A default prior distribution for logistic and other regression models[*]

Andrew Gelman,[†] Aleks Jakulin,[‡] Maria Grazia Pittau,[§] and Yu-Sung Su[¶]

January 26, 2008

Abstract

We propose a new prior distribution for classical (non-hierarchical) logistic regression models, constructed by first scaling all nonbinary variables to have mean 0 and standard deviation 0.5, and then placing independent Student-t prior distributions on the coefficients. As a default choice, we recommend the Cauchy distribution with center 0 and scale 2.5, which in the simplest setting is a longer-tailed version of the distribution attained by assuming one-half additional success and one-half additional failure in a logistic regression. Cross-validation on a corpus of datasets shows the Cauchy class of prior distributions to outperform existing implementations of Gaussian and Laplace priors.

We recommend this prior distribution as a default choice for routine applied use. It has the advantage of always giving answers, even when there is complete separation in logistic regression (a common problem, even when the sample size is large and the number of predictors is small) and also automatically applying more shrinkage to higher-order interactions. This can be useful in routine data analysis as well as in automated procedures such as chained equations for missing-data imputation.

We implement a procedure to fit generalized linear models in R with the Student-t prior distribution by incorporating an approximate EM algorithm into the usual iteratively weighted least squares. We illustrate with several examples, including a series of logistic regressions predicting voting preferences, a small bioassay experiment, and an imputation model for a public health data set.

Keywords: Bayesian inference, generalized linear model, least squares, hierarchical model, linear regression, logistic regression, multilevel model, noninformative prior distribution

---

[†]Department of Statistics and Department of Political Science, Columbia University, New York, `gelman@stat.columbia.edu`, `www.stat.columbia.edu/∼gelman`

[‡]Department of Statistics, Columbia University, New York

[§]Department of Economics, University of Rome

[¶]Department of Political Science, City University of New York

1 Introduction

Nonidentifiability is a common problem in logistic regression. In addition to the problem of collinearity, familiar from linear regression, discrete-data regression can also become unstable from *separation*, which arises when a linear combination of the predictors is perfectly predictive of the outcome (Albert and Anderson, 1984, Lesaffre and Albert, 1989). Separation is surprisingly common in applied logistic regression, especially with binary predictors, and, as noted by Zorn (2005), is often handled inappropriately. For example, a common "solution" to separation is to remove predictors until the resulting model is identifiable, but, as Zorn (2005) points out, this typically results in removing the strongest predictors from the model.

An alternative approach to obtaining stable logistic regression coefficients is to use Bayesian inference. Various prior distributions have been suggested for this purpose, most notably a Jeffreys prior distribution (Firth, 1993), but these have not been set up for reliable computation and are not always clearly interpretable as prior information in a regression context. Here we propose a new, proper prior distribution that produces stable, regularized estimates while still being vague enough to be used as a default in routine applied work.

2 A default prior specification for logistic regression

A challenge in setting up any default prior distribution is getting the scale right: for example, suppose we are predicting vote preference given age (in years). We would not want the same prior distribution if the age scale were shifted to months. But discrete predictors have their own natural scale (most notably, a change of 1 in a binary predictor) that we would like to respect.

On one hand, scale-free prior distributions such as Jeffreys' do not include enough prior information; on the other, what prior information can be assumed for a generic model? Our key idea is that actual effects tend to fall within a limited range. For logistic regression, a change of 5 moves a probability from 0.01 to 0.5, or from 0.5 to 0.99. We rarely encounter situations where a shift in input x corresponds to the probability of outcome y changing from 0.01 to 0.99, hence we are willing to assign a prior distribution that assigns low probabilities to changes of 10 on the logistic scale.

2.1 Standardizing input variables to a commonly-interpretable scale

The first step of the model is to standardize the input variables (Gelman, 2007):

- Binary inputs are shifted to have a mean of 0 and to differ by 1 in their lower and upper conditions. (For example, if a population is 10% African-American and 90% other, we would define the centered "African-American" variable to take on the values 0.9 and −0.1.)

- Other inputs are shifted to have a mean of 0 and scaled to have a standard deviation of 0.5. This scaling puts continuous variables on the same scale as symmetric binary inputs (which, taking on the values ±0.5, have standard deviation 0.5).

Following Gelman and Pardoe (2007), we distinguish between regression *inputs* and *predictors*. For example, in a regression on age, sex, and their interaction, there are four predictors (the constant term, age, sex, and age × sex), but just two inputs: age and sex. It is the input variables, not the predictors, that are standardized.

2.2   A weakly informative t family of prior distributions

The second step of the model is to define prior distributions for the coefficients of the predictors. We use the Student-t family with mean 0, degrees-of-freedom parameter ν, and scale s, with ν and s chosen to provide minimal prior information to constrain the coefficients to lie in a reasonable range.

One way to pick a default value of ν and s is to consider the baseline case of one-half of a success and one-half of a failure for a single binomial trial with probability $p = \text{logit}^{-1}(\theta)$—that is, a logistic regression with only a constant term. The corresponding likelihood is $e^{\theta/2}/(1 + e^{\theta})$, which is close to a t density function with 7 degrees of freedom and scale 2.5 (Liu, 2004). We shall choose a slightly more conservative choice, the Cauchy, or t_1, distribution, again with a scale of 2.5. Figure 1 shows the three density functions: they all give preference to values less than 5, with the Cauchy allowing the occasional possibility of very large values (a point to which we return in Section 5).

We assign independent Cauchy prior distributions with center 0 and scale 2.5 to each of the coefficients in the logistic regression except the constant term. When combined with the standardization, this implies that the absolute difference in logit probability should be less then 5, when moving from one standard deviation below the mean, to one standard deviation above the mean, in any input variable.

If we were to apply this prior distribution to the constant term as well, we would be stating that the success probability is probably between 1% and 99% for units that are average in all the inputs. Depending on the context (for example, epidemiologic modeling
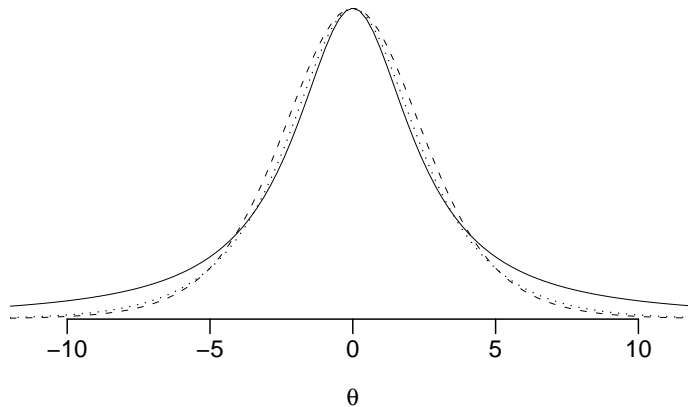
3

Figure 1: (solid line) Cauchy density function with scale 2.5, (dashed line) t_7 density function with scale 2.5, (dotted line) likelihood for θ corresponding to a single binomial trial of probability $\text{logit}^{-1}(\theta)$ with one-half success and one-half failure. All these curves favor values below 5 in absolute value; we choose the Cauchy as our default model because it allows the occasional probability of larger values.

of rare conditions, as in Greenland, 2001), this might not make sense, so as a default we apply a weaker prior distribution—a Cauchy with center 0 and scale 10, which implies that we expect the success probability for an average case to be between 10^{-9} and $1 - 10^{-9}$.

