ defined later. Since

$$\beta = X^{-1}\alpha$$

the required prior for β is

$$\begin{aligned} \beta &\sim \text{Normal}\left(\mathbf{0},\, W\Sigma W'\right) \\ W &= X^{-1}. \end{aligned}$$

For the effects coding, X is just the identity matrix (remember that X only has $K-1$ rows, omitting X_K), and so the prior covariance matrix for β is just Σ itself.

We seek to construct the most diffuse, least informative prior distribution for α satisfying

$$\begin{aligned} \text{E}\left(\alpha_k\right) &= 0 \\ \text{Var}\left(\alpha_k\right) &= \sigma_k^2 \end{aligned}$$

for all k, $1 \leq k \leq K$. We do so using the *method of maximum entropy* [1, 2, 3]: our prior will be the maximum-entropy distribution satisfying the given constraints.

The entropy of a distribution is a measure of how much information the distribution provides about the variable(s) in question; the greater the entropy, the greater the uncertainty and the less informative the distribution. The entropy of a distribution with pdf $p(\alpha)$ is defined as

$$-\int p(\alpha)\log(p(\alpha)/m(\alpha))dx = -\text{E}\left(\log(p(\alpha)/m(\alpha))\right)$$

where $m(\alpha)$ is a reference measure chosen to coincide with some notion of maximal ignorance. Note that the entropy is invariant under a change of variables because both the density and the reference measure transform in the same way.

5 Form of the maximum-entropy solution

In general, the maximum-entropy distribution satisfying a set of n constraints

$$\text{E}\left(f_i(\alpha)\right) = C_i$$

has a pdf of form

$$p(\alpha) = \frac{m(\alpha)}{Z}\exp\left(-\sum_{i=1}^{n}\lambda_i f_i\left(\alpha\right)\right)$$

for some n-vector of parameter values λ and corresponding normalizing constant Z. Applying this to the problem at hand, and using the using the uniform measure $m(\alpha) = 1$, we find that the pdf for the maximum-entropy distribution on α having $\mathrm{E}(\alpha_k) = 0$ and $\mathrm{E}\left(\alpha_k^2\right) = \sigma^2$ for all k is

$$p(\alpha) \quad = \quad Z^{-1} \exp\left(-\sum_{k=1}^{K} \nu_k \alpha_k - \sum_{k=1}^{K} \lambda_k \alpha_k^2 \right) \tag{1}$$

for some choice of parameters ν_k and λ_k, and corresponding normalizing constant Z. Keep in mind that

$$\nu_K \alpha_K = -\nu_K \sum_{k=1}^{K-1} \alpha_k$$

and

$$\lambda_K \alpha_K^2 = \lambda_K \left(-\sum_{k=1}^{K-1} \alpha_k \right)^2$$

in (1).

Rather than directly solving for the parameters ν_k and λ_k, we note the following:

1. Since $\log p(\alpha)$ is quadratic in α we can complete the square to re-express $p(\alpha)$ as a multivariate normal density:

   $$\alpha \sim \mathrm{Normal}(\mu, \Sigma)$$

   for some mean vector μ and covariance matrix Σ.

2. Since $\mathrm{E}(\alpha) = \mathbf{0}$ we know that $\mu = \mathbf{0}$.

3. Our constraints are symmetric: if $\tilde{\alpha}$ is any vector obtained from α by reordering its elements, the constraints on α are equivalent to identical constraints on $\tilde{\alpha}$. Therefore the maximum-entropy distribution for α is also symmetric: the distributions for $\tilde{\alpha}$ and α are identical. That is, Σ must remain unchanged after any permutation of its rows and columns.

Item (3) implies that

- the diagonal elements of Σ are all the same; and
- the off-diagonal elements of Σ are all the same.

Combining this with the requirement that $\mathrm{Var}\left(\alpha_k\right) = \sigma^2$ for all k, we see that we must have

$$\Sigma_{jk} = \begin{cases} \sigma^2 & \text{if } j = k \\ \rho\sigma^2 & \text{if } j \neq k \end{cases}$$

for some value ρ.

6 Solving for the common covariance

At this point we have satisfied all of the constraints except for $\mathrm{Var}\left(\alpha_K\right) = \sigma^2$, and we choose ρ accordingly:

$$\begin{aligned} \mathrm{Var}\left(\alpha_K\right) &= \mathrm{Var}\left(-\sum_{k=1}^{K-1} \alpha_k\right) \\ &= \sum_{j,k=1}^{K-1} \mathrm{Cov}\left(\alpha_j, \alpha_k\right) \\ &= \sum_{k=1}^{K-1} \sigma^2 + \sum_{\substack{j,k=1 \\ j \neq k}}^{K-1} \rho\sigma^2 \\ &= \left(K-1\right)\sigma^2 + \left(K-1\right)\left(K-2\right)\rho\sigma^2 \end{aligned}$$

and so we require that

$$\left(K-1\right)\left(1 + \left(K-2\right)\rho\right) = 1.$$

A bit of algebra then gives

$$\rho = -\frac{1}{K-1}.$$

7 Verifying the solution form

Let's verify that the solution form

$$p\left(\alpha\right) = \mathrm{Normal}\left(\alpha \mid \mathbf{0}, \Sigma\right)$$

matches equation (1). Define $\Lambda = \Sigma^{-1}$. Since Σ is invariant under permutations of its rows and columns, the same is true of Λ; thus we know that

$$\Lambda_{jk} = \begin{cases} a & \text{if } j = k \\ b & \text{if } j \neq k \end{cases}$$

for *some* pair of numbers a and b. Then it is clear that

$$-\frac{1}{2}\alpha'\Lambda\alpha = -\sum_{k=1}^{K}\nu_k\alpha_k - \sum_{k=1}^{K}\lambda_k\alpha_k^2$$

if we define $\nu_k = 0$ for all k, $\lambda_K = b/2$, and $\lambda_k = (a-b)/2$ for $k \neq K$.

