Supplement to: Default Priors for the Intercept Parameter in Logistic Regressions by Boonstra, et al.

S1 Properties of the Exponential Power (EP) Distribution

Normalizing Constant

Let $\alpha \sim \text{EP}_k(\sigma)$, so that $\pi(\alpha) \propto \exp\{-|\alpha/\sqrt{2}\sigma|^k\}$, $\alpha \in (-\infty, \infty)$. The normalizing constant is $Z(\sigma, k) = 1/\left[\int_{-\infty}^{\infty} \exp\{-|x/\sqrt{2}\sigma|^k\}\text{d}x\right] = 0.5/\left[\int_0^{\infty} \exp\{-(x/\sqrt{2}\sigma)^k\}\text{d}x\right]$. Define $u = (x/k)^k$, so that $x = ku^{1/k}$ and $\text{d}x = u^{1/k-1}\,\text{d}u$. Then,

$$\begin{aligned}\int_0^{\infty} \exp\{-(x/\sqrt{2}\sigma)^k\}\text{d}x &= \int_0^{\infty} u^{1/k-1} \exp\left\{-u(\sqrt{2}\sigma/k)^{-k}\right\}\text{d}u \\ &= (\sqrt{2}\sigma/k)\Gamma(1/k).\end{aligned}$$

The last integral is the kernel of a gamma density, implying that $U = (|\alpha|/k)^k \sim \text{Gamma}(1/k, (\sqrt{2}\sigma/k)^k)$. Thus, $Z(\sigma, k) = k/[\sqrt{8}\sigma\Gamma(1/k)]$, and

$$\pi(\alpha|\sigma, k) = \frac{k \exp\{-|\alpha/\sqrt{2}\sigma|^k\}}{\sqrt{8}\sigma\Gamma(1/k)}$$

Distribution Function

Using the above relationship and the symmetry of the ExpPow distribution, we have

$$\begin{aligned}\Pr(|\alpha| > x|\sigma, k) = \Pr(U > (x/k)^k|\sigma, k) &= \frac{\Gamma\left(1/k, (x/k)^k/(\sqrt{2}\sigma/k)^k\right)}{\Gamma(1/k)} \\ &= \frac{\Gamma\left(1/k, (x/\sqrt{2}\sigma)^k\right)}{\Gamma(1/k)}\end{aligned}$$

This can be calculated in `R` (R Core Team, 2016) with

```
> pgamma(q = (x/sqrt(2)/sigma)^k,shape = 1/k,lower.tail = FALSE)
```

S2 Generating Scenarios in Simulation Study: Further Details

Scenario 1 $\boldsymbol{X}$ is multivariate Bernoulli with length $p = 4$, where $\boldsymbol{X} = 1[\boldsymbol{X}^* > 0]$; $\boldsymbol{X}^*$ is multivariate normal with length p; each element has mean zero, unit variance, and pairwise correlation 0.50; $\boldsymbol{\beta}^* = \{0.50, \ldots, 0.50\}$; $\alpha^* = -2.5$; $\alpha^* + E(\boldsymbol{X})^\top \boldsymbol{\beta}^* = -1.5$; $n \in \{50, 100, 200, 400\}$

Scenario 2 $\boldsymbol{X}$ is multivariate Bernoulli with length $p = 25$, where $\boldsymbol{X} = 1[\boldsymbol{X}^* > 0]$; $\boldsymbol{X}^*$ is multivariate normal with length p; each element has mean $\Phi^{-1}(0.25) = -0.67$, unit variance, and pairwise correlation 0.15; $\boldsymbol{\beta}^* = \{1.5, 0, \ldots, 0\}$; $\alpha^* = -2$; $\alpha^* + E(\boldsymbol{X})^\top \boldsymbol{\beta}^* = -1.625$; $n \in \{50, 100, 200\}$

Scenario 3 $\boldsymbol{X}$ is multivariate Bernoulli with length $p = 25$, where $\boldsymbol{X} = 1[\boldsymbol{X}^* > 0]$; $\boldsymbol{X}^*$ is multivariate normal with length p; each element has mean $\Phi^{-1}(0.25) = -0.67$, unit variance, and pairwise correlation 0.15; $\boldsymbol{\beta}^* = \{0.06, \ldots, 0.06\}$; $\alpha^* = -2$; $\alpha^* + E(\boldsymbol{X})^\top \boldsymbol{\beta}^* = -1.625$; $n \in \{50, 100, 200\}$