An appealing byproduct of applying the model to rescaled predictors is that it automatically implies more stringent restrictions on interactions. For example, consider three symmetric binary inputs, x_1, x_2, x_3. From the rescaling, each will take on the values $\pm 1/2$. Then any two-way interaction will take on the values $\pm 1/4$, and the three-way interaction can be $\pm 1/8$. But all these coefficients have the same default prior distribution, so the total contribution of the three-way interaction (for example) is $1/4$ that of the main effect. Going from the low value to the high value in any given three-way interaction is, in the model, unlikely to change the logit probability by more than $5 \cdot (1/8 - (-1/8)) = 5/4$ on the logit scale.

3   Computation

In principle, logistic regression with our prior distribution can be computed using the Gibbs and Metropolis algorithms. We do not give details as this is now standard with Bayesian models (see, for example, Carlin and Louis, 2001, Martin and Quinn, 2002, and Gelman et al., 2003). In practice, however, it is desirable to have a quick calculation that returns a point estimate of the regression coefficients and standard errors. Such an approximate calculation works in routine statistical practice and, in addition, recognizes the approximate

nature of the model itself.

We consider three computational settings:

- Classical (non-hierarchical) logistic regression, using our default prior distribution in place of the usual flat prior distribution on the coefficients.

- Multilevel (hierarchical) modeling, in which some the default prior distribution is applied only to the subset of the coefficients that are not otherwise modeled (sometimes called the "fixed effects").

- Chained imputation, in which each variable with missing data is modeled conditional on the other variables with a regression equation, and these models are fit and random imputations inserted iteratively (Van Buuren and Oudshoom, 2000, Raghunathan, Van Hoewyk, and Solenberger, 2001).

In any of these cases, our default prior distribution has the purpose of stabilizing (regularizing) the estimates of otherwise unmodeled parameters. In the first scenario, we typically only want point estimates and standard errors (unless the sample size is so small that the normal approximation to the posterior distribution is inadequate). In the second scenario, it makes sense to embed the computation within the full Markov chain simulation. In the third scenario of missing-data imputation, we would like the flexibility of quick estimates for simple problems with the potential for Markov chain simulation as necessary. Also, because of the automatic way in which the component models are fit in a chained imputation, we would like a computationally stable algorithm that returns reasonable answers.

We have implemented these computations by altering the `glm` function in R, creating a new function, `bayesglm`, which finds an approximate posterior mode and variance using extensions of the classical generalized linear model computations, as described in the rest of this section. The `bayesglm` function (part of the `arm` package in R) allows the user to specify independent prior distributions for the coefficients in the t family, with the default being Cauchy distributions with center 0 and scale set to 10 (for the regression intercept), 2.5 (for binary predictors), or $2.5/(2 \cdot \text{sd})$, where `sd` is the standard deviation of the predictor in the data (for other numerical predictors). We are also extending the program to fit hierarchical models in which regression coefficients are structured in batches (Gelman et al., 2007).

3.1 Incorporating the prior distribution into classical logistic regression computations

Working in the context of the logistic regression model,

$$\Pr(y_i = 1) = \text{logit}^{-1}(X_i\beta), \tag{1}$$

we adapt the classical maximum likelihood algorithm to obtain approximate posterior inference for the coefficients β, in the form of an estimate $\hat{\beta}$ and covariance matrix V_β.

The standard logistic regression algorithm—upon which we build—proceeds by approximately linearizing the derivative of the log-likelihood, solving using weighted least squares, and then iterating this process, each step evaluating the derivatives at the latest estimate $\hat{beta}$ (see, for example, McCullagh and Nelder, 1989). At each iteration, the algorithm determines pseudo-data z_i and psuedo-variances $(\sigma_i^z)^2$ based on the linearization of the derivative of the log-likelihood:

$$
\begin{aligned}
z_i &= X_i\hat{\beta} + \frac{(1 + e^{X_i\hat{\beta}})^2}{e^{X_i\hat{\beta}}}\left(y_i - \frac{e^{X_i\hat{\beta}}}{1 + e^{X_i\hat{\beta}}}\right) \\
(\sigma_i^z)^2 &= \frac{1}{n_i}\frac{(1 + e^{X_i\hat{\beta}})^2}{e^{X_i\hat{\beta}}}.
\end{aligned}
\tag{2}
$$

and then performs weighted least squares, regressing z on X with weight vector $(\sigma^z)^{-2}$. The resulting estimate $\hat{\beta}$ is used to update the computations in (2), and the iteration proceeds until approximate convergence.

Computation with a specified normal prior distribution

The simplest informative prior distribution assigns normal prior distributions for the components of β:

$$\beta_j \sim \text{N}(\mu_j, \sigma_j^2), \quad \text{for } j = 1, \dots, J.$$

This information can be effortlessly included in the classical algorithm by simply altering the weighted least-squares step, augmenting the approximate likelihood with the prior distribution (see, for example, Section 14.8 of Gelman et al., 2003). If the model has J coefficients β_j with independent $\text{N}(\mu_j, \sigma_j^2)$ prior distributions, then we add J pseudo-data points and perform weighted linear regression on "observations" z_*, "explanatory variables" X_*, andweight vector w_*, where

$$z_* = \begin{pmatrix} z \\ \mu \end{pmatrix}, \quad X_* = \begin{pmatrix} X \\ I_J \end{pmatrix}, \quad w_* = (\sigma^z, \sigma)^{-2}. \tag{3}$$

6

Here, z_* and w_* are vectors of length $n+J$ and X_* is a $(n+J) \times J$ matrix. With the augmented X_*, this regression is identified, and thus the resulting estimate $\hat{beta}$ is well defined and has finite variance, even if the original data have collinearity or separation that would result in nonidentifiability of the maximum likelihood estimate.

The full computation is then iteratively weighted least squares, starting with a guess of β (for example, independent draws from the unit normal distribution), then computing the derivatives of the log-likelihood to compute z and σ_z, then using weighted least squares on the pseudodata (3) to yield an updated estimate of β, then recomputing the derivatives of the log-likelihood at this new value of β, and so forth, converging to the estimate $\hat{\beta}$. The covariance matrix V_β is simply the inverse second derivative matrix of the log-posterior density evaluated at $\hat{\beta}$—that is, the usual normal-theory uncertainty estimate for an estimate not on the boundary of parameter space.

