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; and [c]Yau Mathematical Sciences Center, Tsinghua University, Beijing 100084, China

Contributed by Donald B. Rubin, April 2, 2020 (sent for review December 2, 2019; reviewed by James O. Berger and Hal Stern)

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 with 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 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 is in ref. 1, which was based on an earlier proposal by Rubin in a 1983 report for the US 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 maximum likelihood 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 US 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 10 for some 1970 codes, and many of these logistic regression models faced complete separation. The successful approach in ref. 1 was to use the prior distribution

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

Significance

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.

Author contributions: D.B.R. and S.C.K. designed research; D.H., N.S., D.B.R., and S.C.K. performed research; D.H. contributed new reagents/analytic tools; D.H., D.B.R., and S.C.K. analyzed data; and D.H., N.S., D.B.R., and S.C.K. wrote the paper.

Reviewers: J.O.B., Duke University; and H.S., University of California, Irvine.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: rubin@stat.harvard.edu or kou@stat.harvard.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1920913117/-/DCSupplemental.

First published May 15, 2020.

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

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 independent of the covariates:

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

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**:

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

where $N_{j,1}, N_{j,0}$ are the numbers of $(1, x_j^*)$ and $(0, x_j^*)$, respectively, in the observed data. In this construction, the total amount of synthetic data is taken to be p, the dimension of $\boldsymbol{\beta}$ (*SI Appendix*, Remark 2.2 has 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 (GLMs). 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 *SI Appendix*, section 4).

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 pseudodata 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 and 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 proposed 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 $Y_i \overset{i.i.d.}{\sim} f(y \,|\, \theta)$ governed by unknown parameter θ, where i.i.d. stands for independent and identically distributed. Suppose a model $g(y \,|\, \psi)$ with unknown parameter ψ, whose dimension is smaller than that of θ, is stably fitted from $\boldsymbol{Y}$ and results in a predictive distribution $g_*(y^* \,|\, \boldsymbol{Y})$ for future data drawn from $g(y \,|\, \psi)$. The synthetic data-generating distribution $g_*(y^* \,|\, \boldsymbol{Y})$ 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 $f(y \,|\, \theta)$, but it does not necessarily have to be. Examples: (1) If a Bayesian analysis of the simpler model can be carried out easily, $g_*(y^* \,|\, \boldsymbol{Y})$ can be taken to be the posterior predictive distribution under the simpler model. (2) Alternatively, one can obtain a point estimate $\hat{\psi}$, and $g_*(y^* \,|\, \boldsymbol{Y}) = g(y^* \,|\, \hat{\psi})$ can be the plug-in predictive distribution. (3) If two simpler estimated models are $g_*^{(1)}(y^* \,|\, \boldsymbol{Y})$ and $g_*^{(2)}(y^* \,|\, \boldsymbol{Y})$, then $g_*(y^* \,|\, \boldsymbol{Y})$ can be taken to be a mixture $w\, g_*^{(1)}(y^* \,|\, \boldsymbol{Y}) + (1-w)\, g_*^{(2)}(y^* \,|\, \boldsymbol{Y})$ 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 $\ell(\theta \,|\, Y^*) = \prod_{i=1}^M f(Y_i^* \,|\, \theta)$. 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 density:

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

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:

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

Here, the expectation $\mathbb{E}_{g_*} \{\log f(Y^* \,|\, \theta)\}$ in Eq. **4** is taken with respect to $Y^* \sim g_*(Y^* \,|\, \boldsymbol{Y})$. The dependence of $g_*(Y^* \,|\, \boldsymbol{Y})$ 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 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 $\boldsymbol{\beta}$:

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

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 $f(\boldsymbol{Y} \,|\, \mathbb{X}, \boldsymbol{\beta}) = \prod_{i=1}^n f(Y_i \,|\, \boldsymbol{X}_i, \boldsymbol{\beta})$.

Suppose a simpler model $g(y \,|\, \boldsymbol{X}, \boldsymbol{\psi})$ with unknown parameter $\boldsymbol{\psi}$ is stably fitted from $(\boldsymbol{Y}, \mathbb{X})$ and results in a synthetic data-generating distribution $g_*(y \,|\, \boldsymbol{x}, \boldsymbol{Y}, \mathbb{X})$. 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 $g_*(y \,|\, \boldsymbol{x}, \boldsymbol{Y}, \mathbb{X})$, the catalytic prior first draws a set of synthetic data $\{(Y_i^*, \boldsymbol{X}_i^*)\}_{i=1}^M$ from

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

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 $\ell(\boldsymbol{\beta} \,|\, \boldsymbol{Y}^*, \mathbb{X}^*)$ equals $\prod_{i=1}^M f(Y_i^* \,|\, \boldsymbol{X}_i^*, \boldsymbol{\beta})$. 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:

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

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

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

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 Eqs. **6** and **7** with respect to $\boldsymbol{\beta}$. 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

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

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 GLM assumes that, given a covariate vector $\boldsymbol{X}$, the response Y has the following density with respect to some base probability measure:

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

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 Eqs. **7** and **8**, we have

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

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 $\mathbb{E}_{g_*}[t(Y^*) \,|\, \boldsymbol{X}^*]$ 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 submodel 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 $\lim_{M \to \infty} \frac{1}{M} (\mathbb{X}^*)^\top \mathbb{X}^* = \Sigma_X$, the population catalytic prior is

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

More details about this example can be found in *SI Appendix*.

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

[†] Note that in the previous example of 1970 to 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.

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 Appendix*.

Generating Synthetic Responses. The synthetic data-generating distribution can be specified by fitting a simple model $G_\Psi = \{g(y\,|\,\boldsymbol{x}, \boldsymbol{\psi}) : \boldsymbol{\psi} \in \Psi\}$ 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; *SI Appendix* has 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 submodel with only main effects. If the main-effect model is overfitted as well, we could use a mixed synthetic data-generating distribution, such as $g_*(y\,|\,\boldsymbol{x}, \boldsymbol{Y}, \mathbb{X}) = 0.5\, g_{*,1}(y\,|\,\boldsymbol{x}, \boldsymbol{Y}, \mathbb{X}) + 0.5\, g_{*,0}(y\,|\,\boldsymbol{x}, \boldsymbol{Y}, \mathbb{X})$, 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 submodel 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* 4 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 $M \geq \frac{4p^3}{\epsilon^2}\log(\frac{p}{\delta})$, then the Kullback–Leibler (KL) 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 four times the dimension of $\boldsymbol{\beta}$ (based on *Theorem* 1 and *Proposition* 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 $\Lambda(\tau)$. *SI Appendix*, section 2.C.1 has 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

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

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: $D(y_0, \hat{\mu}) = \boldsymbol{1}_{y_0 \neq \hat{\mu}} = \hat{\mu} - 2y_0\hat{\mu} + y_0$ 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 $f(\boldsymbol{y}\,|\,\mathbb{X}, \widehat{\boldsymbol{\beta}}^0)$, 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

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

where $\widehat{\mu}_{\tau,i} = \mu(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}(\tau))$ and $\hat{\mu}_{\tau,i}^{boot} = \mu(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}^{boot}(\tau))$. *SI Appendix* has a detailed derivation. In practice, the term $\mathrm{Cov}(\lambda(\hat{\mu}_{\tau,i}^{boot}),\, Y_i^{boot})$ 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. *SI Appendix* has a mathematical argument.

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

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

In particular, when squared 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

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

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

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

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 $(\tau, \boldsymbol{\beta})$:

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

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

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

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

$$\Gamma_{\alpha,\gamma}(\tau) = \tau^{\frac{p+\alpha}{2}-1} e^{-\tau(\kappa+\gamma^{-1})} \qquad \textbf{[16]}$$

and for other models,

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

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 (*Theorem* 3). 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* 3 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 toward 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 Appendix*. 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 $\hat{\mu}_0\cdot\boldsymbol{1}_M$, 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

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

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 one or zero. ζ proportion of the coordinates of $\boldsymbol{\beta}^{(0)}$ is randomly selected and set to 1, and the remaining $1-\zeta$ proportion is set to 0, where ζ is the level of nonsparsity 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: $\boldsymbol{\beta}_0=c_1,\quad \boldsymbol{\beta}_{1:(p-1)}=c_2\boldsymbol{\beta}^{(0)}_{1:(p-1)}$, 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, $r=\mathbb{E}_X(\min(\mathbf{P}_{\boldsymbol{\beta}}(Y=1),\mathbf{P}_{\boldsymbol{\beta}}(Y=0)))=\mathbb{E}_X\left(1+\exp(|\boldsymbol{X}^\top\boldsymbol{\beta}|)\right)^{-1}$ is numerically computed by sampling 2,000 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 to evaluate the predictive performance of $\hat{\boldsymbol{\beta}}$. The expectation $\mathbb{E}_{\boldsymbol{X}_0}$ is computed by sampling 1,000 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 coordinatewise 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 unimodal 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 1,600 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, bold 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 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

Table 1. Mean and SE of predictive binomial deviance of different methods

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

The first two columns are the settings of the simulation: ζ is the nonsparsity, 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 SE 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. Bold corresponds to the best-performing method in each simulation scenario.

Table 2. Mean and SE of the difference in predictive binomial deviance between the Cauchy posterior mode and the joint catalytic posterior median

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
1/4	0.2	0.033	0.91
1/4	0.3	0.047	0.86
1/2	0.1	0.007	0.79
1/2	0.2	0.028	0.85
1/2	0.3	0.046	0.84
3/4	0.1	0.003	0.76
3/4	0.2	0.026	0.83
3/4	0.3	0.044	0.82

ζ is the nonsparsity; r is the oracle prediction error.

predictive binomial deviance. *SI Appendix*, section 3.D considers other error measurements, including the classification error and the 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.

Table 3 presents the average coverage probabilities (in percentage) and widths of the 95% nominal intervals for β_j averaging over j. Because all of 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 overcover 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 Appendix*, illustrates that, for logistic regression, the catalytic prior is at least as good as the Cauchy prior. *SI Appendix* 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 that 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 Appendix*.