Scenario 4 $\boldsymbol{X}$ is multivariate Bernoulli with length $p = 25$, where $\boldsymbol{X} = 1[\boldsymbol{X}^* > 0]$; $\boldsymbol{X}^*$ is multivariate normal with length p; each element has mean $\Phi^{-1}(0.05) = -1.64$, unit variance, and pairwise correlation 0.30; $\boldsymbol{\beta}^* = \{\underbrace{3, \ldots, 3}_{10}, \underbrace{0, \ldots, 0}_{15}\}$; $\alpha^* = -6.5$; $\alpha^* + E(\boldsymbol{X})^\top \boldsymbol{\beta}^* = -5$; $n \in \{100, 200, 400\}$

Scenario 5 $\boldsymbol{X}$ is multivariate Bernoulli with length $p = 25$, where $\boldsymbol{X} = 1[\boldsymbol{X}^* > 0]$; $\boldsymbol{X}^*$ is multivariate normal with length p; each element has mean $\Phi^{-1}(0.05) = -1.64$,

unit variance, and pairwise correlation 0.30; $\boldsymbol{\beta}^* = \{\underbrace{3,\ldots,3}_{10},\underbrace{0,\ldots,0}_{15}\}$; $\alpha^* = -4$; $\alpha^* + E(\boldsymbol{X})^\top\boldsymbol{\beta}^* = -2.5$; $n \in \{50, 100, 200\}$

Scenario 6 $\boldsymbol{X}$ is multivariate normal with length $p = 75$; each element has mean zero, unit variance, and pairwise correlation 0.30; $\boldsymbol{\beta}^* = \{2, 0, \ldots, 0\}$; $\alpha^* = \alpha^* + E(\boldsymbol{X})^\top\boldsymbol{\beta}^* = -4$; $n \in \{100, 200, 400\}$

Scenario 7 $\boldsymbol{X}$ is multivariate Bernoulli with length $p = 75$, where $\boldsymbol{X} = 1[\boldsymbol{X}^* > 0]$; $\boldsymbol{X}^*$ is multivariate normal with length p; each element has mean $\Phi^{-1}(0.25) = -0.67$, unit variance, and pairwise correlation 0.30; $\boldsymbol{\beta}^* = \{2, 0, \ldots, 0\}$; $\alpha^* = -3.5$; $\alpha^*+E(\boldsymbol{X})^\top\boldsymbol{\beta}^* = -3$; $n \in \{100, 200, 400\}$

Scenario 8 $\boldsymbol{X}$ is multivariate normal with length $p = 150$; each element has mean zero, unit variance, and pairwise correlation 0.10; $\boldsymbol{\beta}^* = \{-0.5, -0.5, 0, \ldots, 0\}$; $\alpha^* = \alpha^* + E(\boldsymbol{X})^\top\boldsymbol{\beta}^* = -3$; $n \in \{100, 200, 400, 600\}$

Scenario 9 $\boldsymbol{X}$ is multivariate normal with length $p = 150$; each element has mean zero, unit variance, and pairwise correlation 0.10; $\boldsymbol{\beta}^* = \{-1/150, \ldots, -1/150\}$; $\alpha^* = \alpha^* + E(\boldsymbol{X})^\top\boldsymbol{\beta}^* = -3$; $n \in \{100, 200, 400, 600\}$

S3 Algorithm 2: Calculate n_{comp}, n_{piv}, or n_{over}

We assume that n_{comp} is to be calculated, and the algorithm is similar for the other separation statistics. For a given hyperplane of dimension $p+1$, say $\boldsymbol{b}$, which yields the one-dimensional linear predictor $\{1, \boldsymbol{X}_i^\top\}\boldsymbol{b}$, $i = 1, \ldots, n$, the minimum number of observations removed to induce separation is calculable in $n\log(n)$ time by using a binary search. The minimum of these minima over all possible hyperplanes is precisely the n_{comp} statistic, but determining this is NP hard (Hoffgen et al., 1995; Christmann and Rousseeuw, 2001). The algorithm we used also reports a minimum of minima but searches over a subset of candidate hyperplanes that is intended to be likely to contain the best-separating hyperplane, i.e. correspond to n_{comp}. Thus, by construction, the value returned by our algorithm will always bound the true value of n_{comp} from above, but this may not be tight, i.e. we can only guarantee it is an upper-bound.