Approximate EM algorithm with a t prior distribution

If the coefficients β_j have t prior distributions with centers μ_j and scales s_j,[1] we can program a similar procedure, using the formulation

$$\beta_j \sim \mathrm{N}(\mu_j, \sigma_j^2), \quad \sigma_j^2 \sim \text{Inv-}\chi^2(\nu_j, s_j^2) \tag{4}$$

and averaging over the β_j's at each step, treating them as missing data and performing one step of the EM algorithm to estimate the σ_j's. Once enough iterations have been performed to reach approximate convergence, we get an estimate and covariance matrix for the vector parameter β the estimated σ_j's.

We initialize the algorithm by setting each σ_j to the value s_j (the scale of the prior distribution) and, as before, starting with a guess of β. Then, at each step of the algorithm, we update σ by maximizing the expected value of its (approximate) log-posterior density,

$$
\begin{aligned}
\log p(\beta, \sigma | y) \quad \approx \quad & -\frac{1}{2} \sum_{i=1}^{n} \frac{1}{(\sigma_i^z)^2} (z_i - X_i\beta)^2 - \frac{1}{2} \sum_{j=1}^{J} \left(\frac{1}{\sigma_j^2} (\beta_j - \mu_j)^2 + \log(\sigma_j^2) \right) \\
& - p(\sigma_j | \nu_j, s_j) + \text{constant}.
\end{aligned} \tag{5}
$$

Each iteration of the algorithm proceeds as follows:

1. Based on the current estimate of β, perform the normal approximation to the log-likelihood and determine the vectors z and σ^z using (2), as in classical logistic regression computation.

---

[1]As discussed earlier, we set $\mu_j = 0$, $s_j = 2.5$, $\nu_j = 1$ as a default, but we describe the computation more generally in terms of arbitrary values of these parameters.

2. Approximate E-step: first run the weighted least squares regression based on the augmented data (3) to get an estimate $\hat{\beta}$ with variance matrix V_β. Then determine the expected value of the log-posterior density by replacing the terms $(\beta_j - \mu_j)^2$ in (5) by

$$\mathrm{E}\left((\beta_j - \mu_j)^2 | \sigma, y\right) \approx (\hat{\beta}_j - \mu_j)^2 + (V_\beta)_{jj}, \qquad (6)$$

which is only approximate because we are averaging over a normal distribution that is only an approximation to the generalized linear model likelihood.

3. M-step: maximize the (approximate) expected value of the log-posterior density (5) to get the estimate,

$$\hat{\sigma}_j^2 = \frac{(\hat{\beta}_j - \mu_j)^2 + (V_\beta)_{jj} + \nu_j s_j^2}{1 + \nu_j}, \qquad (7)$$

which corresponds to the (approximate) posterior mode of σ_j^2 given a single measurement with value (6) and an Inv-$\chi^2(\nu_j, s_j^2)$ prior distribution.

4. Recompute the derivatives of the log-posterior density given the current $\hat{\beta}$, set up the augmented data (3) using the estimated $\hat{\sigma}$ from (7), and repeat steps 1,2,3 above.

At convergence of the algorithm, we summarize the inferences using the latest estimate $\hat{\beta}$ and covariance matrix V_β.

4   Examples

4.1   A series of regressions predicting vote preferences

Regular users of logistic regression know that separation can occur in routine data analyses, even when the sample size is large and the number of predictors is small. The left column of Figure 2 shows the estimated coefficients for logistic regression predicting probability of Republican vote for President for a series of elections. The estimates look fine except in 1964, where there is complete separation, with all black respondents supporting the Democrats. (Fitting in R actually yields finite estimates, as displayed in the graph, but these are essentially meaningless, being a function of how long the iterative fitting procedure goes before giving up.)

The other three columns of Figure 2 show the coefficient estimates using our default Cauchy prior distribution for the coefficients, along with the t_7 and normal distributions. (In all cases, the prior distributions are centered at 0, with scale parameters set to 10 for the constant term and 2.5 for all other coefficients.) All three prior distributions do a reasonable
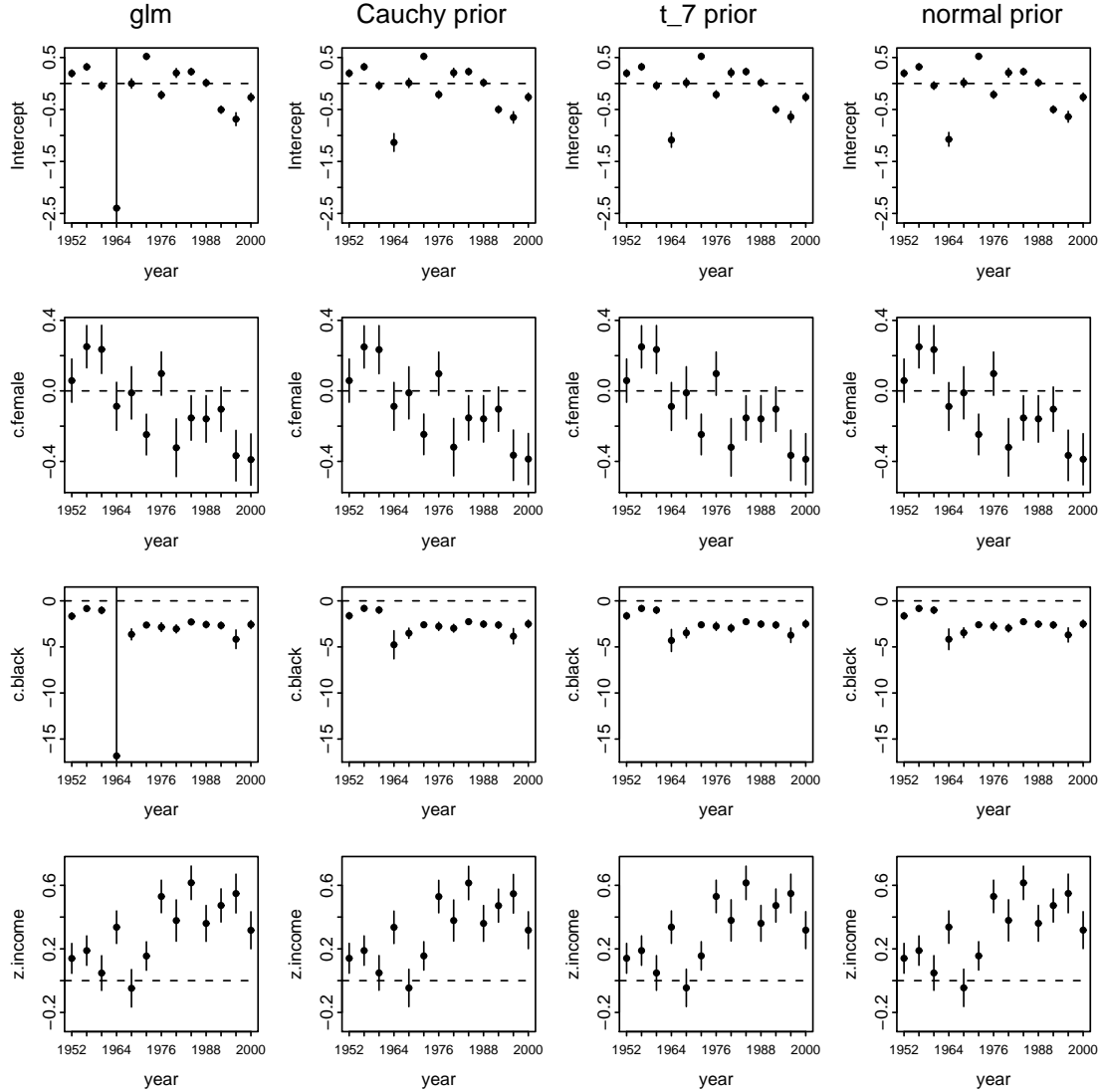
Figure 2: The left column shows the estimated coefficients (± 1 standard error) for a logistic regression predicting probability of Republican vote for President given sex, race, and income, as fit separately to data from the National Election Study for each election 1952 through 2000. (The binary inputs `female` and `black` have been centered to have means of zero, and the numerical variable `income` (originally on a 1–5 scale) has been centered and then rescaled by dividing by two standard deviations.)

