Catalytic Prior Distributions with Application to Generalized Linear Models

Dongming Huang[a], **Nathan Stein**[b], **Donald B. Rubin**[a,c,1], **and S. C. Kou**[a,1]

[a]Department of Statistics, Harvard University, Cambridge, MA 02138; [b]Spotify, New York, NY 10011; [c]Tsinghua University, Beijing 100084, China

This manuscript was compiled on April 2, 2020

A catalytic prior distribution is designed to stabilize a high-dimensional "working model" by shrinking it toward a "simplified model." The shrinkage is achieved by supplementing the observed data with a small amount of "synthetic data" generated from a predictive distribution under the simpler model. We apply this framework to generalized linear models, where we propose various strategies for the specification of a tuning parameter governing the degree of shrinkage and study resultant theoretical properties. In simulations, the resulting posterior estimation using such a catalytic prior outperforms maximum likelihood estimation from the working model and is generally comparable or superior to existing competitive methods in terms of frequentist prediction accuracy of point estimation and coverage accuracy of interval estimation. The catalytic priors have simple interpretations and are easy to formulate.

Bayesian priors | synthetic-data | stable estimation | predictive distribution | regularization

The prior distribution is a unique and important feature of Bayesian analysis, yet in practice, it can be difficult to quantify existing knowledge into actual prior distributions; thus, automated construction of prior distributions can be desirable. Such prior distributions should stabilize posterior estimation in situations when maximum likelihood (ML) behaves problematically, which can occur when sample sizes are small relative to the dimensionality of the models. Here we propose a class of prior distributions designed to address such situations. Henceforth we call the complex model that the investigator wishes to use to analyze the data the "working model."

Often with real working models and datasets, the sample sizes are relatively small, and a likelihood-based analysis is unstable, whereas a likelihood-based analysis of the same dataset using a simpler but less rich model can be stable. Catalytic priors* effectively supplement the observed data with a small amount of *synthetic data* generated from a suitable predictive distribution, such as the posterior predictive distribution under the simpler model. In this way, the resulting posterior distribution under the working model is pulled toward the posterior distribution under the simpler model, resulting in estimates and predictions with better frequentist properties. The name for these priors arises because a catalyst is something that stimulates a reaction to take place that would not take place (or not as effectively) without it, but only an insubstantial amount of the catalyst is needed. When the information in the observed data is substantial, the catalytic prior has a minor influence on the resulting inference, because the information in the synthetic data is small relative to the information in the observed data.

We are not the first to suggest such priors, but we embed the suggestion within a general framework designed for a broad range of examples. One early suggestion for the applied use of such priors was in Ref. (1), which was based on an earlier proposal by Rubin in a 1983 report for the U.S. Census Bureau (reprinted as an appendix in Ref. (2)). Such a prior was also used in a Bayesian analysis of data with noncompliance in a randomized trial (3).

As in both of these earlier references, consider logistic regression as an example:

$$y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta} \sim \text{Bernoulli}\left(1/(1+\exp(-\boldsymbol{x}_i^\top \boldsymbol{\beta}))\right), \quad i = 1, \ldots, n,$$

where, for the ith data point $(y_i, \boldsymbol{x}_i)$, $y_i \in \{0, 1\}$ is the response, and $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{i,p-1})^\top$ represents p covariates, with unknown coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^\top$. The ML estimate (MLE) of $\boldsymbol{\beta}$ is infinite when there is complete separation (4, 5) of the observed covariate values in the two response categories, which can occur easily when p is large relative to n. Earlier attempts to address this problem, such as using Jeffrey's prior (6–9), are not fully satisfactory. This problem arises commonly in practice, for example, Ref. (1) studied the mapping of industry and occupation (I/O) codes in the 1970 U.S. Census to the 1980 Census codes, where both coding systems had hundreds of categories. The I/O classification system changed drastically from the 1970 Census to the 1980 Census, and a single 1970 code could map into as many as 60 possible 1980 codes. For each 1970 code, the 1980 code was considered as missing and multiply-imputed based on covariates. The imputation models were nested (dichotomous) logistic regression models (10) estimated from a special training sample for which both 1970 and 1980 codes were known. The covariates used in these models were derived from nine different factors (sex, age, race, etc.) that formed a cross-classification with $J = 2,304$ categories. The sample available to estimate the mapping was smaller than ten for some 1970 codes, and many of these logistic regression models

Significance Statement

We propose a strategy for building prior distributions that stabilize the estimation of complex "working models" when sample sizes are too small for standard statistical analysis. The stabilization is achieved by supplementing the observed data with a small amount of synthetic data generated from the predictive distribution of a simpler model. This class of prior distributions is easy to use and allows direct statistical interpretation.

D.H., N.S., D.B.R. and S.C.K. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

*Throughout the paper, we use the ineloquent but compact 'priors' in place of the correct 'prior distributions'

[1]To whom correspondence should be addressed. E-mail: kou@stat.harvard.edu or dbrubin@me.com

DRAFT

faced complete separation. The successful approach in Ref. (1) was to use the prior distribution

$$\pi(\boldsymbol{\beta}) \propto \prod_{j=1}^{J} \left(\frac{e^{\boldsymbol{x}_j^{*\top}\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_j^{*\top}\boldsymbol{\beta}}} \right)^{p\hat{\mu}/J} \left(\frac{1}{1+e^{\boldsymbol{x}_j^{*\top}\boldsymbol{\beta}}} \right)^{p(1-\hat{\mu})/J}, \quad [1]$$

where each $\boldsymbol{x}_j^*$ is a possible covariate vector of the cross-classification; p is the dimension of $\boldsymbol{\beta}$; and $\hat{\mu} = \sum_{i=1}^{n} y_i/n$ is the marginal proportion of ones among the observed responses. In this example, the simpler model has the responses y_i's independent of the covariates:

$$y_i \mid \boldsymbol{x}_i, \mu \sim \text{Bernoulli}\,(\mu) \quad (i = 1, \ldots, n),$$

where $\mu \in (0,1)$ is a probability estimated by $\hat{\mu}$. If we supplement the dataset with $p\hat{\mu}/J$ synthetic data points $(y_j^* = 1, \boldsymbol{x}_j^*)$ and $p(1-\hat{\mu})/J$ synthetic data points $(y_j^* = 0, \boldsymbol{x}_j^*)$ for each $\boldsymbol{x}_j^*$ $(j = 1, \ldots, J)$, then the likelihood function of the augmented dataset has the same form as the posterior distribution with the prior in Eq. (1):

$$\pi(\boldsymbol{\beta} \mid \{(y_i, \boldsymbol{x}_i)\}_{i=1}^{n}) \quad [2]$$
$$\propto \prod_{j=1}^{J} \left(\frac{e^{\boldsymbol{x}_j^{*\top}\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_j^{*\top}\boldsymbol{\beta}}} \right)^{N_{j,1}+p\hat{\mu}/J} \left(\frac{1}{1+e^{\boldsymbol{x}_j^{*\top}\boldsymbol{\beta}}} \right)^{N_{j,0}+p(1-\hat{\mu})/J},$$

where $N_{j,1}, N_{j,0}$ are, respectively, the numbers of $(1, x_j^*)$ and $(0, x_j^*)$ in the observed data. In this construction, the total amount of synthetic data is taken to be p, the dimension of $\boldsymbol{\beta}$; see Remark 2.2 in SI for more discussion. The resulting MLE with the augmented dataset equals the maximum posterior estimator (the value of $\boldsymbol{\beta}$ that maximizes the posterior distribution), and it will always be unique and finite when $\hat{\mu} \in (0,1)$.