For a given dataset of n observations of $\{Y_i, \boldsymbol{X}_i\}_i$ with $\boldsymbol{X}_i$ being length-p, the algorithm proceeds as follows.

1. Calculate the linear predictor from the full multivariable logistic regression of Y against all predictors $\{\boldsymbol{X}_i\}_i$ simultaneously. Set $n_{\text{comp}}^{\text{current}}$ equal to the number observations

needed to induce separation using this linear predictor. If this number is zero, proceed to step 5.

2. Now determine whether a different hyperplane constructed using a subset of observations can induce separation. The intuition is that hyperplanes constructed from subsamples that preferentially exclude the high-influence observations, i.e. those subsamples enriched with low-influence observations, may be able to better separate the data than those hyperplanes that are sensitive to high-influence observations. Calculate a length-n vector of sampling weights $\{w_i\}$, where each w_i is equal to the square root of the reciprocal of the absolute difference between the corresponding observation's linear predictor from the full multivariable logistic model and that calculated after leaving out the observation. This is used as a simple measure of that observation's influence in estimation.

3. For $k = 1, \ldots, MAXk$,

   (a) if $\binom{n}{n-k} < NDRAWS$, then, construct every possible subset of $\binom{n}{n-k}$ observations.

   (b) Otherwise, sample $NDRAWS$ subsets of size $n - k$ as follows. Identify the $NDRAWS/2$ subsets of observations with the largest partial sums of weights w_i from step 3. For the remaining $NDRAWS/2$ subsets, randomly sample observations in proportion to each weight w_i. Remove any duplicated sampled subsets from among the $NDRAWS$ subsets.

   For every subset of observations above, calculate the linear predictor for all n observations using the multivariable logistic regression of Y against all $\boldsymbol{X}$'s fit to the subset of size $n - k$. Whenever the minimum number removed to induce separation falls below $n_{\text{comp}}^{\text{current}}$, set $n_{\text{comp}}^{\text{current}}$ equal to this number. If $n_{\text{comp}}^{\text{current}} = 0$, step out of the loop and proceed to step 5.

4. Return the final value of $n_{\text{comp}}^{\text{current}}$ as the estimate of n_{comp} and the corresponding hyperplane that resulted in this greatest separation.

We extended this algorithm to also approximate n_{piv} (the minimum number of observations necessary to induce pivotal separation among the remaining subsets, defined in Section 2.1) and n_{over} (the minimum number of observations necessary to remove to induce quasi-complete separation among the remaining subset (Christmann and Rousseeuw, 2001)). Applied to five datasets considered in Christmann and Rousseeuw (2001), and using $MAKk = 8$ and $NDRAWS = 1000$ (for small n) or $NDRAWS = 200$ (for large n), Algorithm 2 was able to match or improve upon (i.e. identify a verifiable lower upperbound of n_{comp} or n_{over}) an alternative algorithm considered in that paper (see Table 4 in the manuscript). We were not able obtain a sixth dataset ('Hemophilia') considered by those authors

In the simulation study and analysis in the manuscript, we used $MAKk = 8$ and a larger value of $NDRAWS = 5000$.

S4 Supplemental Tables and Figure

Table S1: Values of the extreme boundaries $\text{logit}^{-1}(s_n)$ and the scale parameter σ_n of the EP distribution with $k = 2$ or $k = 4$ and the Logistic distribution, for varying sample sizes n, as calculated by **Algorithm 1** with $\delta = 1$ and $q = 0.01$.