There is complete separation in 1964 (with none of black respondents supporting the Republican candidate, Barry Goldwater), leading to a coefficient estimate of $-\infty$ that year. (The particular finite values of the estimate and standard error are determined by the number of iterations used by `glm` function in R before stopping.)

(other columns) Estimated coefficients (± 1 standard error) for the same model fit each year using independent Cauchy, t_7, and normal prior distributions, each with center 0 and scale 2.5. All three prior distributions do a reasonable job at stabilizing the estimates for 1964, while leaving the estimates for other years essentially unchanged.

9

Dose, x_i (log g/ml)	Number of animals, n_i	Number of deaths, y_i
−0.86	5	0
−0.30	5	1
−0.05	5	3
0.73	5	5

```
# from glm:
            coef.est coef.se
(Intercept) -0.1      0.7
z.x         10.2      6.4
  n = 4, k = 2
  residual deviance = 0.1, null deviance = 15.8 (difference = 15.7)

# from bayesglm (Cauchy priors, scale 10 for const and 2.5 for other coef):
            coef.est coef.se
(Intercept) -0.2      0.6
z.x          5.4      2.2
  n = 4, k = 2
  residual deviance = 1.1, null deviance = 15.8 (difference = 14.7)
```

Figure 3: Data from a bioassay experiment, from Racine et al. (1986), and estimates from classical maximum likelihood and Bayesian logistic regression with the recommended default prior distribution. We show raw computer output to illustrate how our approach would be used in routine practice.

The big change in the estimated coefficient for z.x when going from glm to bayseglm may seem surprising at first, but upon reflection we prefer the second estimate with its lower coefficient for x, which is based on downweighting the most extreme possibilities that are allowed by the likelihood.

job at stabilizing the estimated coefficient for race for 1964, while leaving the estimates for other years essentially unchanged. This example illustrates how we could use our Bayesian procedure in routine practice.

4.2 A small bioassay experiment

We next consider a small-sample example in which the prior distribution makes a difference for a coefficient that is already identified. The example comes from Racine et al. (1986), who used a problem in bioassay to illustrate how Bayesian inference can be applied with small samples. The top part of Figure 3 presents the data, from twenty animals that were exposed to four different doses of a toxin. The bottom parts of Figure 3 show the resulting logistic regression, as fit first using maximum likelihood and then using our default Cauchy prior distributions with center 0 and scale 10 (for the constant term) and 2.5 (for the coefficient of dose). Following our general procedure, we have rescaled dose to have mean 0 and standard

deviation 0.5.

With such a small sample, the prior distribution actually makes a difference, lowering the estimated coefficient of standardized dose from 10.2 ± 6.4 to 5.4 ± 2.2. Such a large change might seem disturbing, but for the reasons discussed above, we would doubt the effect to be as large as 10.2 on the logistic scale, and the analysis shows these data to be consistent with the much smaller effect size of 5.4. The large amount of shrinkage simply confirms how weak the information is that gave the original maximum likelihood estimate.

4.3   A set of chained regressions for missing-data imputation

Multiple imputation (Rubin, 1987, 1996) is another context in which regressions with many predictors are fit in an automatic way. Van Buuren and Oudshoom (2000) and Raghunathan, Van Hoewyk, and Solenberger (2001) discuss implementations of the chained equation approach, in which variables with missingness are imputed one at a time, each conditional on the imputed values of the other variables, in an iterative random process that is used to construct multiple imputations. In chained equations, logistic regressions or similar models can be used to impute binary variables, and when the number of variables is large, separation can arise. Our prior distribution yields stable computations in this setting, as we illustrate in with example from our current applied research.

Separation occurred in the case of imputing virus loads in a longitudinal sample of HIV-positive homeless persons (Messeri et al., 2006). The imputation analysis incorporated a large number of predictors, including demographic and health-related variables, and often with high rates of missingness. Inside the multiple imputation chained equation procedure, logistic regression was used to impute the binary variables. It is generally recommended to include a rich set of predictors when imputing missing values (Rubin, 1996). However, in this example, including all the dichotomous predictors leads to many instances of separation.

For one example from our analysis, separation arose when estimating, for each HIV-positive persons in the sample, the probability of attendance in a group therapy called `haart`. The top part of Figure 4 shows the model as estimated using the `glm` function in R fit to the observed cases in the first year of the data set: the coefficient for `nonhaartcombo.W1` is essentially infinity, and the regression also gives an error message indicating nonidentifiability. The bottom part of Figure 4 shows the fit using our recommended Bayesian procedure (this time, for simplicity, not recentering and rescaling the inputs, most of which are actually binary).

In the chained imputation procedure, the classical `glm` fits were nonidentifiable at many

```
# from glm:
                coef.est coef.sd                 coef.est coef.sd
(Intercept)         0.07    1.41   h39b.W1          -0.10 0.03
age.W1              0.02    0.02   pcs.W1           -0.01 0.01
mcs37.W1           -0.01    0.32   nonhaartcombo.W1 -20.99 888.74
unstabl.W1         -0.09    0.37   b05.W1           -0.07 0.12
ethnic.W3          -0.14    0.23   h39b.W2           0.02 0.03
age.W2              0.02    0.02   pcs.W2           -0.01 0.02
mcs37.W2            0.26    0.31   haart.W2          1.80 0.30
nonhaartcombo.W2    1.33    0.44   unstabl.W2        0.27 0.42
b05.W2              0.03    0.12   h39b.W3           0.00 0.03
age.W3             -0.01    0.02   pcs.W3            0.01 0.01
mcs37.W3           -0.04    0.32   haart.W3          0.60 0.31
nonhaartcombo.W3    0.44    0.42   unstabl.W3       -0.92 0.40
b05.W3             -0.11    0.11
  n = 508, k = 25
  residual deviance = 366.4, null deviance = 700.1 (difference = 333.7)


# from bayesglm (Cauchy priors, scale 10 for const and 2.5 for other coefs):
                coef.est coef.sd                 coef.est coef.sd
(Intercept)        -0.84    1.15   h39b.W1          -0.08 0.03
age.W1              0.01    0.02   pcs.W1           -0.01 0.01
mcs37.W1           -0.10    0.31   nonhaartcombo.W1 -6.74 1.22
unstabl.W1         -0.06    0.36   b05.W1            0.02 0.12
ethnic.W3           0.18    0.21   h39b.W2           0.01 0.03
age.W2              0.03    0.02   pcs.W2           -0.02 0.02
mcs37.W2            0.19    0.31   haart.W2          1.50 0.29
nonhaartcombo.W2    0.81    0.42   unstabl.W2        0.29 0.41
b05.W2              0.11    0.12   h39b.W3          -0.01 0.03
age.W3             -0.02    0.02   pcs.W3            0.01 0.01
mcs37.W3            0.05    0.32   haart.W3          1.02 0.29
nonhaartcombo.W3    0.64    0.40   unstabl.W3       -0.52 0.39
b05.W3             -0.15    0.13
```