How to use the synthetic-data perspective for constructing general prior distributions, which we called catalytic prior distributions, is our focus. We mathematically formulate the class of catalytic priors and apply them to generalized linear models. We show that a catalytic prior is proper and yields stable estimates under mild conditions. Simulation studies indicate the frequentist properties of the model estimator using catalytic priors are comparable, and sometimes superior, to existing competitive estimators. Such a prior has the advantages that it is often easier to formulate and it allows for simple implementation from standard software.

We also provide an interpretation of the catalytic prior from an information theory perspective (detailed in Section 4 of SI).

Related Priors. The practice of using synthetic data or pseudo data to define prior distributions has a long history in Bayesian statistics (11). It is well-known that conjugate priors for exponential families can be viewed as the likelihood of pseudo-observations (12). Some authors have suggested formulating priors by obtaining additional pseudo data from experts' knowledge (13–15), which is not easy to use in practice when data have many dimensions or when numerous models or experts are being considered. Refs. (16, 17) proposed to use a conjugate Beta-distribution prior with specifically chosen values of covariates to approximate a multivariate Gaussian prior for the regression coefficients in a logistic regression model. A complication of this approach is that the augmented dataset may contain impossible values for a covariate. Another approach is the *expected-posterior prior* (18–20), where the prior is defined as the average posterior distribution over a set of imaginary data sampled from a simple predictive model. This approach is designed to address the challenges in Bayesian model selection. Other priors have been proposed to incorporate information from previous studies. Particularly, the *power prior* (21–23) formulates an informative prior generated by a power of the likelihood function of historical data. One limitation of this power prior is that its properness requires the covariate matrix of historical or current data to have full column rank (22). Recently, the *power-expected-posterior prior* was proposed to alleviate the computational challenge of expected-posterior priors for model selection (24, 25). It incorporates the ideas of both the expected-posterior prior and the power prior, but it cannot be applied when the dimension of the working model is larger than the sample size. Some other priors suggested in the literature have appearances similar to catalytic priors. Ref. (26) propose the *reference prior* that maximizes the mutual information between the data and the parameter, resulting in a prior density function that looks similar to that of a catalytic prior but is essentially different. Ref. (27) proposed a prior based on the idea of matching loss functions, which, although operationally similar to the catalytic prior, is conceptually different because it requires a subjective initial choice for the distribution of the data. In Ref. (28), the class of *penalized complexity priors* for hierarchical model components is based on penalizing the complexity induced by the deviation from a simpler model. The simpler model there needs to be nested in the working model, which is not required by the catalytic prior.

Generic Formulation of Catalytic Priors

Catalytic Prior in the Absence of Covariates. Consider the data, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, being analyzed under a working model $Y_i \overset{i.i.d.}{\sim} f(y \mid \theta)$ governed by unknown parameter θ. Suppose a model $g(y \mid \psi)$ with unknown parameter ψ, whose dimension is smaller than that of θ, is stably fitted from $\boldsymbol{Y}$ and results in a predictive distribution $g_*(y^* \mid \boldsymbol{Y})$ for future data drawn from $g(y \mid \psi)$. *The synthetic-data generating distribution* $g_*(y^* \mid \boldsymbol{Y})$ is used to generate the synthetic data $\{Y_i^*\}_{i=1}^{M}$, where M is the synthetic-sample size and the asterisk superscript is used to indicate synthetic data.

The synthetic-data generating distribution can be specified by fitting a model simpler than $f(y \mid \theta)$, but it does not necessarily have to be. Examples: (1) If a Bayesian analysis of the simpler model can be carried out easily, $g_*(y^* \mid \boldsymbol{Y})$ can be taken to be the posterior predictive distribution under the simpler model. (2) Alternatively, one can obtain a point estimate $\hat{\psi}$, and $g_*(y^* \mid \boldsymbol{Y}) = g(y^* \mid \hat{\psi})$ can be the plug-in predictive distribution. (3) If two simpler estimated models are $g_*^{(1)}(y^* \mid \boldsymbol{Y})$ and $g_*^{(2)}(y^* \mid \boldsymbol{Y})$, then $g_*(y^* \mid \boldsymbol{Y})$ can be taken to be a mixture $w\, g_*^{(1)}(y^* \mid \boldsymbol{Y}) + (1-w)\, g_*^{(2)}(y^* \mid \boldsymbol{Y})$ for some $w \in (0,1)$.

The likelihood function of θ under the working model based on the synthetic data $\{Y_i^*\}_{i=1}^{M}$ is $\ell(\theta \mid Y^*) = \prod_{i=1}^{M} f(Y_i^* \mid \theta)$. Because these synthetic data are not really observed data, we down-weight them by raising this likelihood to a power τ/M, where $\tau > 0$ is a tuning parameter called the *prior weight*. This leads to the *catalytic prior* that has an unnormalized

DRAFT

density:

$$\pi_{cat,M}(\theta \mid \tau) \;\propto\; \left\{\prod_{i=1}^{M} f(Y_i^* \mid \theta)\right\}^{\tau/M}, \qquad [3]$$

which depends on the randomly drawn synthetic data $\{Y_i^*\}_{i=1}^M$. The *population catalytic prior* is formally the limit of Eq. (3) as M goes to infinity:

$$\pi_{cat,\infty}(\theta \mid \tau) \;\propto\; \exp\left[\tau \mathbb{E}_{g_*}\left\{\log f(Y^* \mid \theta)\right\}\right]. \qquad [4]$$

Here the expectation $\mathbb{E}_{g_*}\{\log f(Y^* \mid \theta)\}$ in Eq. (4) is taken with respect to $Y^* \sim g_*(Y^* \mid \boldsymbol{Y})$. The dependence of $g_*(Y^* \mid \boldsymbol{Y})$ on the observed $\boldsymbol{Y}$ emphasizes that the catalytic prior is data-dependent, like that used in Box and Cox (29) for power transformations.

The posterior density using the catalytic prior is mathematically proportional to the likelihood with both the observed data and the weighted synthetic data. Thus, we can implement the required Bayesian inference using standard software. For instance, the maximum posterior estimate (posterior mode) is the same as the MLE using the weighted augmented data and can be computed by existing MLE procedures, which can be a computational advantage, as illustrated in Ref. (1).