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, 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. *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* $1-2\exp(-cM)$, *the synthetic covariate matrix* $\mathbb{X}^*$ *has full column rank.*

Population catalytic priors are also proper.

Theorem 2. *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 $\pi_{\alpha,\gamma}(\tau,\boldsymbol{\beta})$ in Eq. **15** and the role of the hyperparameters.

Theorem 3. *Suppose* α *and* γ *are positive. If* $\Gamma_{\alpha,\gamma}(\tau)$ *equals Eq.* **16** *for linear regression or equals Eq.* **17** *for other GLMs, then under the same condition as Theorem* 1, (1) *the joint prior is proper*; (2) *for any* $m\in(0,\alpha)$, *the* mth *moment of* $\boldsymbol{\beta}$ *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 changes the prior. We present here a simplified version of the theoretical result; precise and detailed statements are in *SI Appendix*.

Table 3. Average coverage probability (percentage) and width of 95% posterior intervals under the catalytic prior with $\hat{\tau}_{boot}$, the joint catalytic prior, and Cauchy prior

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

ζ is the nonsparsity; r is the oracle prediction error.

Theorem 4. *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*

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

Data Availability. All of the data used in the article are simulation data. The details, including the models to generate the simulation data, are described in *Illustration of Methods* and *SI Appendix*, section 3.

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 submodel 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 with 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 and (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. The research of D.B.R. is supported in part by NSF Grant IIS-1409177, NIH Grant 1R01AI140854, and Office of Naval Research Grant N00014-17-1-2131. The research of S.C.K. is supported in part by NSF Grant DMS-1810914.

1. C. C. Clogg, D. B. Rubin, N. Schenker, B. Schultz, L. Weidman, Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *J. Am. Stat. Assoc.* **86**, 68–78 (1991).
2. D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys* (John Wiley & Sons, 2004), vol. 81.
3. K. Hirano, G. W. Imbens, D. B. Rubin, X. H. Zhou, Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* **1**, 69–88 (2000).
4. N. E. Day, D. F. Kerridge, A general maximum likelihood discriminant. *Biometrics* **23**, 313–323 (1967).
5. A. Albert, J. A. Anderson, On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**, 1–10 (1984).
6. D. Firth, Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38 (1993).
7. G. Heinze, M. Schemper, A solution to the problem of separation in logistic regression. *Stat. Med.* **21**, 2409–2419 (2002).
8. D. B. Rubin, N. Schenker, Logit-based interval estimation for binomial data using the Jeffrey's prior. *Socio. Methodol.* **17**, 131–144 (1987).
9. M. H. Chen, J. G. Ibrahim, S. Kim, Properties and implementation of Jeffrey's prior in binomial regression models. *J. Am. Stat. Assoc.* **103**, 1659–1664 (2008).
10. L. A. Goodman, The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries: Ra Fisher memorial lecture. *J. Am. Stat. Assoc.* **63**, 1091–1131 (1968).
11. I. J. Good, *Good Thinking: The Foundations of Probability and Its Applications* (University of Minnesota Press, 1983).
12. H. Raiffa, R. Schlaifer, *Applied Statistical Decision Theory* (Harvard University, Boston, MA, 1961).
13. J. B. Kadane, J. M. Dickey, R. L. Winkler, W. S. Smith, SC. Peters, Interactive elicitation of opinion for a normal linear model. *J. Am. Stat. Assoc.* **75**, 845–854 (1980).
14. E. J. Bedrick, R. Christensen, W. Johnson, A new perspective on priors for generalized linear models. *J. Am. Stat. Assoc.* **91**, 1450–1460 (1996).
15. E. J. Bedrick, R. Christensen, W. Johnson, Bayesian binomial regression: Predicting survival at a trauma center. *Am. Statistician* **51**, 211–218 (1997).
16. S. Greenland, R. Christensen, Data augmentation priors for Bayesian and semi-Bayes analyses of conditional-logistic and proportional-hazards regression. *Stat. Med.* **20**, 2421–2428 (2001).
17. S. Greenland, Putting background information about relative risks into conjugate prior distributions. *Biometrics* **57**, 663–670 (2001).
18. K. Iwaki, Posterior expected marginal likelihood for testing hypotheses. *J. Econ. Asia Univ.* **21**, 105–134 (1997).
19. R. M. Neal, "Transferring prior information between models using imaginary data" (Tech. Rep. 0108, Department of Statistics, University of Toronto, Toronto, Canada, 2001).
20. J. M. Pérez, J. O. Berger, Expected-posterior prior distributions for model selection. *Biometrika* **89**, 491–512 (2002).
21. J. G. Ibrahim, M. H. Chen, Power prior distributions for regression models. *Stat. Sci.* **15**, 46–60 (2000).
22. M. H. Chen, J. G. Ibrahim, Q. M. Shao, Power prior distributions for generalized linear models. *J. Stat. Plann. Inference* **84**, 121–137 (2000).
23. J. G. Ibrahim, M. H. Chen, D. Sinha, On optimality properties of the power prior. *J. Am. Stat. Assoc.* **98**, 204–213 (2003).
24. D. Fouskakis, I. Ntzoufras, D. Draper, Power-expected-posterior priors for variable selection in Gaussian linear models. *Bayesian Anal.* **10**, 75–107 (2015).
25. D. Fouskakis, I. Ntzoufras, K. Perrakis, Power-expected-posterior priors for generalized linear models. *Bayesian Anal.* **13**, 721–748 (2018).
26. J. M. Bernardo, Reference posterior distributions for Bayesian inference. *J. Roy. Stat. Soc. B* **41**, 113–128 (1979).
27. P. J. Brown, S. G. Walker, Bayesian priors from loss matching. *Int. Stat. Rev.* **80**, 60–82 (2012).
28. D. Simpson, H. Rue, A. Riebler, T. G. Martins, S. H. Sørbye, Penalising model component complexity: A principled, practical approach to constructing priors. *Stat. Sci.* **32**, 1–28 (2017).
29. G. E. Box, D. R. Cox, An analysis of transformations. *J. Roy. Stat. Soc. B* **26**, 211–243 (1964).
30. B. Efron, The estimation of prediction error: Covariance penalties and cross-validation. *J. Am. Stat. Assoc.* **99**, 619–632 (2004).
31. A. Gelman, A. Jakulin, M. G. Pittau, Y. S. Su, A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2**, 1360–1383 (2008).
32. R. E. Kass, A. E. Raftery, Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).

www.pnas.org

Supplementary Information for

Catalytic Prior Distributions with Applications to Generalized Linear Models

Dongming Huang, Nathan Stein, Donald B. Rubin, S.C. Kou

Corresponding Authors: S. C. Kou, Donald Rubin.
E-mails: kou@stat.harvard.edu, dbrubin@me.com

This PDF file includes:

Supplementary text
Tables S1 to S9
References for SI reference citations

Supporting Information Text

Here is an outline of the supplementary materials.

Section 1 shows the catalytic priors for linear regression, in which case they have closed-form expressions.

Section 2 gives details about the prior specification.

Section 3 provides additional simulation results.

Section 4 provides an information theory/optimization viewpoint and discusses an interesting extension.

Section 5 studies the theoretical properties of catalytic priors and proves the theorems of the main text.

1. Analytic Form of Catalytic Prior for Linear Regression

Consider a linear regression model $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathcal{I}_n)$ with known σ. The first column of $\mathbb{X}$, say $\mathbb{X}_0$, is a vector of 1's. The other columns $\mathbb{X}_j$ $(1 \le j \le p-1)$ are centered at 0. Suppose the synthetic-data generating model is a sub-model whose design matrix is a sub-matrix of $\mathbb{X}$, say $\mathbb{X}P_S^\top$, where the matrix $P_S \in \mathbb{R}^{s\times p}$ have rows $\{\boldsymbol{e}_i^\top : i \in S\}$. If the simpler model is fitted by the MLE, then the synthetic-data generating distribution is $Y^*|(\boldsymbol{X}^* = \boldsymbol{x}) \sim N(\boldsymbol{x}^\top \tilde{\boldsymbol{\beta}}_0, \sigma^2)$, where $\tilde{\boldsymbol{\beta}}_0 = P_S^\top (P_S \mathbb{X}^\top \mathbb{X} P_S^\top)^{-1} P_S \mathbb{X}^\top \boldsymbol{Y}$.

As a special case of GLM, one can use $\mathbb{X}^*\tilde{\boldsymbol{\beta}}_0$, the expected value of the synthetic response vector $\boldsymbol{Y}^*$, in the catalytic prior formula (i.e., replacing $\boldsymbol{Y}^*$ by its predictive mean $\mathbb{X}^*\tilde{\boldsymbol{\beta}}_0$) as explained in the section *Catalytic Prior for GLM* (see Eq. (9)) of the main text. The catalytic prior then has an analytic form

$$\pi_{cat,M}(\boldsymbol{\beta} \mid \tau) = \exp\left(-\frac{\tau}{2}\left\{\log(2\pi\sigma^2) + \frac{1}{\sigma^2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_0)^\top \left(\frac{1}{M}(\mathbb{X}^*)^\top \mathbb{X}^*\right)(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_0)\right\}\right) / Z_{\tau,M}, \qquad [1.1]$$

where the normalizing constant is

$$Z_{\tau,M} = \left(\frac{2\pi\sigma^2}{\tau}\right)^{p/2} \det{}^{-1/2}\left(\frac{1}{M}(\mathbb{X}^*)^\top \mathbb{X}^*\right) \exp\left(-\frac{\tau}{2}\log(2\pi\sigma^2)\right). \qquad [1.2]$$

This prior is proper when $(\mathbb{X}^*)^\top \mathbb{X}^*$ is positive definite, which is easily satisfied for $M > p$ even if the observed sample size is small. For any fixed value of $\tau > 0$, this catalytic prior is Gaussian:

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

In the limit of $M \to \infty$, the population catalytic prior is

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

where $\Sigma_{\boldsymbol{X}} = \lim_{M\to\infty} \frac{1}{M}(\mathbb{X}^*)^\top \mathbb{X}^*$. The key difference between the population catalytic prior and the commonly used Gaussian prior is that its mean $\tilde{\boldsymbol{\beta}}_0$ is determined by fitting the synthetic-data generating model to the observed data. If each synthetic covariate vector $\boldsymbol{X}_i^*$ is drawn from the independent resampling distribution, then $\Sigma_{\boldsymbol{X}}$ becomes a diagonal matrix with the jth entry being $\widehat{\sigma^2_{X,j}} := n^{-1}\mathbb{X}_j^\top \mathbb{X}_j$.