		EP($k = 2$)	EP($k = 4$)	Logistic
n	$\text{logit}^{-1}(s_n)$	σ_n	σ_n	σ_n
250	6.21	2.41	3.52	1.17
500	6.91	2.68	3.91	1.30
1000	7.60	2.95	4.30	1.44
2000	8.29	3.22	4.70	1.57
4000	8.99	3.49	5.09	1.70
8000	9.68	3.76	5.48	1.83

Table S2: Median AUCs (×100), across 200 datasets, using the posterior mean of $\boldsymbol{\beta}$ for all combinations of six priors on the intercept parameter α and two priors on $\boldsymbol{\beta}$. Larger numbers are better. In each row, values in *italics*correspond to the largest median AUC, and values in **bold** had a larger AUC than the italicized value in at least 33% of the datasets. All metrics were calculated separately for each prior on $\boldsymbol{\beta}$. The columns n_{comp} and n_{piv} give the median values of these statistics across the 200 simulated datasets.

			median		$\boldsymbol{\beta} \sim \text{HS}(\tilde{p}_{\text{eff}})$						$\boldsymbol{\beta} \sim \text{Logis}(1)$					
Scenario	p	n	$n_{\text{comp.}}$	$n_{\text{piv.}}$	$t_3(10)$	$t_\infty(10)$	$\text{EP}_2(\sigma_n)$	$\text{EP}_4(\sigma_n)$	$\text{EP}_{10}(\sigma_n)$	$\text{Logis}(\sigma_n)$	$t_3(10)$	$t_\infty(10)$	$\text{EP}_2(\sigma_n)$	$\text{EP}_4(\sigma_n)$	$\text{EP}_{10}(\sigma_n)$	$\text{Logis}(\sigma_n)$
1	4	50	9	9	**66.3**	**66.4**	**66.4**	**66.3**	**66.4**	***66.5***	**65.5**	***65.5***	**65.5**	**65.5**	**65.5**	**65.5**
1	4	100	20	20	**66.9**	***67.1***	**67.0**	**67.0**	**67.1**	**67.1**	***66.6***	**66.5**	**66.5**	**66.5**	**66.5**	**66.5**
1	4	200	40	40	**67.7**	**67.7**	**67.7**	***67.8***	**67.7**	**67.7**	**67.4**	***67.4***	**67.4**	**67.4**	**67.4**	**67.4**
1	4	400	84	84	***67.9***	**67.9**	**67.9**	**67.9**	**67.9**	**67.9**	**67.8**	**67.8**	**67.8**	***67.8***	**67.8**	**67.8**
2	25	50	0	1	**56.5**	**56.5**	**57.8**	**57.5**	**57.3**	***58.0***	**56.6**	**56.5**	**56.6**	**56.5**	**56.4**	***56.7***
2	25	100	2	7.5	**62.7**	**63.1**	***63.1***	**62.9**	**62.8**	**63.1**	**58.5**	**58.5**	**58.5**	**58.5**	**58.4**	***58.6***
2	25	200	22	24.5	**65.3**	**65.3**	**65.3**	***65.3***	**65.3**	**65.3**	**61.2**	***61.3***	**61.3**	**61.2**	**61.2**	**61.3**
3	25	50	0	1	**53.0**	**52.9**	***53.0***	**52.9**	**53.0**	**53.0**	**52.8**	***52.8***	**52.8**	**52.7**	**52.8**	**52.8**
3	25	100	2	8	**52.6**	**52.6**	**52.5**	**52.6**	**52.6**	***52.6***	**52.5**	**52.5**	**52.6**	***52.6***	**52.5**	**52.5**
3	25	200	22	25	**53.7**	**53.7**	**53.8**	***53.8***	**53.8**	**53.7**	**52.8**	**52.7**	**52.8**	**52.8**	***52.8***	**52.8**
4	25	100	0	0	**80.0**	78.9	**82.2**	***82.4***	**82.2**	**82.1**	***91.7***	**91.7**	91.1	91.2	91.2	91.3