Catalytic Prior with Covariates. Let $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ be the set of n pairs of a scalar response Y_i and a p-dimensional covariate vector $\boldsymbol{X}_i$; Y_i depends on $\boldsymbol{X}_i$ in the working model with unknown parameter $\boldsymbol{\beta}$:

$$Y_i \mid \boldsymbol{X}_i, \boldsymbol{\beta} \sim f(y \mid \boldsymbol{X}_i, \boldsymbol{\beta}), i = 1, 2 \ldots, n. \qquad [5]$$

Let $\boldsymbol{Y}$ be the vector $(Y_1, \ldots, Y_n)^\top$, and $\mathbb{X}$ be the matrix $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top$. The likelihood of these data is $f(\boldsymbol{Y} \mid \mathbb{X}, \boldsymbol{\beta}) = \prod_{i=1}^n f(Y_i \mid \boldsymbol{X}_i, \boldsymbol{\beta})$.

Suppose a simpler model $g(y \mid \boldsymbol{X}, \boldsymbol{\psi})$ with unknown parameter $\boldsymbol{\psi}$ is stably fitted from $(\boldsymbol{Y}, \mathbb{X})$ and results in a synthetic-data generating distribution $g_*(y \mid \boldsymbol{x}, \boldsymbol{Y}, \mathbb{X})$. Note that $g_*(\cdot)$ here is analogous to its use earlier except that now, in addition to the observed data, it is also conditioned on $\boldsymbol{x}$. The synthetic covariates $\boldsymbol{X}^*$ will be drawn from a distribution $Q(\boldsymbol{x})$, which we call the *synthetic-covariate generating distribution*. We will discuss the choice of $Q(\boldsymbol{x})$ shortly.

Given the distributions $Q(\boldsymbol{x})$ and $g_*(y \mid \boldsymbol{x}, \boldsymbol{Y}, \mathbb{X})$, the catalytic prior first draws a set of synthetic data $\{(Y_i^*, \boldsymbol{X}_i^*)\}_{i=1}^M$ from

$$\boldsymbol{X}_i^* \overset{i.i.d.}{\sim} Q(\boldsymbol{x}), \quad Y_i^* \mid \boldsymbol{X}_i^* \sim g_*(y \mid \boldsymbol{X}_i^*, \boldsymbol{Y}, \mathbb{X}).$$

Hereafter we write $\boldsymbol{Y}^*$ for the vector of synthetic responses $(Y_1^*, \ldots, Y_M^*)^\top$, and $\mathbb{X}^*$ for the matrix of synthetic covariates $(\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_M^*)^\top$. The likelihood of the working model based on the synthetic data $\ell(\boldsymbol{\beta} \mid \boldsymbol{Y}^*, \mathbb{X}^*)$ equals $\prod_{i=1}^M f(Y_i^* \mid \boldsymbol{X}_i^*, \boldsymbol{\beta})$. Because these synthetic data are not really observed, we down-weight them by raising this likelihood to a power τ/M, which gives the unnormalized density of the catalytic prior with covariates:

$$\pi_{cat,M}(\boldsymbol{\beta} \mid \tau) \propto \left\{\prod_{i=1}^{M} f(Y_i^* \mid \boldsymbol{X}_i^*, \boldsymbol{\beta})\right\}^{\tau/M}. \qquad [6]$$

The population catalytic prior (when $M \to \infty$) has unnormalized density:

$$\pi_{cat,\infty}(\boldsymbol{\beta} \mid \tau) \propto \exp\left(\tau \mathbb{E}_{Q,g_*}\left[\log f(Y^* \mid \boldsymbol{X}^*, \boldsymbol{\beta})\right]\right), \qquad [7]$$

where the expectation $\mathbb{E}_{Q,g_*}$ averages over both $\boldsymbol{X}^*$ and Y^*. Denote by $Z_{\tau,M}$ and $Z_{\tau,\infty}$ the integrals of the right-hand sides of Eq. (6) and Eq. (7) w.r.t. $\boldsymbol{\beta}$. When these integrals are finite, the priors are proper, and $Z_{\tau,M}$ and $Z_{\tau,\infty}$ are their normalizing constants.

An advantage of the catalytic prior is that the corresponding posterior has the same form as the likelihood

$$\begin{aligned}\pi(\boldsymbol{\beta}|\mathbb{X}, \boldsymbol{Y}, \tau) &\propto \pi_{cat,M}(\boldsymbol{\beta}|\tau) f(\boldsymbol{Y}|\mathbb{X}, \boldsymbol{\beta}) \\ &\propto \exp\left(\frac{\tau}{M}\sum_{i=1}^{M}\log(f(Y_i^*|\boldsymbol{X}_i^*, \boldsymbol{\beta}) \right. \\ &\quad \left. + \sum_{i=1}^{n}\log(f(Y_i|\boldsymbol{X}_i, \boldsymbol{\beta})\right),\end{aligned}$$

which makes the posterior inference no more difficult than other standard likelihood-based methods. For example, the posterior mode can be easily computed as a maximum weighted likelihood estimate using standard statistical software. Full posterior inference can also be easily implemented by treating the synthetic data as down-weighted data.

Catalytic Prior for GLMs. A generalized linear model (GLM) assumes that, given a covariate vector $\boldsymbol{X}$, the response Y has the following density w.r.t. some base probability measure:

$$f(y \mid \boldsymbol{X}, \boldsymbol{\beta}) = \exp\left(t(y)\theta - b(\theta)\right), \qquad [8]$$

where $t(y)$ is a sufficient statistic, and θ is the canonical parameter that depends on $\eta = \boldsymbol{X}^\top\boldsymbol{\beta}$ through $\theta = \phi(\eta)$, where $\boldsymbol{\beta}$ is the unknown regression coefficient vector and $\phi(\cdot)$ is a monotone differentiable function. The mean of $t(Y)$ is denoted by $\mu(\eta)$ and is equal to $b'(\phi(\eta))$.

When the working model is a GLM, from Eq. (7) and Eq. (8), we have

$$\begin{aligned}&\mathbb{E}_{Q,g_*}\left[\log f(Y^* \mid \boldsymbol{X}^*, \boldsymbol{\beta})\right] \\ =&\mathbb{E}_Q\left\{\phi(\boldsymbol{\beta}^\top\boldsymbol{X}^*)\mathbb{E}_{g_*}\left[t(Y^*) \mid \boldsymbol{X}^*\right] - b(\phi(\boldsymbol{\beta}^\top\boldsymbol{X}^*))\right\}, \qquad [9]\end{aligned}$$

so that the expectation of the log-likelihood does not depend on particular realizations of the synthetic response, but rather on the conditional mean of the sufficient statistic under the synthetic-data generating distribution. Thus, in the case of a GLM (and exponential family models), instead of a specific realization of the synthetic response, one only needs to use the conditional mean of the sufficient statistic $\mathbb{E}_{g_*}[t(Y^*) \mid \boldsymbol{X}^*]$ to form a catalytic prior. This simplification reduces the variability introduced by synthetic data [†].