Figure 4: A logistic regression fit for missing-data imputation using maximum likelihood (top) and Bayesian inference with default prior distribution (bottom). The classical fit resulted in an error message indicating separation; in contrast, the Bayes fit (using independent Cauchy prior distributions with mean 0 and scale 10 for the intercept and 2.5 for the other coefficients) produced stable estimates. We would not usually summarize results using this sort of table; however, this gives a sense of how the fitted models look in routine data analysis.

places; none of these presented any problem when we switched to our new `bayesglm` function.[2]

5 Data from a large number of logistic regressions

In the spirit of Stigler (1977), we wanted to see how large are logistic regression coefficients in some general population, to get a rough sense of what would be a reasonable default prior distribution. One way to do this is to fit many logistic regressions to available data sets and estimate the underlying distribution of coefficients. Another approach is to examine the cross-validated predictive quality of different types of priors on a corpus of data sets, following the approach of meta-learning in computer science (e.g., Vilalta and Drissi, 2001).

5.1 Cross-validation on a corpus of data sets

The fundamental idea of predictive modeling is that the data are split into two subsets, the training and the test data. The training data are used to construct a model, and the performance of the model on the test data is used to check whether the predictions generalize well. Cross-validation is a way of creating several different partitions. For example, assume that we put aside 1/5 of the data for testing. We divide up the data into 5 pieces of the same size. This creates 5 different partitions, and for each experiment we take one of the pieces as test set and all the others as the training set. In the present section we summarize our efforts in evaluating our prior distribution from the predictive perspective.

Because we can summarize the performance in a single number for a whole data set (using the expected squared error or expected log error), we can work with a larger collection of data sets, as is customary in machine learning. For our needs we have taken a number of data sets from the UCI Machine Learning Repository (Newman et al., 1998), disregarding those whose outcome is a continuous variable (such as "anonymous Microsoft Web data") and those that are given in form of logical theories (such as "artificial characters"). Figure 5 summarized the datasets we used for our cross-validation.

Because we do not want our results to depend on an imputation method, we add an additional predictor for each variable with missing data indicating whether the particular predictor's value is missing. We also use the Fayyad and Irani (1993) method for converting

---

[2]We also tried the `brlr` function in R, which implements the Jeffreys prior distribution of Firth (1993). Unfortunately, we still encountered problems in achieving convergence and obtaining reasonable answers, several times obtaining an error message indicating nonconvergence of the optimization algorithm. We suspect this problem arises because `brlr` uses a general-purpose optimization algorithm that, when fitting regression models, is less stable than iteratively weighted least squares.

Name	Cases	Num	Cat	Pred	Outcome	$\mathbf{Pr}(y=1)$	$\mathbf{Pr(NA)}$	$\overline{|\vec{x}|}$
adult	32561	6	8	133	y=0	0.76	0.01	2.4
mushroom	8124	0	22	95	edible=e	0.52	0	3.0
spam	4601	57	0	105	class=0	0.61	0	3.2
krkp	3196	0	36	37	result=won	0.52	0	2.6
segment	2310	19	0	154	y=5	0.14	0	3.5
titanic	2201	0	3	5	surv=no	0.68	0	0.7
car	1728	0	6	15	eval=unacc	0.70	0	2.0
cmc	1473	2	7	19	Contracept=1	0.43	0	1.9
german	1000	7	13	48	class=1	0.70	0	2.8
tic-tac-toe	958	0	9	18	y=p	0.65	0	2.3
heart	920	7	6	30	num=0	0.45	0.15	2.3
anneal	898	6	32	64	y=3	0.76	0.65	2.4
vehicle	846	18	0	58	Y=3	0.26	0	3.0
pima	768	8	0	11	class=0	0.65	0	1.8
crx	690	6	9	45	A16=-	0.56	0.01	2.3
australian	690	6	8	36	Y=0	0.56	0	2.3
soybean-large	683	35	0	75	y=brown-spot	0.13	0.10	3.2
breast-wisc-c	683	9	0	20	y=2	0.65	0	1.6
balance-scale	625	0	4	16	name=L	0.46	0	1.8
monk2	601	0	6	11	y=0	0.66	0	1.9
wdbc	569	20	0	45	diag=B	0.63	0	3.0
monk1	556	0	6	11	y=0	0.50	0	1.9
monk3	554	0	6	11	y=1	0.52	0	1.9
voting	435	0	16	32	party=dem	0.61	0	2.7
horse-colic	369	7	19	121	outcom=1	0.61	0.20	3.4
ionosphere	351	32	0	110	y=g	0.64	0	3.5
bupa	345	6	0	6	selector=2	0.58	0	1.5
primary-tumor	339	0	17	25	primary=1	0.25	0.04	2.0
ecoli	336	7	0	12	y=cp	0.43	0	1.3
breast-LJ-c	286	3	6	16	recurrence=no	0.70	0.01	1.8
shuttle-control	253	0	6	10	y=2	0.57	0	1.8
audiology	226	0	69	93	y=cochlear-age	0.25	0.02	2.3
glass	214	9	0	15	y=2	0.36	0	1.7
yeast-class	186	79	0	182	func=Ribo	0.65	0.02	4.6
wine	178	13	0	24	Y=2	0.40	0	2.2
hayes-roth	160	0	4	11	y=1	0.41	0	1.5
hepatitis	155	6	13	35	Class=LIVE	0.79	0.06	2.5
iris	150	4	0	8	y=virginica	0.33	0	1.6
lymphography	148	2	16	29	y=2	0.55	0	2.5
promoters	106	0	57	171	y=mm	0.50	0	6.1
zoo	101	1	15	17	type=mammal	0.41	0	2.2
post-operative	88	1	7	14	ADM-DECS=A	0.73	0.01	1.6
soybean-small	47	35	0	22	y=D4	0.36	0	2.6
lung-cancer	32	0	56	103	y=2	0.41	0	4.3
lenses	24	0	4	5	lenses=none	0.62	0	1.4
o-ring-erosion	23	3	0	4	no-therm-d=0	0.74	0	0.7