According to Bayes rule, the posterior distribution for $\boldsymbol{\beta}$ under the catalytic prior is:

$$\begin{aligned} &\boldsymbol{\beta}|(\boldsymbol{Y}, \mathbb{X}) \sim N(\widehat{\boldsymbol{\beta}}, \widehat{\Sigma}), && [1.3] \\ &\text{where } \widehat{\boldsymbol{\beta}} := (\mathbb{X}^\top\mathbb{X} + \frac{\tau}{M}(\mathbb{X}^*)^\top \mathbb{X}^*)^{-1}(\mathbb{X}^\top \boldsymbol{Y} + \frac{\tau}{M}(\mathbb{X}^*)^\top \mathbb{X}^* \tilde{\boldsymbol{\beta}}_0), && [1.4] \\ &\qquad\quad \widehat{\Sigma} := \sigma^2(\mathbb{X}^\top\mathbb{X} + \frac{\tau}{M}(\mathbb{X}^*)^\top \mathbb{X}^*)^{-1}. && [1.5] \end{aligned}$$

Since $\mathbb{X}^*\tilde{\boldsymbol{\beta}}_0 = \boldsymbol{A}\boldsymbol{Y}$, where $\boldsymbol{A} := \mathbb{X}^* P_S^\top (P_S \mathbb{X}^\top \mathbb{X} P_S^\top)^{-1} P_S \mathbb{X}^\top$, $\widehat{\boldsymbol{\beta}}$ can be written as

$$\widehat{\boldsymbol{\beta}} = (\mathbb{X}^\top\mathbb{X} + \frac{\tau}{M}(\mathbb{X}^*)^\top \mathbb{X}^*)^{-1}(\mathbb{X}^\top + \frac{\tau}{M}(\mathbb{X}^*)^\top \boldsymbol{A})\boldsymbol{Y}.$$

The predictive mean vector for the response given the observed covariates is $\widehat{\boldsymbol{\mu}} = \mathbb{X}\widehat{\boldsymbol{\beta}}$ and can be written as $\widehat{\boldsymbol{\mu}} = \boldsymbol{H}\boldsymbol{Y}$ for a matrix $\boldsymbol{H}$. More generally, the form $\widehat{\boldsymbol{\mu}} = \boldsymbol{H}\boldsymbol{Y} + \boldsymbol{c}$ holds for any synthetic-data generating distribution as long as the conditional mean of $\boldsymbol{Y}^*$ depends linearly on $\boldsymbol{Y}$. This result is used in the section *Frequentist Predictive Risk Estimation* of the main text when we specify the weight parameter τ using Eq. (13) of the main text.

As a specific example, if the prior generating model is the intercept-only model, then $\mathbb{X}P_S = \mathbf{1}_n$, $\mathbb{E}_{g_*}[\boldsymbol{Y}^* \mid \mathbb{X}^*] = \bar{Y}\mathbf{1}_M = n^{-1}\mathbf{1}_M\mathbf{1}_n^\top\boldsymbol{Y}$ and

$$\widehat{\boldsymbol{\mu}} = \mathbb{X}\widehat{\boldsymbol{\beta}} = \mathbb{X}(\mathbb{X}^\top\mathbb{X} + \frac{\tau}{M}(\mathbb{X}^*)^\top\mathbb{X}^*)^{-1}(\mathbb{X}^\top + \frac{\tau}{Mn}(\mathbb{X}^*)^\top\mathbf{1}_M\mathbf{1}_n^\top)\boldsymbol{Y}.$$

2. Details about Prior Specification

A. Synthetic-Covariate Generation. In addition to drawing the synthetic covariates from the independent resampling distribution, we discuss three other variants.

The first variant aims at reducing the skewness in a continuous covariate X_j by the idea of *symmetrizing*. Suppose the covariate $(X^0)_j$ is drawn from the empirical distribution of X_j, we randomly flip $(X^0)_j$ around the observed sample mean $\bar{\boldsymbol{X}}_{\cdot j}$. Formally, let ξ_j be independently drawn from -1 and 1 with equal probability, then set $(\boldsymbol{X}^*)_j = \bar{\boldsymbol{X}}_{\cdot j} + \xi_j((X^0)_j - \bar{\boldsymbol{X}}_{\cdot j})$. We call this proposal *the symmetrizing resampling*.

The second variant aims at increasing the diversity of the sampling distribution for a continuous covariate by *smoothing*. An example of smoothing is to perturb each resampled covariate value by adding a small noise centered at zero (with the noise variance chosen appropriately). This variant allows a synthetic covariate to take values different from the observed ones and thus avoids potential rank deficiency in the working model when higher-order terms of the covariate(s) are used.

The third variant aims at balancing a categorical covariate by the idea of *flattening*. A categorical covariate X_j is called imbalanced if some of its observed frequency counts are smaller than $n/(2 \cdot \#K_j)$, where K_j is the set of all categories of X_j. Such imbalance would cause problem in practice, as a categorical variable is often coded into dummy variables that take the value 0 or 1 and then are standardized to have mean 0 and variance 1. In an extreme case when some category only has frequency count being 1, the corresponding standardized dummy variable will take values $\sqrt{n-1}$ and $-1/\sqrt{n-1}$. Such an extreme case would make the catalytic prior behaves poorly. In cases like this, a simple remedy is to mix the empirical distribution with the uniform distribution on K_j to generate a synthetic covariate. Suppose $(X^0)_j$ is drawn from the empirical distribution of X_j, then with probability $1/2$, we either keep $(X^0)_j$ or draw a new $(\boldsymbol{X}^*)_j$ uniformly from K_j. Formally, this sampling can be expressed as

$$\mathbf{P}_j(\boldsymbol{X}_j^* = x) = \frac{1}{2n}\#\{1 \le i \le n : (\boldsymbol{X}_i)_j = x\} + \frac{1}{2 \cdot \#K_j}, \ \forall x \in K_j.$$

We call this proposal *the flattening resampling*. It guarantees that the standardized dummy variables for X_j^* are uniformly bounded by $\sqrt{2 \cdot \#K_j - 1}$ regardless of n.

In general, in the independent sampling, if the observed X_j is neither skewed nor imbalanced and the discreteness of its empirical distribution does not cause rank deficiency of the working model, the coordinate X_j^* is drawn from the empirical distribution of X_j; otherwise, it is recommended to draw X_j^* by using one of the above variants (the symmetrizing, smoothing, and flattening resampling) accordingly.

We want to emphasize that the independent resampling and its variants discussed here are by no means exhaustive. Other ways can be used to generate synthetic covariates. For example, if historical data are available, a synthetic covariate vector can be resampled from the historical covariates. One can also generate synthetic covariates through block resampling: grouping the covariates into disjoint blocks and sampling the covariates within the same block jointly (i.e., sampling the covariate vector in a block). This sampling scheme allows one to accommodate the correlations among the covariates.

B. Synthetic-Responses Generation. The synthetic responses are generated from a stably fitted simple model. The only requirement for a simple model is that the standard estimation for it using either Bayesian or frequentist approach can lead to a well-defined predictive distribution for future data. Let q be the dimension of the simple model and n be the observed sample size. When q/n is small, different fitting procedures often make little difference in the predictive distribution.

We have suggested several choices of the simple model in the main paper: a fixed distribution, an intercept-only model, and also a regression model based on the first few principal components. We present an experiment that compares these choices in Section 3.E. Among candidate simple models, we did not suggest a single choice so that users can also incorporate extra information, such as previous study and domain knowledge.

As discussed in section *Catalytic Prior for GLMs* of the main text, in GLMs and exponential families a synthetic response may be replaced by the expectation of the sufficient statistic. However, sometimes such expectations may not be easy to compute, and in these cases we suggest to sample multiple synthetic responses for each sampled synthetic covariate vector. In the main text, this sampling scheme is termed the *stratified synthetic data generation.* The stratified synthetic data generation is especially useful for binary regression because it usually ensures the properness of the catalytic prior. See Corollary 5.15.

C. Selection of the Prior Weight via Frequentist Predictive Risk Estimation. In this section, we provide details of the prior weight selection via the frequentist approach presented in the main text.

***C.1.* $\Lambda(\tau)$ *in* Eq. (11).** The discussion here closely follows Ref. (1).

Given a discrepancy $D(y_0, \hat{\mu})$, the conditional prediction error of a prediction rule $r(\boldsymbol{x})$ is define as the expected discrepancy between a new response Y_0 and the predictive value $r(\boldsymbol{x}_0)$ given the covariate $\boldsymbol{x}_0$, i.e., $\mathbb{E}_{Y_0|\boldsymbol{x}_0}[D(Y_0, r(\boldsymbol{x}_0))]$.