As a concrete example, consider a linear regression model $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2\mathcal{I}_n)$ with known σ. Suppose the synthetic-data generating model is a sub-model with the estimated parameter $\boldsymbol{\beta}_0^*$, and $\mathbb{X}^*$ is the synthetic covariate matrix. In this case, the catalytic prior with any positive τ has a normal distribution:

$$\boldsymbol{\beta} \sim N\left(\boldsymbol{\beta}_0^*, \frac{\sigma^2}{\tau}\left(\frac{1}{M}(\mathbb{X}^*)^\top\mathbb{X}^*\right)^{-1}\right).$$

If $\lim_{M\to\infty}\frac{1}{M}(\mathbb{X}^*)^\top\mathbb{X}^* = \Sigma_{\boldsymbol{X}}$, the population catalytic prior is

$$\boldsymbol{\beta} \sim N\left(\boldsymbol{\beta}_0^*, \frac{\sigma^2}{\tau}(\Sigma_{\boldsymbol{X}})^{-1}\right).$$

More details about this example can be found in SI.

[†] Note that in the previous example of 1970-1980 I/O code mapping, instead of the raw counts of synthetic responses, their expected values $p\hat{\mu}/J$ and $p(1-\hat{\mu})/J$ were used.

DRAFT

Specifications of the Catalytic Prior

Generating Synthetic Covariates. The synthetic covariate vectors are generated such that $(\mathbb{X}^*)^\top\mathbb{X}^*$ has full rank. Moreover, a synthetic covariate should have the same sample space as a real covariate. The simple choice of resampling the observed covariate vectors would not guarantee the full rank of $(\mathbb{X}^*)^\top\mathbb{X}^*$; for example, if the observed covariates are rank deficient, resampling would still give rank deficient $(\mathbb{X}^*)^\top\mathbb{X}^*$.

Instead, we consider one option for generating synthetic covariates: resample each coordinate of the observed covariates independently. Formally, we define *the independent resampling distribution* by the probability mass function

$$Q_0(\boldsymbol{x}) := \prod_j \left(\frac{1}{n}\#\{1 \le i \le n : (\boldsymbol{X}_i)_j = x_j\}\right),$$

for all $\boldsymbol{x} \in \mathcal{X}$, where $\mathcal{X}$ is the sample space of $\boldsymbol{X}$. We use this distribution for simplicity. Alternatively, if historical data are available, synthetic covariates can be sampled from the historical covariates. Furthermore, if some variables are naturally grouped or highly correlated, one may want to resample these grouped parts together. Other examples are discussed in SI.

Generating Synthetic Responses. The synthetic-data generating distribution can be specified by fitting a simple model $G_\Psi = \{g(y \mid \boldsymbol{x}, \boldsymbol{\psi}) : \boldsymbol{\psi} \in \Psi\}$ to the observed data. The only requirement is that this simple model can be stably fit by the observed data in the sense that the standard estimation of $\boldsymbol{\psi}$, using either a Bayesian or frequentist approach, can lead to a well-defined predictive distribution for future data. Examples include a fixed distribution and an intercept-only model. G_Ψ can also be a regression model based on dimension reduction, such as a principal components analysis; see SI for a numerical example, which also suggests to keep G_Ψ as simple as possible when the observed sample size is small. For a working regression model with interactions, a natural choice of G_Ψ is the sub-model with only main-effects. If the main-effect model is overfitted as well, we could use a mixed synthetic-data generating distribution, such as $g_*(y \mid \boldsymbol{x}, \boldsymbol{Y}, \mathbb{X}) = 0.5\ g_{*,1}(y \mid \boldsymbol{x}, \boldsymbol{Y}, \mathbb{X}) + 0.5\ g_{*,0}(y \mid \boldsymbol{x}, \boldsymbol{Y}, \mathbb{X})$, where $g_{*,1}$ and $g_{*,0}$ are the predictive distributions of the preliminarily fitted main-effect model and intercept-only model, respectively. G_Ψ can also be chosen using additional knowledge, such as a sub-model that includes a few important covariates that have been identified in previous studies, or if domain experts have opinions on the range of possible values of certain model parameters, then the parameter space Ψ can be constrained accordingly.

Sometimes it is beneficial to draw multiple synthetic responses for each sampled synthetic covariate vector. We name this sampling the *stratified synthetic data generation*. It could help reduce variability introduced by synthetic data.

Sample Size of Synthetic Data. Theorem 5 below quantifies how fast the randomness in the catalytic prior diminishes as the synthetic-sample size M increases. One implication is that for linear regression with binary covariates, if $M \ge \frac{4p^3}{\epsilon^2}\log(\frac{p}{\delta})$, then the Kullback-Leibler divergence between the catalytic prior $\pi_{cat,M}$ and its limit $\pi_{cat,\infty}$ is at most ϵ with probability at least $1-\delta$. Such a bound can help choose the magnitude of M. When the prior needs to be proper, we suggest taking M larger than 4 times the dimension of $\boldsymbol{\beta}$ (based on Theorem 1 and Proposition 2 below).

Weight of Synthetic Data. The prior weight τ controls how much the posterior inference relies on the synthetic data because it can be interpreted as the effective prior sample size. Here we provide two guidelines for systematic specifications of τ.

Frequentist Predictive Risk Estimation. Choose a value of τ using the following steps: (1) Compute the posterior mode $\widehat{\boldsymbol{\beta}}(\tau)$ for various values of τ. (2) Choose a discrepancy function $D(y_0, \widehat{\mu})$ that measures how well a prediction $\widehat{\mu}$ predicts a future response y_0. (3) Find an appropriate criterion function $\Lambda(\tau)$ that estimates the expected (in-sample) prediction error, for a future response Y_0 based on $\widehat{\boldsymbol{\beta}}(\tau)$, and (4) Pick the value of τ that minimizes $\Lambda(\tau)$. See Section 2.C.1 in SI for a detailed discussion.

The discrepancy $D(y_0, \widehat{\mu})$ measures the error of a prediction $\widehat{\mu}$ for a future response Y_0 that takes value y_0. We consider here discrepancy functions of the form

$$D(y_0, \widehat{\mu}) := a(\widehat{\mu}) - \lambda(\widehat{\mu})y_0 + c(y_0), \qquad [10]$$

and define $\boldsymbol{D}(\boldsymbol{Y}_0, \widehat{\boldsymbol{\mu}}) := \frac{1}{n}\sum_{i=1}^n D(Y_{0,i}, \widehat{\mu}_i)$. This class is general enough to include squared error, classification error and deviance for GLMs: (a) squared error: $D(y_0, \hat{\mu}) = (y_0 - \hat{\mu})^2 = \hat{\mu}^2 - 2y_0\hat{\mu} + y_0^2$; (b) classification error: $D(y_0, \hat{\mu}) = \mathbf{1}_{y_0 \ne \hat{\mu}} = \hat{\mu} - 2y_0\hat{\mu} + y_0$ for any y_0 and $\hat{\mu}$ in $\{0, 1\}$; (c) deviance for GLMs: $D(y_0, \widehat{\mu}) = b(\widehat{\theta}) - y_0\widehat{\theta} + \sup_\theta(y_0\theta - b(\theta))$, where $\widehat{\theta} = (b')^{-1}(\widehat{\mu})$.