Figure 5: The 46 datasets from the UCI Machine Learning data repository which we used for our cross-validation. Each dataset is described with its name, the number of cases in it (Cases), the number of numerical attributes (Num), the number of categorical attributes (Cat), the number of binary predictors generated from the initial set of attributes by means of discretization (Pred), the event corresponding to the positive binary outcome (Outcome), the percentage of cases having the positive outcome ($p_{y=1}$), the proportion of attribute values that were missing, expressed as a percentage (NA), and the average length of the predictor vector, ($\overline{|\vec{x}|}$).
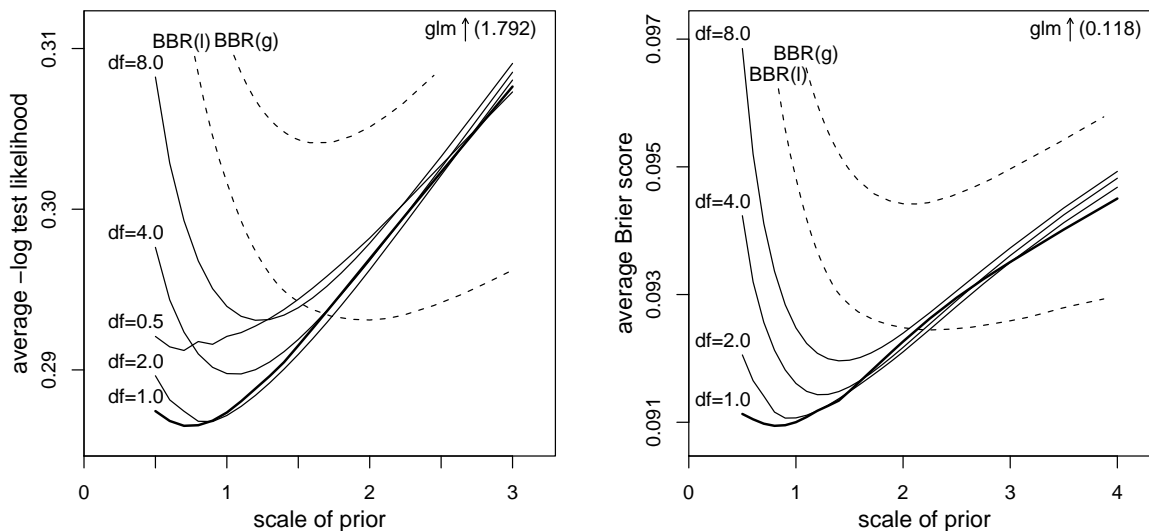
Figure 6: Mean logarithmic score (left plot) and Brier score (right plot), in fivefold cross-validation averaging over the data sets in the UCI corpus, for different independent prior distributions for logistic regression coefficients. Higher value on the y axis indicates a larger error. Each line represents a different degrees-of-freedom parameter for the Student-t prior family. BBR(l) indicates the Laplace prior with the BBR algorithm of Genkin, Lewis, and Madigan (2007), and BBR(g) represents the Gaussian prior. The Cauchy prior distribution with scale 0.75 performs best, while the performance of `glm` (shown in the upper-right corner) is so bad that we could not capture it on our scale. The scale axis corresponds to the square root of variance for the normal and the Laplace distribution.

continuous predictors into discrete ones. To convert a k-level predictor into a set of binary predictors, we create $k − 1$ predictors corresponding to all levels except the most frequent. Finally, for all data sets with multinomial outcomes, we transform into binary by simply comparing the most frequent category to the union of all the others.

5.2 Average predictive errors corresponding to different prior distributions

We use fivefold cross-validation to compare "bayesglm" (our approximate Bayes point estimate) for different default scale and degrees of freedom parameters; recall that degrees of freedom equal 1 and ∞ for the Cauchy and Gaussian prior distributions, respectively. We also compare to classical logistic regression (that is, bayesglm with prior scale set to ∞) and to the BBR (Bayesian binary regression) algorithm of Genkin, Lewis, and Madigan (2007), which adaptively sets the scale for the choice of Laplacian or Gaussian prior distribution.
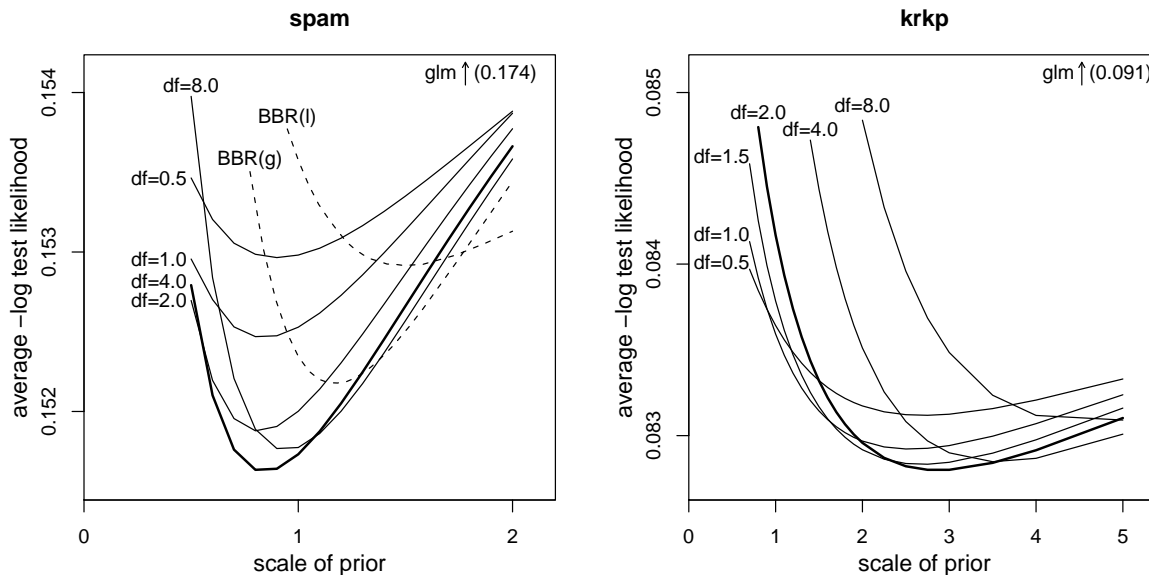
15

Figure 7: Mean logarithmic score for two datasets, "Spam" and "KRKP," from the UCI database. The curves show average cross-validated log-likelihood for estimates based on t prior distributions with different degrees of freedom and different scales. For the "spam" data, the t_4 with scale 0.8 is optimal, whereas for the "krkp" data, the t_2 with scale 2.8 performs best under cross-validation.