4	25	200	1	2	92.5	92.2	**93.6**	***93.6***	**93.6**	**93.5**	**94.8**	***94.8***	**94.7**	**94.7**	**94.8**	**94.8**
4	25	400	2	7	**96.1**	***96.1***	**96.1**	**96.1**	**96.1**	**96.1**	**96.3**	**96.3**	***96.3***	**96.3**	**96.3**	**96.3**
5	25	50	1	1	72.4	72.5	***73.9***	**73.7**	73.5	**73.8**	**83.9**	**83.9**	**84.1**	**84.1**	***84.3***	**84.0**
5	25	100	2	2	79.8	79.8	***80.6***	**80.5**	80.4	**80.6**	**88.0**	**87.9**	**88.0**	***88.2***	**88.0**	**88.0**
5	25	200	6	6	**90.0**	**90.1**	**90.1**	***90.1***	**90.0**	**90.0**	**90.8**	***90.8***	**90.8**	**90.8**	**90.8**	**90.8**
6	75	100	0	0	**85.6**	**85.6**	**86.2**	**86.2**	**86.2**	***86.3***	***74.5***	74.1	70.3	68.6	67.3	72.2
6	75	200	0	1	88.5	88.5	**88.7**	**88.7**	**88.7**	***88.7***	***78.5***	78.4	74.7	72.4	71.4	76.7
6	75	400	0	8	**88.7**	**88.7**	***88.8***	**88.7**	**88.7**	**88.8**	***79.0***	79.0	78.0	77.4	77.2	78.6
7	75	100	0	0	**60.3**	**60.1**	**63.6**	**63.8**	**63.5**	***64.1***	***57.7***	**57.7**	57.0	**56.5**	56.4	57.2
7	75	200	0	2	**67.7**	**67.5**	***69.5***	**69.4**	**69.1**	**69.4**	**58.5**	***58.6***	**57.8**	**57.6**	**57.3**	**58.2**
7	75	400	0.5	12	**71.7**	**71.9**	**72.0**	**71.9**	**72.0**	***72.1***	**60.2**	***60.4***	**60.2**	**59.9**	**59.9**	**60.2**
8	150	100	0	0	**55.7**	**55.7**	***56.5***	**56.5**	**56.4**	**56.4**	**55.7**	***55.7***	**55.1**	**55.0**	**55.0**	**55.5**
8	150	200	0	0	**58.4**	**58.4**	***58.6***	**58.5**	**58.3**	**58.4**	***56.3***	**56.1**	**55.7**	**54.9**	**54.6**	55.9
8	150	400	0	4	***61.3***	**61.2**	**61.0**	**61.1**	**61.1**	**61.2**	***55.6***	**55.6**	**55.0**	**54.6**	**54.4**	**55.5**
8	150	600	0	14	**63.4**	**63.6**	**63.5**	**63.4**	**63.4**	***63.6***	***56.1***	**56.0**	55.7	55.9	55.9	55.9
9	150	100	0	0	**54.7**	**55.2**	**55.6**	**55.8**	***56.1***	**56.0**	**55.0**	**55.1**	***55.2***	**55.1**	**55.0**	**55.0**
9	150	200	0	0	**56.0**	**55.9**	**56.2**	**55.9**	**55.9**	***56.3***	***55.0***	**54.9**	**54.9**	**54.3**	**54.0**	**54.8**
9	150	400	0	3	**56.2**	**56.1**	***56.7***	**56.3**	**56.2**	**56.6**	**54.2**	**54.3**	**54.2**	**53.7**	**53.8**	***54.3***
9	150	600	0	11	**56.9**	**57.0**	**56.9**	**56.9**	**56.9**	***57.1***	***54.2***	**54.1**	**54.0**	**54.1**	**53.9**	**54.1**
Avg. Rank (1-6)					3.77	3.80	3.29	3.36	3.48	3.30	3.05	3.13	3.67	3.77	3.91	3.45

References

Christmann, A. and Rousseeuw, P. (2001). Measuring overlap in binary regression. *Computational Statistics & Data Analysis* **37,** 65–75.

Hoffgen, K., Simon, H., and Vanhorn, K. (1995). Robust trainability of single neurons. *Journal of Computer and System Sciences* **50,** 114–125.