The in-sample prediction error of a prediction rule $r(\cdot)$ is defined as the average conditional prediction error with $\boldsymbol{x}_0$ being uniformly sampled from the observed covariate vectors, i.e.,

$$\operatorname{Err}(r) \;:=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Y_{0,i}|(\boldsymbol{x}_0=\boldsymbol{X}_i)}[D(Y_{0,i}, \hat{\mu}_i)],$$

where $Y_{0,i}|(\boldsymbol{x}_0 = \boldsymbol{X}_i) \overset{d}{=} Y_i|\boldsymbol{X}_i$ and $\hat{\mu}_i := r(\boldsymbol{X}_i)$ is the prediction given by $r(\cdot)$. In reality, the prediction rule $r(\cdot)$ itself and thus Err depend on the observed response $\boldsymbol{Y}$, we need to estimate the expectation of Err. First, consider the plug-in estimate $D(\boldsymbol{Y}, r(\mathbb{X})) = \frac{1}{n}\sum_{i=1}^{n} D(Y_i, \widehat{\mu}_i)$. For the class of discrepancy functions in Eq. (10) of the main text, the bias of this plug-in estimate is

$$\begin{aligned}\mathbb{E}_{\boldsymbol{Y}|\mathbb{X}}[D(\boldsymbol{Y}, r(\mathbb{X}))] - \mathbb{E}_{\boldsymbol{Y}|\mathbb{X}}[\operatorname{Err}(r)] &= -\mathbb{E}_{\boldsymbol{Y}|\mathbb{X}}\left(\frac{1}{n}\sum_{i=1}^{n}\lambda(\widehat{\mu}_i)\left(Y_i - \mathbb{E}_{\boldsymbol{Y}_0|(\boldsymbol{x}_0=\boldsymbol{X}_i)}(Y_{0,i})\right)\right)\\ &= -\frac{1}{n}\sum_{i=1}^{n}\operatorname{Cov}_{\boldsymbol{Y}|\mathbb{X}}(\lambda(\widehat{\mu}_i), Y_i), \end{aligned} \qquad [2.1]$$

where we have used the fact that $\mathbb{E}_{\boldsymbol{Y}_0|(\boldsymbol{x}_0=\boldsymbol{X}_i)}(Y_{0,i}) = \mathbb{E}_{\boldsymbol{Y}|\mathbb{X}}(Y_i)$ and $\mathbb{E}_{\boldsymbol{Y}_0|(\boldsymbol{x}_0=\boldsymbol{X}_i)}[c(Y_{0,i})] = \mathbb{E}_{\boldsymbol{Y}|\mathbb{X}}[c(Y_i)]$, since $\boldsymbol{Y}_0$ is an independent copy of $\boldsymbol{Y}$. Following Ref. (2), define *covariance penalty* $\Omega(r)$ as

$$\Omega \;:=\; \frac{1}{n}\sum_{i=1}^{n}\operatorname{Cov}(\lambda(\widehat{\mu}_i), Y_i).$$

Once an estimate $\widehat{\Omega}$ for Ω is found, Eq. (2.1) suggests to use $D(\boldsymbol{Y}, r(\mathbb{X})) + \widehat{\Omega}$ as an estimate for $\mathbb{E}[\operatorname{Err}]$, which leads to the criterion function $\Lambda(\tau)$ in Eq. (11) of the main text, where Ω is estimated by parametric bootstrap.

C.2. Posterior Mode. We focus on the prediction given by the posterior mode because of its computational simplicity: the posterior mode can be computed using the MLE procedure in many statistical software by weighting each synthetic data point with weight τ/M. Other predictions involving the posterior distribution, such as the plug-in estimate with the posterior mean

$$\mathbf{P}_{\hat{\theta}}(\cdot), \text{ where } \hat{\theta} = \int \theta\ \pi(\theta|\boldsymbol{Y})d\theta,$$

and the Bayesian posterior prediction

$$\int \mathbf{P}_\theta(\cdot)\pi(\theta|\boldsymbol{Y})d\theta,$$

usually rely on intensive Monte Carlo sampling and take a long time to compute due to the high dimensional integrals.

C.3. Computation of $\Lambda(\tau)$**.** Once a criterion function $\Lambda(\tau)$ to estimate the (expected) prediction error is determined, we choose the value of τ that minimizes $\Lambda(\tau)$. The computation of $\Lambda(\tau)$ is often conducted on a pre-determined fixed grid, either a linear grid or a geometric grid. For example, in our simulation example with linear regression, we choose a grid like $\{c_0\theta^k : 0 \le k \le 100\}$ for some $\theta \in (0, 1)$ and integer k; in our experiments with logistic regression, the grid is taken as $\{kp/4 : k = 1, \ldots, 8\}$ for the ease of computation.

C.4. Parametric Bootstrap. To estimate Ω for general prediction rules, the *parametric bootstrap* begins with a preliminary estimator $\widehat{\boldsymbol{\beta}}^0$ and samples $\boldsymbol{Y}^b$ from the distribution $f(\boldsymbol{Y}|\mathbb{X}, \widehat{\boldsymbol{\beta}}^0)$. The bootstrap estimate of Ω is

$$\widehat{\Omega}_{boot} := \frac{1}{n}\sum_{i=1}^{n} \mathrm{Cov}(\widehat{\lambda}_i(\boldsymbol{Y}^b), Y_i^b).$$

For each $i \in \{1, \ldots, n\}$, $\mathrm{Cov}(\widehat{\lambda}_i(\boldsymbol{Y}^b), Y_i^b)$ can be numerically approximated by repeatedly sampling $\boldsymbol{Y}^b$ for B times:

$$\frac{1}{B}\sum_{k=1}^{B} \widehat{\lambda}_i(\boldsymbol{Y}^{b,k})(Y_i^{b,k} - \bar{Y}_i^b), \quad \text{where } \bar{Y}_i^b = \frac{1}{B}\sum_{k=1}^{B} Y_i^{b,k}.$$

The preliminary estimator $\widehat{\boldsymbol{\beta}}^0$ is important for $\widehat{\Omega}_{boot}$ to approximate Ω well. A simple choice is the posterior mode under catalytic prior with some small τ_0.

C.5. Steinian Estimate. Ref. (1) introduced the Steinian estimate of Ω for binary observations. First, define the conditional covariance penalty $cov_{(i)} := \mathbb{E}_{Y_i|\mathbb{X}}\left(\lambda(\widehat{\mu}_i)\cdot(Y_i - \mu_i)\middle|\boldsymbol{Y}_{(-i)}, \mathbb{X}\right)$, where $\boldsymbol{Y}_{(-i)}$ is the response vector that excludes Y_i. Then Ω can be rewritten as the expected summation of $cov_{(i)}$

$$\Omega = \sum_i \mathbb{E}\, cov_{(i)}.$$

Second, we can estimate $cov_{(i)}$ by flipping the observed response as explained next.

Given i, let us write $\lambda(\widehat{\mu}_i) = \widehat{\lambda}_i(\boldsymbol{Y}_{(-i)}, Y_i)$. Noting that $\mathbf{P}(Y_i = 1|\boldsymbol{Y}_{(-i)}, \mathbb{X}) = \mathbf{P}(Y_i = 1|\mathbb{X}) = \mu_i$, it follows that

$$cov_{(i)} = \mu_i(1-\mu_i)\left(\widehat{\lambda}_i(\boldsymbol{Y}_{(-i)}, 1) - \widehat{\lambda}_i(\boldsymbol{Y}_{(-i)}, 0)\right).$$

For any $y \in \{0, 1\}$ and any function f, it always holds that $f(1) - f(0) = (2y - 1)[f(y) - f(1 - y)]$. Hence,

$$cov_{(i)} = \mu_i(1-\mu_i)(2Y_i - 1)\left(\widehat{\lambda}_i(\boldsymbol{Y}_{(-i)}, Y_i) - \widehat{\lambda}_i(\boldsymbol{Y}_{(-i)}, 1 - Y_i)\right).$$

Here $\widehat{\lambda}_i(\boldsymbol{Y}_{(-i)}, Y_i)$ is simply $\lambda(\widehat{\mu}_i)$, and $\widehat{\lambda}_i(\boldsymbol{Y}_{(-i)}, 1 - Y_i)$ can be computed by first obtaining the posterior mode $\widehat{\boldsymbol{\beta}}^{\triangle i}(\tau)$ based on data $(\boldsymbol{X}, \boldsymbol{Y}^{\triangle i})$, where $\boldsymbol{Y}^{\triangle i}$ is a copy of $\boldsymbol{Y}$ but with only Y_i replaced by $1 - Y_i$, and then computing the predictive mean $\widehat{\mu}_{\tau,i}^{\triangle i} = \mu(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}^{\triangle i}(\tau))$, yielding $\widehat{\lambda}_i(\boldsymbol{Y}_{(-i)}, 1 - Y_i) = \lambda(\widehat{\mu}_{\tau,i}^{\triangle i})$.

Using a plug-in estimate for $\mu_i(1 - \mu_i)$, say $\widehat{\mu}_i^0(1 - \widehat{\mu}_i^0)$, where $\widehat{\mu}_i^0 = \mu(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}^0)$, we obtain the Steinian estimate:

$$\widehat{\Omega}_{stein} := \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_i^0(1 - \widehat{\mu}_i^0)(2Y_i - 1)\left(\lambda(\widehat{\mu}_{\tau,i}) - \lambda(\widehat{\mu}_{\tau,i}^{\triangle i})\right),$$

arriving at Eq. (14) of the main text. Note that calculating the Steinian estimate requires n extra fitting procedures, each of which uses a different response vector $\boldsymbol{Y}^{\triangle i}$. The computational advantage of the Steinian estimate over the bootstrap estimate is significant when n is much smaller than the number of bootstrap replications.

C.6. The Initial Value τ_0. Both the parametric bootstrap and the Steinian estimates for Ω require a small τ_0 to begin with. On the one hand, we want τ_0 to be small so that the prior does not make $\widehat{\Omega}$ biased. On the other hand, it cannot be too small in order to provide reasonable initial estimate. Consider logistic regression with complete separation, where the model fits the data perfectly, as an example. If τ_0 is too small, the parametric bootstrap distribution degenerates and the bootstrap sample $\boldsymbol{Y}^b$ will always be identical to the observed $\boldsymbol{Y}$. In this case $\widehat{\Omega}_{boot} \approx 0$ and is not a good estimate of Ω. Similarly, the Steinian with too small τ_0 results in $\widehat{\sigma_i^2} \approx 0$ and $\widehat{\Omega}_{stein} \approx 0$.