Figure 6 shows the results, displaying average logarithmic and Brier score losses for different choices of prior distribution.[3] The Cauchy prior distribution with scale 0.75 performs best, on average. Classical logistic regression ("glm"), which corresponds to prior degrees of freedom and prior scale both set to ∞, performs the worst: with no regularization, maximum likelihood occasionally gives extreme estimates, which then result in large penalties in the cross-validation. In fact, the log and Brier scores for classical logistic regression would be even worse except that the `glm` function in R stops after a finite number of iterations, thus giving estimates that are less extreme than they would otherwise be.

The Cauchy prior distribution with scale 0.75 is a good consensus choice, but for any particular dataset, other prior distributions can perform better. To illustrate, Figure 7 shows the cross-validation errors for individual data sets in the corpus for the Cauchy prior distribution with different choices of the degrees-of-freedom and scale parameter. The Cauchy (that is, t_1 with scale 1) performs reasonably well in both cases, and much better

---

[3]Given the vector of predictors $\vec{x}$, the true outcome y and the predicted probability $p_y = f(\vec{x})$ for y, the Brier score is defined as $(1 - p_y)^2/2$ and the logarithmic score is defined as $-\log p_y$. Because of cross-validation, the probabilities were built without using the predictor-outcome pairs $(\vec{x}, y)$, so we are protected against overfitting.

16

than classical glm, but the optimal prior distribution is difference for each particular dataset.

5.3 Choosing a weakly-informative prior distribution

The Cauchy prior distribution with scale 0.75 performs the best, yet we recommend as a default a larger scale of 2.5. Why? The argument is that, following the usual principles of noninformative or weakly informative prior distributions, we are including in our model less information than we actually have. This approach is generally considered "conservative" in statistical practice (Gelman and Jakulin, 2007). In the case of logistic regression, the evidence suggests that the Cauchy distribution with scale 0.75 captures the underlying variation in logistic regression coefficients in a corpus of data sets. We use a scale of 2.5 to weaken this prior information and bring things closer to the traditional default choice of maximum likelihood. True logistic regression coefficients are almost always quite a bit less than 5 (if predictors have been standardized), and so this Cauchy distribution actually contains less prior information than we really have. From this perspective, the uniform prior distribution is the most conservative, but sometimes too much so (in particular, for datasets that feature separation, coefficients have maximum likelihood estimates of infinity), and this new prior distribution is still somewhat conservative, thus defensible to statisticians. Any particular choice of prior distribution is arbitrary; we have motivated ours based on the notion that extremely large coefficients are unlikely, and as a longer-tailed version of the model corresponding to one-half success and one-half failure, as discussed in Section 2.2.

The BBR procedure of Genkin, Lewis, and Madigan (adapted from the regularization algorithm of Zhang and Oles, 2001) employs a heuristic for determining the scale of the prior: the scale corresponds to $k/\mathrm{E}[\vec{x}\dot{\vec{x}}]$ where k is the number of dimensions in $\vec{x}$. This heuristic assures some invariance with respect to the scaling of the input data. All the predictors in our experiments took either the value of 0 or of 1, and we did not perform additional scaling. The average value of the heuristic across the datasets was approximately 2.0, close to the optimum. However, the heuristic scale for individual datasets resulted in worse performance than using the corpus optimum. We interpret this observation as supporting our corpus-based approach for determining the parameters of the prior.

6 Discussion

We recommend using, as a default prior model, independent Cauchy distributions on all logistic regression coefficients, each centered at 0 and with scale parameter 10 for the constant term and 2.5 for all other coefficients. Before fitting this model, we center each binary

input to have mean 0 and rescale each numeric input to have mean 0 and standard deviation 0.5. When applying this procedure to classical logistic regression, we fit the model using an adaptation of the standard iteratively weighted least squares computation, using the posterior mode as a point estimate and the curvature of the log-posterior density to get standard errors. More generally, the prior distribution can be used as part of a fully Bayesian computation in more complex settings such as hierarchical models.

6.1 Other models

This paper has focused on logistic regression, but the same idea could be used for other models.

Linear regression. Our algorithm is basically the same for linear regression, except that weighted least squares is an exact rather than approximate maximum penalized likelihood, and also a step needs to be added to estimate the data variance. In addition, we would preprocess y by rescaling the outcome variable to have mean 0 and standard deviation 0.5 before assigning the prior distribution (or, equivalently, multiply the prior scale parameter by the standard deviation of the data).

Other generalized linear models. Again, the same algorithm is unchanged, except that the pseudo-data and pseudo-variances in (2), which are derived from the first and second derivatives of the log-likelihood, are changed (see Section 16.4 of Gelman et al., 2003). For Poisson regression and other models with the logarithmic link, we would not often expect effects larger than 5 on the logarithmic scale, and so the prior distributions given in this article would seem like a reasonable default choice. In addition, for models such as the negative binomial that have dispersion parameters, these can be estimated using an additional step as is done when estimating the data-level variance in normal linear regression.

Multilevel (hierarchical) modeling. Although not the main topic of this paper, hierarchical logistic regression models can be fit in a similar approximate manner using an extension of the above EM algorithm to average over hyperparameters; see Gelman et al. (2007).

Avoiding nested looping when inserting into larger models. In more elaborate multilevel models or in applications such as chained imputation (discussed in Section 4.3),

18

it should be possible to speed the computation by threading, rather than nesting, the loops. For example, suppose we are fitting an imputation by iteratively regressing u on v, w, then v on u, w, then w on u, v. Instead of doing a full iterative weighted least squares at each iteration, then we could perform one step of weighted least squares at each step, thus taking less computer time to ultimately converge by not wasting time by getting hyper-precise estimates at each step of the stochastic algorithm.

6.2   Related work

Our key idea is to use minimal prior knowledge, specifically that a typical change in an input variable would be unlikely to correspond to a change as large as 5 on the logistic scale (which would move the probability from 0.01 to 0.50 or from 0.50 to 0.99). This is related to the method of Bedrick, Christensen, and Johnson (1996) of setting a prior distribution by eliciting the possible distribution of outcomes given different combinations of regression inputs, and the method of Witte, Greenland, and Kim (1998) and Greenland (2001) of assigning prior distributions by characterizing expected effects in weakly informative ranges ("probably near null," "probably moderately positive," and so on). Our method differs from these related approaches in being more of a generic prior constraint rather than information specific to a particular analysis. As such, we would expect our prior distribution to be more appropriate for automatic use, with these other methods suggesting ways to add more targeted prior information when necessary. One approach for going further, discussed by MacLehose et al. (2006) and Dunson, Herring, and Engel (2006), is to use mixture prior distributions for logistic regressions with large numbers of predictors. These models use batching in the parameters, or attempt to discover such batching, in order to identify more important predictors and shrink others.

Another area of related work is the choice of parametric family for the prior distribution. We have chosen the t family, focusing on the Cauchy as a conservative choice. Genkin, Lewis, and Madigan (2007) consider the Laplace (double-exponential) distribution, which has the property that its posterior mode estimates can be shrunk all the way to zero. This is an appropriate goal in projects such as text categorization (the application in that article) in which data storage is an issue, but is less relevant in social science analysis of data that have already been collected.