The criterion function $\Lambda(\tau)$ is an estimate of the expectation of the (in-sample) prediction error. Such an estimate can be obtained by using the parametric bootstrap. Take a bootstrap sample of the response vector $\boldsymbol{Y}^{boot}$ from the distribution $f(\boldsymbol{y} \mid \mathbb{X}, \widehat{\boldsymbol{\beta}}^0)$, where $\widehat{\boldsymbol{\beta}}^0 = \widehat{\boldsymbol{\beta}}(\tau_0)$ is a preliminary estimate, and denote by $\widehat{\boldsymbol{\beta}}^{boot}(\tau)$ the posterior mode based on data $(\boldsymbol{Y}^{boot}, \mathbb{X})$ with the catalytic prior. The bootstrap criterion function is given by

$$\Lambda(\tau) = \boldsymbol{D}(\boldsymbol{Y}, \widehat{\boldsymbol{\mu}}_\tau) + \frac{1}{n}\sum_{i=1}^n \mathrm{Cov}(\lambda(\hat{\mu}^{boot}_{\tau,i}), Y_i^{boot}), \qquad [11]$$

where $\widehat{\mu}_{\tau,i} = \mu(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}(\tau))$ and $\hat{\mu}^{boot}_{\tau,i} = \mu(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}^{boot}(\tau))$. See SI for a detailed derivation. In practice, the term $\mathrm{Cov}(\lambda(\hat{\mu}^{boot}_{\tau,i}), Y_i^{boot})$ is numerically computed by sampling $\boldsymbol{Y}^{boot}$ repeatedly. Based on our experiments with linear and logistic models, the default choices of the initial values can be $\tau_0 = 1$ for linear regression and $\tau_0 = p/4$ for other cases. See SI for a mathematical argument.

The costly bootstrap repetition step to numerically compute $\mathrm{Cov}(\lambda(\hat{\mu}^{boot}_{\tau,i}), Y_i^{boot})$ can be avoided in two special cases (see SI for more discussion):

1. If Y_i follows a normal distribution and $\lambda(\widehat{\mu}_{\tau,i})$ is smooth in y_i, then the *Stein's unbiased risk estimate* yields

$$\Lambda(\tau) = \boldsymbol{D}(\boldsymbol{Y}, \widehat{\boldsymbol{\mu}}_\tau) + \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(Y_i)\mathbb{E}\frac{\partial\lambda(\widehat{\mu}_{\tau,i})}{\partial y_i}. \qquad [12]$$

   In particular, when square error is considered and if $\widehat{\boldsymbol{\mu}}_\tau$ can be written as $\widehat{\boldsymbol{\mu}}_\tau = \boldsymbol{H}_\tau \cdot \boldsymbol{Y} + \boldsymbol{c}_\tau$, the risk estimate is

$$\Lambda(\tau) = \|\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_\tau\|^2 + \frac{2}{n}\sum_{i=1}^n \mathrm{Var}(Y_i)\boldsymbol{H}_\tau(i,i). \qquad [13]$$

DRAFT

2. When responses are binary, say 0 or 1, let $\boldsymbol{Y}^{\triangle i}$ be a copy of $\boldsymbol{Y}$ but with Y_i replaced by $1-Y_i$, and let $\widehat{\boldsymbol{\beta}}^{\triangle i}(\tau)$ be the posterior mode based on data $(\boldsymbol{X}, \boldsymbol{Y}^{\triangle i})$ with the catalytic prior. The *Steinian* estimate (30) is given by

$$\boldsymbol{D}(\boldsymbol{Y}, \widehat{\boldsymbol{\mu}}_\tau) + \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_i^0(1-\widehat{\mu}_i^0)(2Y_i-1)\left(\lambda(\widehat{\mu}_{\tau,i}) - \lambda(\widehat{\mu}_{\tau,i}^{\triangle i})\right), \tag{14}$$

where $\widehat{\mu}_i^0 = \mu(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}^0)$, and $\widehat{\mu}_{\tau,i}^{\triangle i} = \mu(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}^{\triangle i}(\tau))$.

Bayesian Hyperpriors. An alternative way to specify the prior weight τ is to consider a joint catalytic prior for $(\tau, \boldsymbol{\beta})$:

$$\pi_{\alpha,\gamma}(\tau, \boldsymbol{\beta}) \propto \Gamma_{\alpha,\gamma}(\tau) \left\{ \prod_{i=1}^{M} f(Y_i^* \mid \boldsymbol{X}_i^*, \boldsymbol{\beta}) \right\}^{\tau/M}, \tag{15}$$

where $\Gamma_{\alpha,\gamma}(\tau)$ is a function defined as follows for positive scalar hyperparameters α and γ. Denote

$$\kappa := \sup_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{M}\sum_{i=1}^{M} \log f(Y_i^* \mid \boldsymbol{X}_i^*, \boldsymbol{\beta}).$$

For linear regression, the function $\Gamma_{\alpha,\gamma}(\tau)$ can be taken to be

$$\Gamma_{\alpha,\gamma}(\tau) = \tau^{\frac{p+\alpha}{2}-1} e^{-\tau(\kappa+\gamma^{-1})}. \tag{16}$$

and for other models,

$$\Gamma_{\alpha,\gamma}(\tau) = \tau^{p+\alpha-1} e^{-\tau(\kappa+\gamma^{-1})}. \tag{17}$$

The form of $\Gamma_{\alpha,\gamma}(\tau)$ is chosen mainly for practical convenience; by separating the dependence on p and κ, we have meaningful interpretations for α and γ. For GLMs, prior moments of $\boldsymbol{\beta}$ up to order α exist, and γ controls the exponential decay of the prior density of τ; see Theorem 4. For linear regression, the marginal prior for $\boldsymbol{\beta}$ induced by Eq. (15) is a multivariate t-distribution centered around the MLE for the synthetic data with covariance matrix $\frac{2\sigma^2}{\alpha\gamma}\cdot(\frac{1}{M}(\mathbb{X}^*)^\top\mathbb{X}^*)^{-1}$ and degrees of freedom α. The analysis in Theorem 4 reveals how the parameters α and γ affect the joint prior. Roughly speaking, a larger value of α (or γ) tends to pull the working model more towards the simpler model. Admittedly, it appears impossible to have a single choice that works the best in all scenarios. We recommend $(\alpha, \gamma) = (2, 1)$ as a simple default choice based on our numerical experiments.

Illustration of Methods

Logistic Regression. We illustrate the catalytic prior using logistic regression. Another example using linear regression is presented in SI. Here the mean of Y depends on the linear predictor $\eta = \boldsymbol{X}^\top\boldsymbol{\beta}$ through $\mu = e^\eta/(1+e^\eta)$. Suppose the synthetic-data generating model includes only the intercept, so it is Bernoulli(μ_0), where a simple estimate of μ_0 is given by $\hat{\mu}_0 = (1/2 + \sum_{i\le n} Y_i)/(1+n)$. The synthetic response vector $\boldsymbol{Y}^*$ can be taken to be $\hat{\mu}_0 \cdot \mathbf{1}_M$, and each synthetic covariate vector $\boldsymbol{X}_i^*$ is drawn from the independent resampling distribution; this prior is proper when $(\mathbb{X}^*)^\top\mathbb{X}^*$ is positive definite according to Theorem 1.