The choices of $\tau_0 = p/4$ for logistic regression and $\tau_0 = 1$ for linear regression works well empirically. An intuitive explanation is to look at the non-singularity of $\boldsymbol{H}$, the Hessian matrix of the negative log likelihood, at the posterior mode with the augmented data set, which roughly represents the stability of the posterior mode. For linear regression and any value of $\tau > 0$, as long as the synthetic covariate matrix has full column rank, $\boldsymbol{H}$ is strictly positive definite, and its smallest singular value is no smaller than τ/σ^2 times the smallest singular value of $(\mathbb{X}^*)^\top \mathbb{X}^*/M$. Therefore we choose $\tau_0 = 1$ for a heuristic interpretation of adding effectively one synthetic data point. This is not the case for logistic regression. For logistic regression,

$$\begin{aligned}\boldsymbol{H} =& \sum_{i=1}^{n} \mu(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}(\tau))(1-\mu(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}(\tau)))\boldsymbol{X}_i\boldsymbol{X}_i^\top \\ &+ \frac{\tau}{M}\sum_{i=1}^{M} \mu((\boldsymbol{X}_i{}^*)^\top \widehat{\boldsymbol{\beta}}(\tau))(1-\mu((\boldsymbol{X}_i{}^*)^\top \widehat{\boldsymbol{\beta}}(\tau)))\boldsymbol{X}_i^*(\boldsymbol{X}_i{}^*)^\top,\end{aligned}$$

where $\mu(t) = 1/(1+e^{-t})$. When complete separation occurs and τ is small, $\widehat{\boldsymbol{\beta}}(\tau)$ can be so large that all $\mu((\boldsymbol{X}_i{}^*)^\top \widehat{\boldsymbol{\beta}}(\tau))$ and all $\mu(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}(\tau))$ are close to either 0 or 1. In this case, $\boldsymbol{H}$ can be nearly degenerated regardless of $(\mathbb{X}^*)^\top \mathbb{X}^*/M$. Choosing $\tau_0 = p/4$ effectively adds 1/4 synthetic data points for estimating each parameter and thus prevents such a degenerate situation from occuring in logistic regression.

C.7. Other Prediction Error Estimate. There are other estimates for prediction errors discussed in the literature. Here we briefly mention the cross-validation. A comprehensive introduction can be found in Ref. (3).

The K-fold Cross-validation randomly divides the data into K disjoint subsets. For every $k \leq K$, the model is fitted from all data except those in the kth fold $\boldsymbol{y}^{(k)}$, and prediction $\widehat{\boldsymbol{\mu}}^{(k)}$ for the omitted data $\boldsymbol{y}^{(k)}$ is made based on this fitted model. The kth test error is defined by $D(\boldsymbol{y}^{(k)}, \widehat{\mu}^{(k)})$. The cross-validation estimate is the average of all K test errors, and can be used as a criterion function to select tuning parameters. Note that the cross-validation estimate is based on the idea of making the validation dataset independent of the training dataset and is thus different from the criterion function $\Lambda(\tau)$ introduced in the main text. When we tried to use cross-validation to select τ, we found that it gave poor result. The primary reason is than in the cases of small sample size, the variance of cross-validation estimate is too large. The small-sample-size case is in fact the case when priors, such as catalytic priors, are preferred to be used. Thus, we instead use other model-based methods in the main text. Our experience echoed the discussion in Ref. (3, Chapter 12.3) and Ref. (4).

C.8. Connection with other practices in the literature. We conclude our discussion of the prior weight selection via the frequentist approach by remarking its connection with other practices in the literature.

Remark 2.1. In the power prior development (see Ref. (5)), the weight on the historical data is often taken to be less than 1. Although this is not required by catalytic priors, the weight we put on each synthetic data point, τ/M, is often less than 1 because M (in the order of hundreds of thousands) is typically much larger than τ. □

Remark 2.2. In the example of 1970-1980 I/O code mapping (see Ref. (6)) discussed in the main text, the prior weight there was chosen to be the same as the dimension of $\boldsymbol{\beta}$, and this choice was found to work well in practice. In light of the catalytic priors, we can now give a justification for this observation. The synthetic responses used in this example were the sample mean $\hat{\mu}$, and the synthetic covariate vectors satisfied $M^{-1}\sum_{i=1}^{M} \boldsymbol{X}_i^*(\boldsymbol{X}_i{}^*)^\top = \boldsymbol{\Sigma}_* \in \mathbb{R}^{p\times p}$. Let $\beta_0^* = \log(\frac{\hat{\mu}}{1-\hat{\mu}})$. Then $\boldsymbol{\beta}^* = (\beta_0^*, 0, \ldots, 0)^\top$ is the parameter of the distribution that generates the synthetic responses. If we approximate the log-density in Eq. (1) in the main text by its second order expansion, then the prior distribution is approximately a normal distribution $N(\boldsymbol{\beta}^*, \boldsymbol{H}_0^{-1})$, where

$$\boldsymbol{H}_0 = \frac{\tau}{M}\sum_{i=1}^{M}\hat{\mu}(1-\hat{\mu})\boldsymbol{X}_i^*(\boldsymbol{X}_i{}^*)^\top = \tau\hat{\mu}(1-\hat{\mu})\boldsymbol{\Sigma}_*.$$

Hence, the average prior variance of $(\boldsymbol{X}^*)^\top\boldsymbol{\beta}$ is approximately

$$\begin{aligned}&\frac{1}{M}\sum_{k=1}^{M}(\boldsymbol{X}^*)_k^\top\boldsymbol{H}_0^{-1}\boldsymbol{X}_k^* = \frac{1}{\tau\hat{\mu}(1-\hat{\mu})}\frac{1}{M}\sum_{k=1}^{M}(\boldsymbol{X}^*)_k^\top\boldsymbol{\Sigma}_*^{-1}\boldsymbol{X}_k^*\\ =&\frac{1}{\tau\hat{\mu}(1-\hat{\mu})}\mathrm{Tr}\left(\boldsymbol{\Sigma}_*^{-1}\frac{1}{M}\sum_{k=1}^{M}\boldsymbol{X}_k^*(\boldsymbol{X}_k{}^*)^\top\right) = \frac{1}{\tau\hat{\mu}(1-\hat{\mu})}\mathrm{Tr}\left(\boldsymbol{I}_p\right) = \frac{p}{\tau\hat{\mu}(1-\hat{\mu})}.\end{aligned}$$

This expression gives a justification for choosing $\tau = p$ as in Ref. (6). By doing so, the average prior variance of $(\boldsymbol{X}^*)^\top\boldsymbol{\beta}$ is approximately constant across models that include different covariates, and we, therefore, expect robust performance of this specific choice. □

D. Bayesian Joint Priors.

D.1. Computation. The Gibbs algorithm can be used to sample from the posterior distribution under the joint prior $\pi_{\alpha,\gamma}(\tau, \boldsymbol{\beta})$. The Gibbs algorithm iteratively samples one component from the conditional distribution holding the other component fixed.

Given $\boldsymbol{\beta}$, an update of τ can be sampled from the Gamma distribution

$$\begin{aligned}\pi_{\alpha,\gamma}(\tau|\boldsymbol{\beta}, \boldsymbol{Y}, \mathbb{X}) &\propto \Gamma_{\alpha,\gamma}(\tau)\mathrm{exp}(\frac{\tau}{M}\sum_{i=1}^{M}\log(f(Y_i^*|\boldsymbol{\beta}^\top\boldsymbol{X}_i{}^*))\\ &= \tau^{c-1}\mathrm{exp}\left(-\tau(\kappa + \frac{1}{\gamma} - \frac{1}{M}\sum_{i=1}^{M}\log(f(Y_i^*|\boldsymbol{\beta}^\top\boldsymbol{X}_i{}^*)))\right),\end{aligned}$$

where $c = (p+\alpha)/2$ for linear regression, and $c = p + \alpha$ for other models.

Given τ, an update of $\boldsymbol{\beta}$ should be drawn from $\pi(\boldsymbol{\beta}|\tau, Y, \mathbb{X}) \propto f(\boldsymbol{Y}^*|\mathbb{X}^*, \boldsymbol{\beta})^{\tau/M}f(\boldsymbol{Y}|\mathbb{X}, \boldsymbol{\beta})$. It can be sampled by various methods such as the Metropolis-Hasting algorithm and the Hamiltonian Monte Carlo (HMC). We recommend to use HMC with random step size and with adaptive variances for the momentum variables. Before running HMC within Gibbs, the adaptive variances are set at the diagonal entries of the inverse Hessian matrix of the negative log density at the posterior mode. The initial point of such a MCMC step should be the most recent sample of $\boldsymbol{\beta}$.

D.2. Default Choice of (α, γ). It is generally difficult to find a specification of (α, γ) that works optimally in all cases. Nevertheless, a guideline for a reasonable choice of (α, γ) is: (i) The posterior distribution will give meaningful answer about the estimand in most cases; (ii) the predictive performance is not significantly worse than the alternative methods for finite sample within the range of population that we are mostly interested in, and (iii) the coverage rate of the interval estimate should be close to or higher than the nominal coverage within the range of interesting populations.

We recommend the simple choice of $(\alpha, \gamma) = (2, 1)$ as the default based on our numerical experiments on linear regression and logistic regression models. We found this choice yields competitive estimation in most cases. Of course, our experiments were not exhaustive and did not cover all possible underlying distributions.

3. Additional Simulations

We provide additional numerical experiments on synthetic data and compare the performance of catalytic prior to other methods.

A. Logistic Regression with Interaction Terms. In this simulation, we consider logistic regression with interaction terms. Suppose there are q covariates. Then there are q main effects and $q(q-1)/2$ two-way interaction terms in total, so the total number of regression coefficients (i.e., the number of parameters) is $p = 1 + q + q(q-1)/2$. In this example, we generate the covariates as described in the *Illustration of Methods* Section of the main text. The regression coefficient $\boldsymbol{\beta}$ is specified similarly to the main text but with a different $\boldsymbol{\beta}^{(0)}$. We first randomly select half of the main effect terms in $\boldsymbol{\beta}^{(0)}$ to be 1 with the rest to be 0, and then randomly select $\lfloor q(q-1)/2 \times \zeta \rfloor$ interaction terms to be 1 with the rest to be 0, where ζ is the level of *non-sparsity* and is set at 0.1, 0.2, 0.4. The amplitude of $\boldsymbol{\beta}$ is specified through the oracle prediction error r in the same way as the main text, whose levels are 0.1, 0.2, 0.3. We set $q = 8$, which gives $p = 37$, and the observed sample size $n = 60$ in this example. Following the experiment in the main text, we use the predictive binomial deviance to evaluate the prediction performance of an estimator $\hat{\boldsymbol{\beta}}$.