8 Verifying the covariance matrix

Let's verify that the matrix Σ we have defined is in fact positive definite, and hence a legitimate covariance matrix. We begin by writing Σ as

$$\Sigma = (1-\rho)\sigma^2 I + \rho\sigma^2 J$$

where I is the $(K-1)$-dimensional identity matrix and $J = \mathbf{1}\mathbf{1}'$ is the $(K-1)$-dimensional matrix of all 1's. For any $\alpha \neq \mathbf{0}$ we then have

$$\begin{aligned}
\alpha'\Sigma\alpha &= \alpha'\left((1-\rho)\sigma^2 I + \rho\sigma^2 J\right)\alpha \\
&= (1-\rho)\sigma^2\alpha'\alpha + \rho\sigma^2\alpha'\mathbf{1}\mathbf{1}'\alpha \\
&= (1-\rho)\sigma^2\left\|\alpha\right\|^2 + \rho\sigma^2\left(\alpha'\mathbf{1}\right)^2 \\
&= (1-\rho)\,\sigma^2\left\|\alpha\right\|^2 + \rho\sigma^2\left(\left\|\alpha\right\|\cdot\left\|\mathbf{1}\right\|\cdot\cos\theta\right)^2 \\
&= \left\|\alpha\right\|^2\sigma^2\left(1-\rho+\rho(K-1)\cos^2\theta\right) \\
&= \left\|\alpha\right\|^2\sigma^2\left(1+\frac{1}{K-1}-\cos^2\theta\right) \\
&\geq \left\|\alpha\right\|^2\sigma^2\frac{1}{K-1} \\
&> 0
\end{aligned}$$

where θ is the angle between the vectors α and $\mathbf{1}$. Hence Σ is positive definite.

9 Verifying symmetry

To verify that our solution is symmetric as desired, we need to show that

1. $\mathrm{Var}\left(\alpha_k\right)$ is the same for all effects α_k, and
2. $\mathrm{Cov}\left(\alpha_j, \alpha_k\right)$ is the same for all pairs of effects α_j and α_j, $j \neq k$.

Item (1) is true by construction: all effects have the same variance σ^2.

Item (2) is true by construction when $j, k \neq K$: $\mathrm{Cov}(\alpha_j, \alpha_k) = \rho\sigma^2$ in this case. It remains only to show that $\mathrm{Cov}(\alpha_j, \alpha_K) = \rho\sigma^2$ for any $j \neq K$. We have

$$\begin{aligned}
\mathrm{Cov}(\alpha_j, \alpha_K) &= \mathrm{Cov}\left(\alpha_j, -\sum_{k=1}^{K-1} \alpha_k\right) \\
&= -\sum_{k=1}^{K-1} \mathrm{Cov}(\alpha_j, \alpha_k) \\
&= -\sigma^2 - (K-2)\rho\sigma^2 \\
&= \left(-1 + \frac{K-2}{K-1}\right)\sigma^2 \\
&= \rho\sigma^2
\end{aligned}$$

and so item (2) is also true.

10 Conclusion

In summary, to achieve a symmetric prior on the effects α_k, with a prior mean of 0 and prior covariance of σ^2 for each α_k, we use the following prior on the vector of regression coefficients β:

$$\begin{aligned}
\beta &\sim \mathrm{Normal}\left(\mathbf{0}, W\Sigma W'\right) \\
W &= X^{-1} \\
\Sigma_{jk} &= \begin{cases} \sigma^2 & \text{if } j = k \\ \rho\sigma^2 & \text{if } j \neq k \end{cases} \\
\rho &= -\frac{1}{K-1}
\end{aligned}$$

where

- Σ and X are both square matrices with $K-1$ rows/columns,
- X_k (row k of X) is the encoding for level $k \neq K$,
- the encoding for level K is

$$-\sum_{k=1}^{K-1} X_k,$$

- and no row of X is a linear combination of the other rows of X.

References

[1] Jaynes, Edwin T. (1957). "Information Theory and Statistical Mechanics," *Physical Review*, Series II 106 (4): 620–630.

[2] Jaynes, Edwin T. (1957). "Information Theory and Statistical Mechanics II," *Physical Review*, Series II 108 (2): 171–190.

[3] Jaynes, Edwin T. (2003). *Probability Theory: The Logic of Science*, Cambridge University Press, pp. 351–355.

[4] Lenk, Peter and Bryan Orme (2009). "The Value of Informative Priors in Bayesian Inference with Sparse Data," *J. of Marketing Research* 46 (6): 832–845.