Numerical Example. We first generate the observed covariates $\boldsymbol{X}_i$ by drawing a Gaussian random vector $\boldsymbol{Z}_i$ whose components have mean 0, variance 1 and common correlation $\rho = 0.5$; set

$$\boldsymbol{X}_{i,j} = \begin{cases} 2\cdot\mathbf{1}_{\boldsymbol{Z}_{i,j}>0} - 1, & 2j < p \\ \boldsymbol{Z}_{i,j}, & 2j \ge p. \end{cases}$$

This process yields covariate vectors that have dependent components and have both continuous and discrete components as one would encounter in practical logistic regression problems. We consider three different sparsity levels and three different amplitudes of the regression coefficient $\boldsymbol{\beta}$ in the underlying model. More precisely, $\boldsymbol{\beta}$ is specified through scaling an initial coefficient $\boldsymbol{\beta}^{(0)}$ that accommodates different levels of sparsity. Each coordinate of $\boldsymbol{\beta}^{(0)}$ is either 1 or 0. ζ proportion of the coordinates of $\boldsymbol{\beta}^{(0)}$ are randomly selected and set to 1, and the remaining $1-\zeta$ proportion are set to 0, where ζ is the level of *non-sparsity* and is set at 1/4, 1/2, 3/4. This factor controls how many covariates actually affect the response. Then the amplitude of $\boldsymbol{\beta}$ is specified indirectly: $\boldsymbol{\beta}_0 = c_1$, $\boldsymbol{\beta}_{1:(p-1)} = c_2\boldsymbol{\beta}^{(0)}_{1:(p-1)}$, where parameters (c_1, c_2) are chosen such that the oracle classification error r (the expected classification error of the classifier given by the true $\boldsymbol{\beta}$) is equal to 0.1, 0.2, 0.3. Here $r = \mathbb{E}_{\boldsymbol{X}}\left(\min(\mathbf{P}_{\boldsymbol{\beta}}(Y=1), \mathbf{P}_{\boldsymbol{\beta}}(Y=0))\right) = \mathbb{E}_{\boldsymbol{X}}\left(1+\exp(|\boldsymbol{X}^\top\boldsymbol{\beta}|)\right)^{-1}$ is numerically computed by sampling 2000 extra covariate vectors. The value of r represents how far apart the class $Y=1$ is from the class $Y=0$, and small values of r correspond to large amplitudes of $\boldsymbol{\beta}$.

In this example, the number of covariates is 16, so the dimension of $\boldsymbol{\beta}$ is $p = 17$, and the sample size is $n = 30$. We use the predictive binomial deviance, $\mathbb{E}_{\boldsymbol{X}_0}\left[D(\mu(\boldsymbol{X}_0^\top\boldsymbol{\beta}), \mu(\boldsymbol{X}_0^\top\hat{\boldsymbol{\beta}}))\right]$, where $D(a,b) = a\log(a/b) + (1-a)\log((1-a)/(1-b))$ measures the discrepancy between two Bernoulli distributions with probability a and b respectively, to evaluate the predictive performance of $\hat{\boldsymbol{\beta}}$. The expectation $\mathbb{E}_{\boldsymbol{X}_0}$ is computed by sampling 1000 extra independent copies of $\boldsymbol{X}_0$ from the same distribution that generates the observed covariates.

To specify catalytic priors, we use the generating distributions for synthetic data just described, and fix M at 400. The first estimator of $\boldsymbol{\beta}$ is the posterior mode of $\boldsymbol{\beta}$ with $\tau = \hat{\tau}_{boot}$ selected by predictive risk estimation via the bootstrap with deviance discrepancy (denoted as *Cat.Boot.*). This estimator can be computed as the MLE with the weighted augmented data. The second estimator of $\boldsymbol{\beta}$ is the coordinate-wise posterior median of $\boldsymbol{\beta}$ with the joint prior $\pi_{\alpha=2,\gamma=1}$ (denoted as *Cat.Joint*). The posterior median is used here because there is no guarantee that the posterior distribution of $\boldsymbol{\beta}$ is uni-modal in this case. These estimators are compared with two alternatives: the MLE and the posterior mode with the Cauchy prior (31) (calculated by the authors' R package `bayesglm`).

Table 1 presents the average predictive binomial deviance over 1600 simulations in each cell. The column *Comp.Sep.* shows how often complete separation occurs in the datasets; when complete separation occurs, the MLE does not exist but a pseudo-MLE can be algorithmically computed if the change in the estimate is smaller than 10^{-8} within 25 iterations; the column of MLE averages across only the cases where either MLE or pseudo-MLE exists. In Table 1, the boldface corresponds to the best performing method under each simulation scenario. Based on this table, the catalytic prior with $\hat{\tau}_{boot}$ predicts the best and the MLE predicts the worst in all cases considered. Although the Cauchy prior seems to perform close

DRAFT

Setting				Performance of Methods			
ζ	r	Comp. Sep.		Cat. Boot.	Cat. Joint	Cauchy	MLE (pseudo)
1/4	0.1	100%	Mean	**1.692**	1.772	1.793	2.081
			SE $\times 10^3$	(6.8)	(6.7)	(6.7)	(8.7)
	0.2	98%	Mean	**0.675**	0.769	0.802	1.123
			SE $\times 10^3$	(5.2)	(5.0)	(5.0)	(7.2)
	0.3	91%	Mean	**0.297**	0.399	0.445	0.751
			SE $\times 10^3$	(2.3)	(2.0)	(1.9)	(7.3)
2/4	0.1	100%	Mean	**1.661**	1.742	1.749	2.048
			SE $\times 10^3$	(3.9)	(3.8)	(3.8)	(5.0)
	0.2	98%	Mean	**0.648**	0.743	0.771	1.107
			SE $\times 10^3$	(2.5)	(2.2)	(2.0)	(3.4)
	0.3	92%	Mean	**0.287**	0.392	0.438	0.748
			SE $\times 10^3$	(2.1)	(1.8)	(1.7)	(7.1)
3/4	0.1	100%	Mean	**1.664**	1.746	1.749	2.052
			SE $\times 10^3$	(4.0)	(3.9)	(3.8)	(4.9)
	0.2	99%	Mean	**0.649**	0.745	0.771	1.104
			SE $\times 10^3$	(2.5)	(2.2)	(2.0)	(3.4)
	0.3	91%	Mean	**0.287**	0.391	0.435	0.738
			SE $\times 10^3$	(2.1)	(1.9)	(1.7)	(7.3)

Table 1. Mean and standard error of predictive binomial deviance of different methods. The first two columns are the settings of the simulation: ζ is the non-sparsity, and r is the oracle prediction error. The column of *Comp.Sep.* shows how often complete separation occurs in the datasets. The last four columns report the mean and standard error of the predictive binomial deviance of the different methods, which are the catalytic posterior mode with $\hat{\tau}_{boot}$, denoted by *Cat.Boot.*, the posterior median under joint catalytic prior, denoted by *Cat.Joint*, the Cauchy posterior mode, denoted by *Cauchy*, and the MLE. The boldface corresponds to the best performing method in each simulation scenario.