Setting			Performance of Methods				
		Comp.		Cat.	Cat.	Cauchy	MLE
ζ	r	Sep.		Boot.	Joint		(pseudo)
0.1	0.1	97%	Mean	**1.759**	1.851	1.846	2.058
			SE $\times 10^3$	(5.4)	(5.3)	(5.3)	(21.9)
	0.2	85%	Mean	**0.704**	0.827	0.842	1.166
			SE $\times 10^3$	(2.9)	(2.5)	(2.4)	(8.1)
	0.3	70%	Mean	**0.305**	0.440	0.480	0.838
			SE $\times 10^3$	(1.7)	(1.5)	(1.3)	(2.1)
0.2	0.1	96%	Mean	**1.756**	1.846	1.848	2.119
			SE $\times 10^3$	(4.9)	(4.8)	(4.8)	(22.7)
	0.2	86%	Mean	**0.708**	0.826	0.849	1.171
			SE $\times 10^3$	(2.6)	(2.3)	(2.2)	(5.4)
	0.3	71%	Mean	**0.306**	0.438	0.483	0.843
			SE $\times 10^3$	(1.7)	(1.5)	(1.3)	(2.0)
0.4	0.1	96%	Mean	**1.772**	1.861	1.868	2.163
			SE $\times 10^3$	(4.9)	(4.8)	(4.8)	(22.0)
	0.2	84%	Mean	**0.713**	0.829	0.858	1.180
			SE $\times 10^3$	(2.5)	(2.2)	(2.2)	(5.0)
	0.3	69%	Mean	**0.307**	0.439	0.487	0.845
			SE $\times 10^3$	(1.7)	(1.5)	(1.4)	(2.1)

Table S1. (Logistic regression with interaction terms) Mean and standard error of average predictive binomial deviance of the catalytic posterior mode with $\hat{\tau}_{boot}$, the posterior median under the joint catalytic prior, the Cauchy posterior mode, and the MLE. ζ is the non-sparsity. r is the oracle prediction error. The column of *Comp.Sep.* shows how often complete separation occurs in the data sets. The boldface corresponds to the best performing method in each simulation scenario.

Table S1 presents the average predictive binomial deviance over 1600 simulations in each cell. The first column 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. Table S1 shows that the catalytic posterior mode with $\hat{\tau}_{boot}$ predicts the best in all cases considered, and the posterior median under the joint catalytic prior predicts better than the Cauchy prior except the cases when $r = 0.1$ (in these cases it still predicts comparably to the Cauchy prior). These three estimators all predict much better than the MLE.

Setting			Performance of Methods		
ζ	r		Cat.Boot	Cat.Joint	Cauchy
0.1	0.1	Cover	94.3	92.3	95.4
		Width	3.0	2.4	3.3
	0.2	Cover	94.0	97.7	99.3
		Width	2.0	2.1	3.0
	0.3	Cover	95.3	97.1	98.2
		Width	1.4	1.9	2.8
0.2	0.1	Cover	91.7	88.7	94.3
		Width	2.9	2.3	3.2
	0.2	Cover	93.3	97.4	99.3
		Width	2.0	2.1	3.0
	0.3	Cover	95.0	97.1	98.4
		Width	1.4	1.9	2.8
0.4	0.1	Cover	89.4	85.0	94.3
		Width	2.9	2.3	3.2
	0.2	Cover	92.7	97.3	99.3
		Width	2.0	2.1	3.0
	0.3	Cover	95.0	97.1	98.4
		Width	1.4	1.9	2.8

Table S2. (Logistic regression with interaction terms) Average coverage probability (%) and width of 95% posterior intervals for β under the catalytic prior with $\hat{\tau}_{boot}$, the joint catalytic prior, and the Cauchy prior. ζ is the non-sparsity. r is the oracle prediction error.

Table S2 presents the average coverage probabilities (in percentage) and widths of the 95% posterior intervals for β_j averaging over j. It is seen that that the coverage probabilities of the posterior intervals given by catalytic priors are close to the nominal probability in most cases except for $(\zeta, r) = (0.2, 0.1)$ and $(\zeta, r) = (0.4, 0.1)$. In contrast, the intervals given by Cauchy prior tend to have higher coverage probability but are much wider (about 1.5 times as wide). Note that because all the intervals associated with 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.

B. Model Misspecification. In this simulation, the settings, including the covariates distribution, regression coefficients and the sample size, are the same as those in the main text (i.e., without the interaction terms), but the response Y is actually generated from $\mathbf{P}(Y = 1 \mid \boldsymbol{X}^\top \boldsymbol{\beta}) = F_c(\boldsymbol{X}^\top \boldsymbol{\beta})$, where $F_c(\cdot)$ is the CDF of a standard Cauchy distribution. The working model, however, is still the logistic regression, so this is a case of model misspecification. Although the estimated parameters no longer have clear interpretations due to the model misspecification, we can still evaluate the prediction performance of various methods using the predictive binomial deviance as before.

Setting			Performance of Methods				
		Comp.		Cat.	Cat.	Cauchy	MLE
ζ	r	Sep.		Boot.	Joint		(pseudo)
1/4	0.1	100%	Mean	$\mathbf{0.988}$	1.073	1.097	1.401
			SE $\times 10^3$	(4.3)	(4.3)	(4.1)	(5.3)
	0.2	97%	Mean	$\mathbf{0.524}$	0.620	0.655	0.984
			SE $\times 10^3$	(3.0)	(2.7)	(2.6)	(3.8)
	0.3	90%	Mean	$\mathbf{0.272}$	0.375	0.422	0.715
			SE $\times 10^3$	(2.2)	(1.9)	(1.8)	(7.2)
2/4	0.1	100%	Mean	$\mathbf{0.963}$	1.049	1.062	1.377
			SE $\times 10^3$	(2.3)	(2.3)	(2.0)	(2.2)
	0.2	97%	Mean	$\mathbf{0.508}$	0.607	0.638	0.970
			SE $\times 10^3$	(2.3)	(2.0)	(1.8)	(3.7)
	0.3	91%	Mean	$\mathbf{0.265}$	0.370	0.416	0.717
			SE $\times 10^3$	(2.1)	(1.8)	(1.7)	(7.0)
3/4	0.1	100%	Mean	$\mathbf{0.965}$	1.051	1.060	1.377
			SE $\times 10^3$	(2.4)	(2.4)	(2.1)	(2.2)
	0.2	98%	Mean	$\mathbf{0.510}$	0.609	0.639	0.972
			SE $\times 10^3$	(2.3)	(2.0)	(1.8)	(3.4)
	0.3	89%	Mean	$\mathbf{0.262}$	0.367	0.412	0.706
			SE $\times 10^3$	(2.1)	(1.8)	(1.7)	(7.6)

Table S3. (Model misspecification) Mean and standard error of predictive binomial deviance of the catalytic posterior mode with $\hat{\tau}_{boot}$ and the posterior median under the joint catalytic prior, the Cauchy posterior mode, and the MLE. ζ is the non-sparsity. r is the oracle prediction error. The column of *Comp.Sep.* shows how often the complete separation occurs in the data sets. The boldface corresponds to the best performing method in each simulation scenario.

Table S3 shows that both catalytic prior specifications (the catalytic posterior mode with $\hat{\tau}_{boot}$ and the posterior median under the joint catalytic prior) predict better than Cauchy and MLE in all cases considered. The stable performance of the catalytic prior estimators under model misspecification illustrates the robustness of the catalytic priors.

C. Linear Regression. In this simulation, we consider the linear regression model. The covariates and the vector $\boldsymbol{\beta}^{(0)}$ are generated as in the *Illustration of Methods* Section of the main text. The only difference in specifying the regression coefficient $\boldsymbol{\beta}$ is that $\boldsymbol{\beta} = r \times \boldsymbol{\beta}^{(0)}$, where r is a scaling factor and is set at 1, 2 or 4. The sample size $n = 30$.

For the catalytic prior, we generate the synthetic data using the generating distributions described in Section 1 and fix M at 400. The first two estimators of $\boldsymbol{\beta}$ with the catalytic prior are the posterior modes of $\boldsymbol{\beta}$ with $\tau = \hat{\tau}_{cp}$ and $\tau = \hat{\tau}_{boot}$ selected from the estimated predictive risk via C_p and bootstrap respectively (denoted as $Cat.C_p$ and $Cat.Boot$). The third estimator of $\boldsymbol{\beta}$ with the catalytic prior is the coordinate-wise posterior median of $\boldsymbol{\beta}$ under the joint catalytic prior (denoted as $Cat.Joint$). These catalytic prior based estimators are compared to two alternative methods: the MLE and the Ridge estimator. The Ridge estimate is computed by

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \right),$$

where the penalty parameter λ is tuned by 10-fold cross-validation. Note that the Ridge estimator is identical to the posterior mode under an independent normal prior: $\beta_j \sim N(0, \lambda^{-1})$. We refer to this prior as the *Ridge prior* throughout this section. We use the predictive squared error, $\mathbb{E}_{\boldsymbol{X}_0}(\boldsymbol{X}_0^\top \boldsymbol{\beta} - \boldsymbol{X}_0^\top \hat{\boldsymbol{\beta}})^2$, to evaluate the performance of an estimator $\hat{\boldsymbol{\beta}}$, where the expectation is computed by sampling 1000 extra independent copies of $\boldsymbol{X}_0$ from the same distribution that generates the observed covariates.