In the other direction, our approach (which, in the simplest logistic regression that includes only a constant term, is close to adding one-half success and one-half failure; see Figure 1) can be seen as a generalization of the work of Agresti and Coull (1988) on using

19

Bayesian techniques to get point estimates and confidence intervals with good small-sample frequency properties. As we have noted earlier, similar penalized likelihood methods using the Jeffreys prior have been proposed by Firth (1993), Heinze and Schemper (2003), and Zorn (2005); Heinze (2006) evaluates the frequency properties of estimates and tests using method. Our approach is similar but is parameterized in terms of the coefficients and thus allows us to make use of prior knowledge on that scale. In simple cases the two methods can give similar results (for example, identical to the first decimal place in the example in Figure 3), with our algorithm being more stable by taking advantage of the existing iteratively weighted least squares algorithm.

6.3 Concerns

A theoretical concern is that our prior distribution is defined on centered and scaled input variables; thus it implicitly depends on the data. As more data arrive, the linear transformations used in the centering and scaling will change, thus changing the implied prior distribution as defined on the original scale of the data. A natural extension here would be to formally make the procedure hierarchical, for example defining the j-th input variable x_{ij} as having a population mean μ_j and standard deviation σ_j, then defining the prior distributions for the corresponding predictors in terms of scaled inputs of the form $z_{ij} = (x_{ij} - \mu_j)/(2\sigma_j)$. We did not go this route, however, because modeling all the input variables corresponds to a potentially immense effort which is contrary to the spirit of this method, which is to be a quick automatic solution. In practice, we do not see the dependence of our prior distribution on data as a major concern, although we imagine it could cause difficulties when sample sizes are very small.

Modeling the coefficient of a scaled variable is analogous to parameterizing a simple regression through the correlation, which depends on the distribution of x as well as the regression of y on x. Changing the values of x can change the correlation, and thus the implicit prior distribution, even though the regression is not changing at all (assuming an underlying linear relationship). That said, this is the cost of having an informative prior distribution: some scale must be used, and the scale of the data seems like a reasonable default choice.

Finally, one might argue that the Bayesian procedure, by always giving an estimate, obscures nonidentifiability and could lead the user into a false sense of security. To this objection we would reply (following Zorn, 2005): first, one is always free to also fit using maximum likelihood, and second, separation corresponds to information in the data, which

is ignored if the offending predictor is removed and awkward to handle if it is included with an infinite coefficient (see, for example, the estimates for 1964 in the first column of Figure 2). Given that we do not expect to see effects as large as 10 on the logistic scale, it is appropriate to use this information. As we have seen in specific examples and also in the corpus of datasets, this weakly-informative prior distribution yields estimates that make more sense and perform better predictively, compared to maximum likelihood, which is still the standard approach for routine logistic regression in theoretical and applied statistics.

References

Agresti, A., and Coull, B. A. (1998). Approximate is better than exact for interval estimation of binomial proportions. *American Statistician* **52**, 119–126.

Albert, A., and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**, 1–10.

Bedrick, E. J., Christensen, R., and Johnson, W. (1996). A new perspective on priors for generalized linear models. *Journal of the American Statistical Association* **91**, 1450–1460.

Carlin, B. P., and Louis, T. A. (2001). *Bayes and Empirical Bayes Methods for Data Analysis*, second edition. London: CRC Press.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* **39**, 1–38.

Dunson, D. B., Herring, A. H., and Engel, S. M. (2006). Bayesian selection and clustering of polymorphisms in functionally-related genes. *Journal of the American Statistical Association*, under revision.

Fayyad, U. M., and Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-93*. Chambery, France: Morgan Kauffman.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 514–534.

Gelman, A. (2007). Scaling regression inputs by dividing by two standard deviations. Technical report, Department of Statistics, Columbia University.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, second edition. London: CRC Press.

Gelman, A., and Jakulin, A. (2007). Bayes: liberal, radical, or conservative? *Statistica Sinica* **17**, 422–426.

Gelman, A., and Pardoe, I. (2007). Average predictive comparisons for models with non-linearity, interactions, and variance components. *Sociological Methodology*.

Gelman, A., Pittau, M. G., Yajima, M., and Su, Y. S. (2007). An approximate EM algorithm for multilevel generalized linear models. Technical report, Department of Statistics, Columbia University.

Genkin, A., Lewis, D. D., and Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics* **49**, 291–304.

Greenland, S. (2001). Putting background information about relative risks into conjugate prior distributions. *Biometrics* **57**, 663–670.

Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in Medicine*.

Heinze, G., and Schemper, M. (2003). A solution to the problem of separation in logistic regression. *Statistics in Medicine* **12**, 2409–2419.

Lesaffre, E., and Albert, A. (1989). Partial separation in logistic discrimination. *Journal of the Royal Statistical Society B* **51**, 109–116.

Liu, C. (2004). Robit regression: a simple robust alternative to logistic and probit regression. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, ed. A. Gelman and X. L. Meng, 227–238. London: Wiley.

MacLehose, R. F., Dunson, D. B., Herring, A. H., and Hoppin, J. A. (2006). Bayesian methods for highly correlated exposure data. *Epidemiology*, under revision.

Martin, A. D., and Quinn, K. M. (2002). MCMCpack. `scythe.wustl.edu/mcmcpack.html`

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*, second edition. London: Chapman and Hall.

Newman, D. J., Hettich, S., Blake, C. L., and Merz, C. J. (1998). UCI Repository of machine learning databases. Department of Information and Computer Sciences, University of California, Irvine. `www.ics.uci.edu/∼mlearn/MLRepository.html`

Racine, A., Grieve, A. P., Fluhler, H., and Smith, A. F. M. (1986). Bayesian methods in practice: experiences in the pharmaceutical industry (with discussion). *Applied Statis-*

tics **35**, 93–150.

Raghunathan, T. E., Van Hoewyk, J., and Solenberger, P. W. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* **27**, 85–95.

Rubin, D. B. (1978). Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse (with discussion). *Proceedings of the American Statistical Association, Survey Research Methods Section*, 20–34.

Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association* **91**, 473–520.

Stigler, S. M. (1977). Do robust estimators work with real data? *Annals of Statistics* **5**, 1055–1098.

Van Buuren, S., and Oudshoom, C. G. M. (2000). MICE: Multivariate imputation by chained equations (S software for missing-data imputation).
`web.inter.nl.net/users/S.van.Buuren/mi/`

Vilalta, R., and Drissi, Y. (2002). A perspective view and survey of metalearning. *Artificial Intelligence Review*, **18** (2), 77–95,

Witte, J. S., Greenland, S., Kim, L. L. (1998). Software for hierarchical modeling of epidemiologic data. *Epidemiology* **9**, 563–566.

Zhang, T., and Oles, F. J. (2001). Text categorization based on regularized linear classification methods. *Information Retrieval* **4**, 5–31.

Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis* **13**, 157–170.