Difference between the error of Cauchy and that of Cat.Joint			
ζ	r	Mean	SE $\times 10^3$
1/4	0.1	0.021	0.98
	0.2	0.033	0.91
	0.3	0.047	0.86
1/2	0.1	0.007	0.79
	0.2	0.028	0.85
	0.3	0.046	0.84
3/4	0.1	0.003	0.76
	0.2	0.026	0.83
	0.3	0.044	0.82

Table 2. Mean and standard error of the difference in predictive binomial deviance between the Cauchy posterior mode and the joint catalytic posterior median. ζ is the non-sparsity; r is the oracle prediction error.

to the joint catalytic prior, Table 2 shows that the prediction based on the joint catalytic prior is statistically significantly better than that of the Cauchy prior (Table 2 directly calculates the difference of the prediction errors between the Cauchy prior and the joint catalytic prior and shows that the difference is significantly positive with Bonferroni-corrected p-value smaller than 0.02). Tables 1 and 2 focus on predictive binomial deviance. Section 3.D in SI considers other error measurements, including the classification error and the AUC (Area Under Curve), where a similar conclusion can be drawn regarding the performance of different methods: predictions based on catalytic priors are generally much better than those based on the MLE and are often better than those based on the Cauchy prior.

Setting			Performance of Methods		
ζ	r		Cat.Boot	Cat.Joint	Cauchy
1/4	0.1	Cover	90.5%	88.1%	90.1%
		Width	3.5	2.9	3.3
	0.2	Cover	93.3%	97.2%	98.0%
		Width	2.8	2.7	3.0
	0.3	Cover	95.0%	97.6%	97.6%
		Width	2.2	2.4	2.8
2/4	0.1	Cover	89.8%	85.7%	86.2%
		Width	3.5	2.9	3.2
	0.2	Cover	93.4%	97.5%	98.4%
		Width	2.7	2.7	3.0
	0.3	Cover	95.7%	97.7%	97.7%
		Width	2.1	2.4	2.8
3/4	0.1	Cover	89.4%	85.6%	86.1%
		Width	3.5	2.9	3.2
	0.2	Cover	93.9%	97.6%	98.6%
		Width	2.7	2.7	3.0
	0.3	Cover	95.9%	97.8%	97.8%
		Width	2.1	2.4	2.7

Table 3. Average coverage probability (%) and width of 95% posterior intervals under the catalytic prior with $\hat{\tau}_{boot}$, the joint catalytic prior, and Cauchy prior. ζ is the non-sparsity; r is the oracle prediction error.

Table 3 presents the average coverage probabilities (in percentage) and widths of the 95% nominal intervals for β_j averaging over j. Because all the intervals given by the MLE have widths too large to be useful (thousands of times wider than those given by the other methods), we do not report them in this table. The intervals from the other three priors are reasonably short in all cases and have coverage rates not far from the nominal levels. Specifically, the intervals given by the Cauchy prior and the joint catalytic prior tend to over-cover when the true $\boldsymbol{\beta}$ has small amplitudes ($r = 0.2$ or 0.3) and tend to under-cover when $\boldsymbol{\beta}$ has large amplitudes ($r = 0.1$), whereas the intervals given by the catalytic prior with $\hat{\tau}_{boot}$ perform more consistently. This example, together with more results given in SI, illustrates that, for logistic regression, the catalytic prior is at least as good as the Cauchy prior. The SI also illustrates the performance of the catalytic prior in linear regression, where it is at least as good as ridge regression. Catalytic priors thus appear to provide a general framework for prior construction over a broad range of models.

Theoretical Properties of Catalytic Priors

We show the properness and the convergence of a catalytic prior when the working model is a GLM. Without loss of generality, we assume the sufficient statistic in the GLM formula Eq. (8) is $t(y) = y$; otherwise, we can let the response be $Y' = t(Y)$ and proceed. We assume that every covariate has at least two different observed values. Denote by $\mathcal{Y}$ the nonempty interior of the convex hull of the support of the model density in Eq. (8). Our results apply to any positive prior weight τ.

Properness. A proper prior is needed for many Bayesian inferences, such as model comparison using Bayes factors (32). We show that catalytic priors, population catalytic priors, and joint catalytic priors are generally proper, with proofs in SI.

Theorem 1. *Suppose (1) $\phi(\cdot)$ satisfies $\inf_{\eta\neq 0} |\phi(\eta)/\eta| > 0$, (2) the synthetic covariate matrix $\mathbb{X}^*$ has full column rank,*

DRAFT

and (3) each synthetic response Y_i^* lies in $\mathcal{Y}$ or there exists a linearly independent subset $\{X_{i_k}^*\}_{k=1}^p$ of the synthetic covariate vectors such that the average of synthetic responses with the same $X_{i_k}^*$ lies in $\mathcal{Y}$. Then the catalytic prior is proper for any $\tau > 0$.

The condition $\inf_{\eta \neq 0} |\phi(\eta)/\eta| > 0$ is satisfied for the canonical link for any GLM, and also for the commonly used probit link and the complementary log-log link in binary regression. The condition that $\mathbb{X}^*$ has full column rank holds with high probability according to the following result.

Proposition 2. *If each synthetic covariate vector is drawn from the independent resampling distribution, then there exists a constant $c > 0$ that only depends on the observed $\mathbb{X}$ such that for any $M > p$, with probability at least $1 - 2exp(-cM)$, the synthetic covariate matrix $\mathbb{X}^*$ has full column rank.*

Population catalytic priors are also proper.

Theorem 3. *Suppose (1) $\phi(\cdot)$ satisfies $\inf_{\eta \neq 0} |\phi(\eta)/\eta| > 0$, (2) the synthetic covariate vector is drawn from the independent resampling distribution, and (3) there exists a compact subset $\mathcal{Y}^{com} \subset \mathcal{Y}$ such that $\mathbf{P}(Y^* \in \mathcal{Y}^{com}) = 1$. Then the population catalytic prior is proper for any $\tau > 0$.*

The following result shows the properness of the joint prior $\pi_{\alpha,\gamma}(\tau, \beta)$ in Eq. (15) and the role of the hyperparameters.

Theorem 4. *Suppose α and γ are positive. If $\Gamma_{\alpha,\gamma}(\tau)$ equals* Eq. (16) *for linear regression or equals* Eq. (17) *for other generalized linear models. Then under the same condition as Theorem 1, (1) the joint prior is proper; (2) for any $m \in (0, \alpha)$, the m^{th} moment of β exists; (3) $\lim_{\tau \to \infty} \frac{1}{\tau} \log h_{\alpha,\gamma}(\tau) = -1/\gamma < 0$, where $h_{\alpha,\gamma}(\tau)$ denotes the marginal prior on τ.*

Convergence to the Population Catalytic Prior. When synthetic-sample size, M, is large enough, the randomness in the synthetic data will not affect the catalytic prior regardless of the observed real sample size because, as a distribution of the parameters, the catalytic prior converges to the population catalytic prior.