Setting		Performance of Methods					
			Cat.	Cat.	Cat.	Ridge	MLE
ζ	r		Cp	Boot.	Joint		
0.25	1	Mean	0.228	0.228	**0.217**	0.236	0.282
		SE $\times 10^3$	(2.8)	(2.8)	(2.6)	(3.0)	(3.4)
	2	Mean	**0.264**	**0.264**	**0.264**	0.270	0.282
		SE $\times 10^3$	(3.2)	(3.2)	(3.2)	(3.4)	(3.4)
	4	Mean	**0.277**	**0.277**	**0.277**	0.281	0.281
		SE $\times 10^3$	(3.4)	(3.4)	(3.4)	(3.5)	(3.5)
0.5	1	Mean	0.246	0.246	**0.241**	0.253	0.279
		SE $\times 10^3$	(3.0)	(3.0)	(2.9)	(3.3)	(3.6)
	2	Mean	**0.270**	**0.270**	**0.270**	0.275	0.280
		SE $\times 10^3$	(3.1)	(3.1)	(3.1)	(3.2)	(3.3)
	4	Mean	**0.277**	**0.277**	0.278	0.281	0.280
		SE $\times 10^3$	(3.2)	(3.2)	(3.2)	(3.3)	(3.3)
0.75	1	Mean	0.256	0.256	**0.254**	0.260	0.277
		SE $\times 10^3$	(3.3)	(3.3)	(3.3)	(3.3)	(3.5)
	2	Mean	**0.272**	**0.272**	0.273	0.274	0.276
		SE $\times 10^3$	(3.6)	(3.6)	(3.6)	(3.6)	(3.6)
	4	Mean	**0.276**	**0.276**	0.277	0.279	0.277
		SE $\times 10^3$	(3.4)	(3.4)	(3.4)	(3.4)	(3.4)

Table S4. (Linear regression) Mean and standard error of the average predictive squared errors of the catalytic posterior mode estimates with $\hat{\tau}_{cp}$ and $\hat{\tau}_{boot}$, the posterior median estimate under joint catalytic prior, the Ridge estimate, and the MLE. ζ is the non-sparsity factor. r is the scaling factor. The boldface corresponds to the best performing method in each simulation scenario.

Setting		Performance of Methods					
ζ	r		Cat.Cp	Cat.Boot	Cat.Joint	Ridge	MLE
0.25	1	Cover	92.2	92.4	95.4	91.9	95.2
		Width	1.0	1.0	1.1	1.0	1.4
	2	Cover	94.0	93.7	94.1	93.2	95.2
		Width	1.2	1.2	1.2	1.2	1.4
	4	Cover	94.8	94.5	94.6	94.4	95.2
		Width	1.3	1.3	1.3	1.3	1.4
0.5	1	Cover	93.3	93.3	94.1	92.7	95.4
		Width	1.1	1.1	1.1	1.1	1.4
	2	Cover	94.5	94.4	94.3	93.9	95.4
		Width	1.3	1.3	1.3	1.3	1.4
	4	Cover	95.0	95.1	94.9	94.9	95.4
		Width	1.4	1.4	1.4	1.4	1.4
0.75	1	Cover	94.0	93.9	94.1	93.3	95.1
		Width	1.2	1.2	1.2	1.2	1.4
	2	Cover	94.7	94.8	94.8	94.5	95.2
		Width	1.3	1.3	1.3	1.3	1.4
	4	Cover	95.1	94.8	95.1	94.6	95.1
		Width	1.4	1.4	1.4	1.4	1.4

Table S5. (Linear regression) Average coverage probability (%) and width of 95% posterior intervals under the catalytic prior with $\hat{\tau}_{C_p}$, $\hat{\tau}_{boot}$, the joint catalytic prior, Ridge prior, and the confidence intervals associated with MLE. ζ is the non-sparsity. r is the oracle prediction error.

Table S4 compares the predictive squared error of different methods. We conclude that the estimates given by catalytic prior have generally better prediction performance to the Ridge estimate, and are generally much better than MLE.

Table S5 shows the average coverage probabilities (in percentage) and width of the 95% interval estimates for β_j averaging over j. The first 4 columns show the results for the posterior intervals under the catalytic priors and the Ridge prior, while the last column shows the confidence interval based on MLE. It is seen that the coverage probabilities given by all interval estimates considered here are reasonably close to the nominal in most cases, but when the coefficient vector is sparse ($\zeta = 0.25$) *and* weak ($r = 1$), the intervals given by both the Ridge estimator and the catalytic prior with τ selected by risk estimation tend to be narrow and do not achieve the nominal coverage. The intervals given by the joint catalytic prior and MLE have coverage probabilities close to the nominal level.

D. Other Error Measurements. In the *Illustration of Methods* section of the main text, we used the predictive binomial deviance to evaluate the performance of different estimators. The conclusion drawn there is in fact robust to the choice of error measurement: below we present the results in that experiment in terms of two alternative error measurements and draw a similar conclusion as stated in the main text.

1. **Expected classification error** is defined as $\mathbb{E}_{\boldsymbol{X}_0, Y_0, \hat{Y}_0}[\mathbf{1}_{Y_0 \neq \hat{Y}_0}]$, that is, the expected classification for a future data point $(\boldsymbol{X}_0, Y_0)$ using a probabilistic prediction $\hat{Y}_0$:

$$\hat{Y}_0 = \begin{cases} 1 & \text{w.p.} \quad \hat{\mu}_0 \\ 0 & \text{w.p.} \quad 1 - \hat{\mu}_0 \end{cases} \qquad [3.1]$$

   where $\hat{\mu}_0 = \mu(\boldsymbol{X}_0^\top \hat{\boldsymbol{\beta}})$ is the estimated probability for the future response being 1.

2. **Area Under Curve (AUC)** is often used in practice to evaluate the discrimination accuracy (7). It is defined as the area under the *Receiver Operating Characteristic* (ROC) curve, which is created by drawing the true positive rate against true negative rate for all possible cut-off points from 0 to 1. For a binary classifier, a higher AUC evaluated on the test data set means a better prediction.

Setting			Performance of Methods				
		Comp.		Cat.	Cat.	Cauchy	MLE
ζ	r	Sep.		Boot.	Joint		(pseudo)
1/4	0.1	100%	Mean	0.217	0.215	**0.208**	0.266
			SE $\times 10^3$	(1.2)	(1.2)	(1.1)	(1.8)
	0.2	98%	Mean	0.304	**0.301**	0.306	0.345
			SE $\times 10^3$	(1.1)	(1.1)	(1.1)	(2.1)
	0.3	91%	Mean	0.394	**0.393**	0.401	0.427
			SE $\times 10^3$	(1.0)	(1.0)	(1.0)	(2.2)
2/4	0.1	100%	Mean	0.216	**0.215**	0.221	0.266
			SE $\times 10^3$	(1.1)	(1.1)	(1.1)	(1.7)
	0.2	98%	Mean	0.302	**0.300**	0.310	0.347
			SE $\times 10^3$	(1.1)	(1.0)	(1.0)	(2.2)
	0.3	92%	Mean	0.394	**0.393**	0.403	0.427
			SE $\times 10^3$	(1.0)	(1.0)	(0.9)	(2.4)
3/4	0.1	100%	Mean	**0.215**	**0.215**	0.226	0.269
			SE $\times 10^3$	(1.1)	(1.1)	(1.1)	(1.7)
	0.2	99%	Mean	0.302	**0.299**	0.312	0.344
			SE $\times 10^3$	(1.1)	(1.0)	(1.0)	(2.0)
	0.3	91%	Mean	**0.392**	**0.392**	0.402	0.427
			SE $\times 10^3$	(1.0)	(1.0)	(0.9)	(2.4)

Table S6. (Logistic regression with main effect) Mean and standard error of average predictive classification error of the catalytic posterior mode with $\hat{\tau}_{boot}$, the posterior median under the joint catalytic prior, the Cauchy posterior mode and the MLE. ζ is the non-sparsity. r is the oracle prediction error. The column of *Comp.Sep.* shows how often complete separation occurs in the data sets. The boldface corresponds to the best performing method in each simulation scenario.

Setting				Performance of Methods			
		Comp.		Cat.	Cat.	Cauchy	MLE
ζ	r	Sep.		Boot.	Joint		(pseudo)
1/4	0.1	100%	Mean	0.875	0.876	**0.882**	0.809
			SE $\times 10^3$	(1.2)	(1.2)	(1.1)	(2.1)
	0.2	98%	Mean	0.775	**0.776**	0.767	0.710
			SE $\times 10^3$	(1.2)	(1.2)	(1.3)	(2.8)
	0.3	91%	Mean	**0.657**	0.654	0.641	0.602
			SE $\times 10^3$	(1.3)	(1.3)	(1.3)	(3.1)
2/4	0.1	100%	Mean	**0.877**	**0.877**	0.869	0.811
			SE $\times 10^3$	(1.1)	(1.1)	(1.1)	(2.0)
	0.2	98%	Mean	0.777	**0.778**	0.763	0.710
			SE $\times 10^3$	(1.2)	(1.2)	(1.2)	(2.9)
	0.3	92%	Mean	**0.658**	0.654	0.639	0.601
			SE $\times 10^3$	(1.3)	(1.3)	(1.3)	(3.4)
3/4	0.1	100%	Mean	**0.877**	**0.877**	0.864	0.807
			SE $\times 10^3$	(1.0)	(1.0)	(1.1)	(2.0)
	0.2	99%	Mean	0.777	**0.778**	0.760	0.713
			SE $\times 10^3$	(1.2)	(1.2)	(1.2)	(2.7)
	0.3	91%	Mean	**0.660**	0.656	0.640	0.601
			SE $\times 10^3$	(1.3)	(1.2)	(1.2)	(3.3)

Table S7. (Logistic regression with main effect) Mean and standard error of average predictive AUC. A higher predictive AUC means a better prediction. See the caption of Table S6. The boldface corresponds to the best performing method in each simulation scenario.