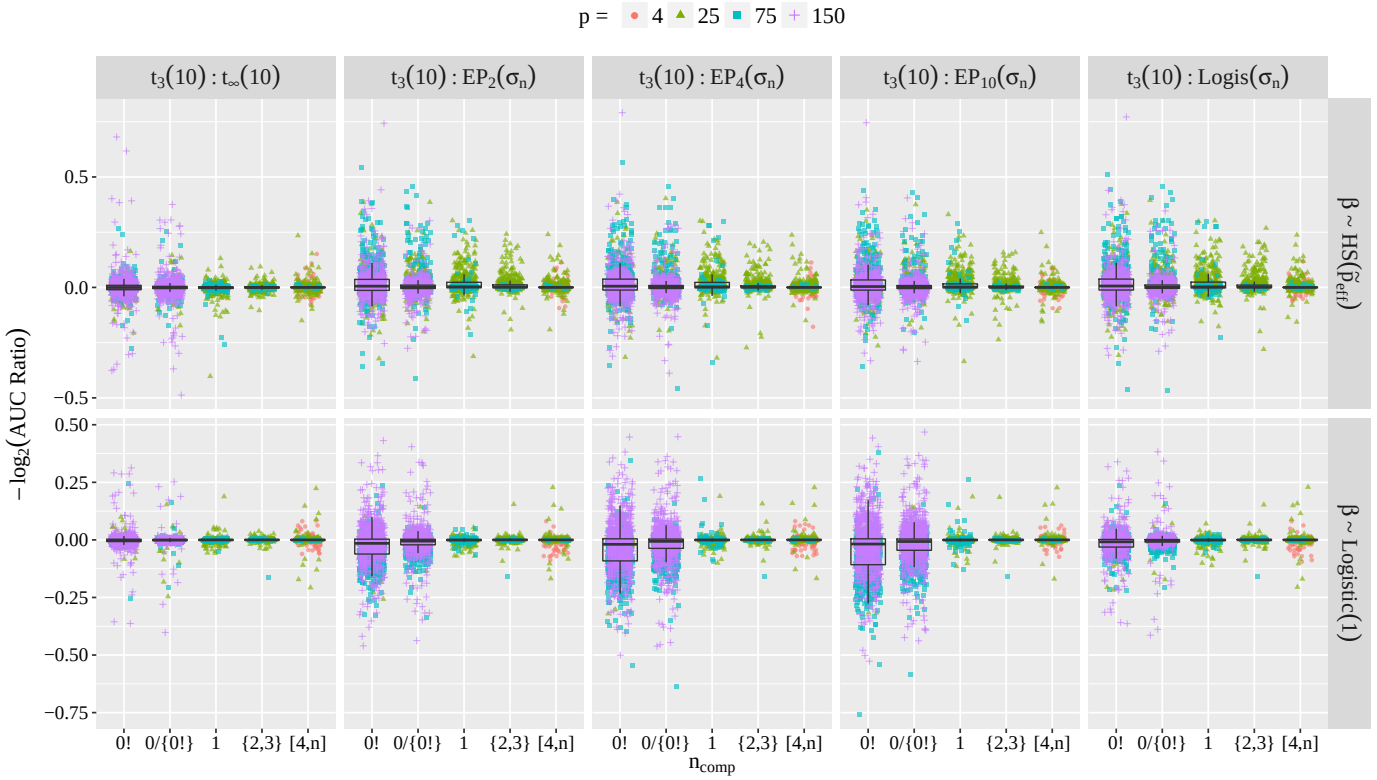

Figure S1: Comparison of $t_3(10)$ prior on α against five alternative priors on α (columns) under two different priors on $\boldsymbol{\beta}$ (rows). Each point represents an individual simulated dataset. The y-axis gives area under the curve (AUC) ratios on the negative-$\log_2$ scale when using the posterior of mean of $\boldsymbol{\beta}$ to classify observations, and the x-axis defines groups based upon separation: "0!" indicates pivotal separation ($n_{\text{piv}} = 0$); "0/{0!}" indicates complete but not pivotal separation ($n_{\text{piv}} > n_{\text{comp}} = 0$); and the remaining categories correspond to value(s) of n_{comp}. **Positive values on the y-axis indicate that the given prior on α yielded better classification of observations than a $t_3(10)$ prior on α.** Different plot characters are used to indicate p, the number of predictors. In total, each panel contains 6000 points (30 unique scenario-sample size configuration times 200 simulated datasets per configuration).

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.