We can quantify how fast the catalytic prior, as a random distribution, converges to the population catalytic prior by establishing an explicit upper bound on the distance between these two distributions in terms of M. This result shows how large M needs to be so that the randomness in the synthetic data no longer influentially change the prior. We present here a simplified version of the theoretical result; for precise and detailed statements, see SI.

Theorem 5. *Under mild regularity conditions,*

1. *For any given τ and p, there exists a constant C_1, such that for any small positive ϵ_0, ϵ_1, and any $M \geq C_1 \left(1 + \log^2(\frac{1}{\epsilon_1})\right) \frac{1}{\epsilon_1^2} \log(\frac{1}{\epsilon_0})$, with probability at least $1-\epsilon_0$ the total variation distance between the catalytic prior and the population catalytic prior is bounded by*

$$d_{TV}(\pi_{cat,\infty}, \pi_{cat,M}) \leq \epsilon_1.$$

2. *If the working model is linear regression with Gaussian noise, then there exists a constant C_2 that only depends on the observed covariates, such that for any $\epsilon_0 > 0$ and any $M > \frac{16}{9} C_2^2 p \log(\frac{p}{\epsilon_0})$, with probability at least $1 - \epsilon_0$, the KL divergence between the catalytic prior and the population catalytic prior with any $\tau > 0$ is bounded by*

$$KL(\pi_{cat,\infty}, \pi_{cat,M}) \leq 2C_2 \sqrt{\frac{1}{M} p^3 \log(\frac{p}{\epsilon_0})}.$$

Data Availability. All the data used in the article are simulation data. The details, including the models to generate the simulation data, are described in the *Illustration of Methods* section and the *Additional Simulations* section of the SI.

Discussion

The class of catalytic prior distributions stabilizes the estimation of a relatively complicated working model by augmenting the actual data with synthetic data drawn from the predictive distribution of a simpler model (including but not limited to a sub-model of the working model). Our theoretical work and simulation-based evidence suggest that the resulting inferences using standard software, which treat the augmented data just like actual data, have competitive and sometimes clearly superior frequency operating characteristics, compared to inferences based on alternatives that have been previously proposed. Moreover, catalytic priors are generally easier to formulate because they are based on hypothetical smoothed data that resemble the actual data. Two tuning constants, M and τ, require selection, and wise choices for them appear to be somewhat model dependent, for example, differing for linear and logistic regressions, both of which are considered here. We anticipate that catalytic priors will find broad application, especially as more complex Bayesian models are fit to more and more complicated datasets. Some open questions for future investigation include: (1) how to apply the catalytic priors to model selection, (2) how to study the asymptotic properties when both the sample size and the dimension of the working model go to infinity — in such a regime, it is also interesting to investigate what the simple model should be in order to achieve good bias-variance tradeoffs.

ACKNOWLEDGMENTS. D.B.R.'s research is supported in part by National Science Foundation Grant IIS-1409177, National Institutes of Health Grant 1R01AI140854, and Office of Naval Research Grant N00014-17-1-2131. S.C.K.'s research is supported in part by National Science Foundation Grant DMS-1810914.

1. Clogg CC, Rubin DB, Schenker N, Schultz B, Weidman L (1991) Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association* 86(413):68–78.
2. Rubin DB (2004) *Multiple imputation for nonresponse in surveys.* (John Wiley & Sons) Vol. 81.
3. Hirano K, Imbens GW, Rubin DB, Zhou XH (2000) Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* 1(1):69–88.
4. Day NE, Kerridge DF (1967) A general maximum likelihood discriminant. *Biometrics* pp. 313–323.
5. Albert A, Anderson JA (1984) On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 71(1):1–10.
6. Firth D (1993) Bias reduction of maximum likelihood estimates. *Biometrika* 80(1):27–38.
7. Heinze G, Schemper M (2002) A solution to the problem of separation in logistic regression. *Statistics in Medicine* 21(16):2409–2419.
8. Rubin DB, Schenker N (1987) Logit-based interval estimation for binomial data using the Jeffrey's prior. *Sociological Methodology* 17:131–144.
9. Chen MH, Ibrahim JG, Kim S (2008) Properties and implementation of Jeffrey's prior in binomial regression models. *Journal of the American Statistical Association* 103(484):1659–1664.
10. Goodman LA (1968) The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries: Ra fisher memorial lecture. *Journal of the American Statistical Association* 63(324):1091–1131.
11. Good IJ (1983) *Good thinking: The foundations of probability and its applications.* (U of Minnesota Press).
12. Raiffa H, Schlaifer R (1961) *Applied statistical decision theory.* (Wiley Cambridge).
13. Kadane JB, Dickey JM, Winkler RL, Smith WS, Peters SC (1980) Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association* 75(372):845–854.

DRAFT

14. Bedrick EJ, Christensen R, Johnson W (1996) A new perspective on priors for generalized linear models. *Journal of the American Statistical Association* 91(436):1450–1460.
15. Bedrick EJ, Christensen R, Johnson W (1997) Bayesian binomial regression: Predicting survival at a trauma center. *The American Statistician* 51(3):211–218.
16. Greenland S, Christensen R (2001) Data augmentation priors for Bayesian and semi-Bayes analyses of conditional-logistic and proportional-hazards regression. *Statistics in Medicine* 20(16):2421–2428.
17. Greenland S (2001) Putting background information about relative risks into conjugate prior distributions. *Biometrics* 57(3):663–670.
18. Iwaki K (1997) Posterior expected marginal likelihood for testing hypotheses. *Journal of Economics, Asia University* 21:105–134.
19. Neal RM (2001) Transferring prior information between models using imaginary data, (Department of Statistics, University of Toronto), Technical Report 0108.
20. Pérez JM, Berger JO (2002) Expected-posterior prior distributions for model selection. *Biometrika* 89(3):491–512.
21. Ibrahim JG, Chen MH (2000) Power prior distributions for regression models. *Statistical Science* 15(1):46–60.
22. Chen MH, Ibrahim JG, Shao QM (2000) Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference* 84(1-2):121–137.
23. Ibrahim JG, Chen MH, Sinha D (2003) On optimality properties of the power prior. *Journal of the American Statistical Association* 98(461):204–213.
24. Fouskakis D, Ntzoufras I, Draper D (2015) Power-expected-posterior priors for variable selection in Gaussian linear models. *Bayesian Analysis* 10(1):75–107.
25. Fouskakis D, Ntzoufras I, Perrakis K (2018) Power-expected-posterior priors for generalized linear models. *Bayesian Analysis* 13(3):721–748.
26. Bernardo JM (1979) Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)* 41(2):113–128.
27. Brown PJ, Walker SG (2012) Bayesian priors from loss matching. *International Statistical Review* 80(1):60–82.
28. Simpson D, Rue H, Riebler A, Martins TG, Sørbye SH (2017) Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science* 32(1):1–28.
29. Box GE, Cox DR (1964) An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)* 26(2):211–243.
30. Efron B (2004) The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association* 99(467):619–632.
31. Gelman A, Jakulin A, Pittau MG, Su YS (2008) A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* 2(4):1360–1383.
32. Kass RE, Raftery AE (1995) Bayes factors. *Journal of the American Statistical Association* 90(430):773–795.

DRAFT