Tables S6 and S7 present the average predictive classification error and the average predictive AUC of the same experiment as in the *Illustration of Methods* section of the main text. Both measurements are evaluated on a independent test data set of size $1,000$ from the true data population. Based on these tables, both catalytic prior specifications predict better than MLE in all cases considered. Only in the case when the true $\boldsymbol{\beta}$ are very sparse ($\zeta = 0.25$) *and* have large amplitude ($r = 0.1$), the Cauchy prior works slightly better. In all other cases, the predictions given by catalytic priors are better than those of the Cauchy prior. Together with Table 1 of the main text, we conclude that catalytic priors provide much better prediction than MLEs and generally predicts better than or comparable to the Cauchy prior.

E. Synthetic-Data Generating Models with Different Input Dimensions. In this simulation experiment, we examine how the dimensionality/complexity of the synthetic-data generating model affects a catalytic prior, under the experimental setup in the *Illustration of Methods* Section of the main text with varying dimensions of the synthetic-data generating model.

To specify a synthetic-data generating model with dimension k, we use principal components of the covariate matrix to reduce the dimension of the covariate input. For $k > 1$, we can use logistic regression with the first $(k-1)$ principal components plus the intercept, and use maximum likelihood to generate the synthetic response vector $\boldsymbol{Y}^*$. The case with $k = 1$ only includes the intercept and is the same as the synthetic-data generating model of the main text. For $k = 0$, we use the fixed Bernoulli distribution with equal probability. To simplified the analysis, we discard the simulations where complete separation occurs in the first three principal components.

Setting			Dimension of the Simple Model				
ζ	r		0	1	2	3	4
1/4	0.1	Mean	1.676	1.684	1.712	1.736	1.767
		SE $\times 10^3$	(6.7)	(6.7)	(6.7)	(6.9)	(6.9)
	0.2	Mean	0.661	0.668	0.689	0.704	0.725
		SE $\times 10^3$	(5.1)	(5.1)	(5.1)	(5.2)	(5.4)
	0.3	Mean	0.280	0.289	0.302	0.312	0.327
		SE $\times 10^3$	(2.1)	(2.1)	(2.3)	(2.4)	(2.6)
2/4	0.1	Mean	1.655	1.662	1.689	1.713	1.740
		SE $\times 10^3$	(4.0)	(4.0)	(4.0)	(4.2)	(4.3)
	0.2	Mean	0.645	0.652	0.672	0.687	0.706
		SE $\times 10^3$	(2.5)	(2.5)	(2.7)	(2.9)	(3.1)
	0.3	Mean	0.281	0.289	0.302	0.312	0.325
		SE $\times 10^3$	(2.1)	(2.1)	(2.3)	(2.4)	(2.5)
3/4	0.1	Mean	1.664	1.672	1.701	1.720	1.749
		SE $\times 10^3$	(4.1)	(4.1)	(4.2)	(4.3)	(4.5)
	0.2	Mean	0.649	0.655	0.678	0.692	0.712
		SE $\times 10^3$	(2.5)	(2.5)	(2.7)	(2.8)	(3.0)
	0.3	Mean	0.282	0.289	0.303	0.312	0.326
		SE $\times 10^3$	(2.1)	(2.1)	(2.3)	(2.4)	(2.5)

Table S8. Mean and standard error of predictive binomial deviance of the catalytic posterior mode with $\hat{\tau}_{boot}$ using various synthetic-data generating models with different covariate input dimensions. ζ is the non-sparsity. r is the oracle prediction error.

Setting			Dimension of the Simple Model				
ζ	r		0	1	2	3	4
1/4	0.1	Cover	90.8%	90.7%	90%	88.8%	84.8%
		Width	3.4	3.5	3.8	4.3	5.2
	0.2	Cover	94.3%	93.3%	92.8%	91.3%	88.6%
		Width	2.7	2.7	2.9	3.0	3.3
	0.3	Cover	96.1%	95.1%	94.6%	93.7%	92%
		Width	2.1	2.1	2.2	2.2	2.3
2/4	0.1	Cover	89.4%	89.5%	88.7%	87.4%	84.5%
		Width	3.4	3.5	3.8	4.2	5.3
	0.2	Cover	94.2%	93.3%	92.7%	91.1%	88.5%
		Width	2.7	2.8	2.9	3.1	3.4
	0.3	Cover	96.2%	95.5%	95.1%	94.1%	92.3%
		Width	2.1	2.1	2.1	2.2	2.3
3/4	0.1	Cover	89.2%	89.4%	88.6%	87.1%	84%
		Width	3.4	3.5	3.9	4.3	5.3
	0.2	Cover	94.4%	94%	92.9%	91.4%	88.4%
		Width	2.7	2.8	2.9	3.1	3.3
	0.3	Cover	96.3%	95.6%	95.1%	94%	92.1%
		Width	2.1	2.1	2.2	2.2	2.3

Table S9. Average coverage probability (%) and width of 95% posterior intervals under the catalytic prior with $\hat{\tau}_{boot}$ using various synthetic-data generating models with different covariate input dimensions. ζ is the non-sparsity. r is the oracle prediction error.

Table S8 presents the average predictive binomial deviance, and Table S9 shows the average coverage probabilities (in percentage) and width of the 95% interval estimates for β_j averaging over j, both using the posterior mode under a catalytic prior whose τ is the minimizer of the parametric bootstrap risk estimate. As the dimension of the synthetic-data generating model increases, the prediction tends to get worse, and the interval estimates become wider and have lower coverage probabilities. We notice that such degeneration in performance disappears when the observed sample is relatively large, say 10 times larger than the dimension of the working model. This indicates that when the observed sample size is small, one should keep the synthetic-data generating model simple.

4. An Information Theory/Optimization Perspective

A. Interpretation from Information Theory. The Kullback-Leibler (KL) divergence is a measure of how far apart two distributions are. Suppose $g(y)$ and $f(y)$ are the densities of $\mathbf{P}_g$ and $\mathbf{P}_f$ over $\mathcal{Y}$ with respect to a base measure ν, the KL divergence between $\mathbf{P}_g$ and $\mathbf{P}_f$ is defined as

$$\mathrm{KL}\left(g(\cdot), f(\cdot)\right) = \mathrm{KL}\left(\mathbf{P}_g, \mathbf{P}_f\right) = \int_{\mathcal{Y}} g(y) \log\left(\frac{g(y)}{f(y)}\right) d\nu(y).$$

It is straightforward that

$$\mathrm{KL}\left(g(\cdot), f(\cdot)\right) = \mathbb{E}_{Y\sim\mathbf{P}_g} \log\left(\frac{g(Y)}{f(Y)}\right). \qquad [4.1]$$

The population catalytic prior can be mathematically rewritten as exponentiating the KL divergence between the working model and the synthetic-data generating distribution, which is

$$\mathrm{KL}\left(g_*(\cdot \mid \boldsymbol{Y}),\ f(\cdot \mid \theta)\right) = \mathbb{E}_{Y^*}\left[\log \frac{g_*(Y^* \mid \boldsymbol{Y})}{f(Y^* \mid \theta)}\right],$$

as (ignoring the terms that do no involve θ) one can rewrite Eq. (4) of the main text as

$$\pi_{cat,\infty}(\theta \mid \tau) \propto \exp\left(-\tau\ \mathrm{KL}\left(g_*(\cdot \mid \boldsymbol{Y}),\ f(\cdot \mid \theta)\right)\right).$$

This formulation is mathematically similar to that of the *PC prior* (8), but the key differences are (i) the catalytic prior is primarily motivated from the synthetic-data perspective, (ii) the PC prior requires the simpler model to be nested in the working model, whereas the catalytic prior has no such restrictions, and (iii) mathematically, the KL divergence is not symmetric, and the PC prior is penalizing $\mathrm{KL}\left(f(\cdot \mid \theta),\ g_*(\cdot \mid \boldsymbol{Y})\right)$; this leads to a different construction.

In the presence of covariates, we can also mathematically formulate the population catalytic prior in terms of the *expected KL* divergence:

$$\mathbb{E}_{\boldsymbol{X}^*}\left[\mathrm{KL}\left(g_*(\cdot \mid \boldsymbol{X}^*, \boldsymbol{Y}, \mathbb{X}),\ f(\cdot \mid \boldsymbol{X}^*, \boldsymbol{\beta})\right)\right] = \mathbb{E}_{\boldsymbol{X}^*, Y^*}\left[\log \frac{g_*(Y^* \mid \boldsymbol{X}^*, \boldsymbol{Y}, \mathbb{X})}{f(Y^* \mid \boldsymbol{X}^*, \boldsymbol{\beta})}\right],$$

where the expectation is averaging over both the synthetic response and the synthetic-covariates. Eq. (7) of the main text can be written as

$$\pi_{cat,\infty}(\boldsymbol{\beta} \mid \tau) \propto \exp\left\{-\tau\ \mathbb{E}_{\boldsymbol{X}^*}\left[\mathrm{KL}\left(g_*(\cdot \mid \boldsymbol{X}^*, \boldsymbol{Y}, \mathbb{X}),\ f(\cdot \mid \boldsymbol{X}^*, \boldsymbol{\beta})\right)\right]\right\}.$$

This formulation suggests the resultant posterior mode is the solution of the following optimization

$$\min_{\boldsymbol{\beta}} \left\{ -\log f(\boldsymbol{Y} \mid \mathbb{X}, \boldsymbol{\beta}) + \tau\ \mathbb{E}_{\boldsymbol{X}^*}\left[\mathrm{KL}\left(g_*(\cdot \mid \boldsymbol{X}^*, \boldsymbol{Y}, \mathbb{X}),\ f(\cdot \mid \boldsymbol{X}^*, \boldsymbol{\beta})\right)\right] \right\}. \qquad [4.2]$$

We now further show that the posterior mode under a population catalytic prior can be viewed as the maximum of the likelihood function with the constraint that the expected KL divergence between the corresponding distribution and the synthetic-data generating distribution is bounded by a budget.

Consider the following constrained optimization problem

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & -\log f(\boldsymbol{Y} \mid \mathbb{X}, \boldsymbol{\beta}) \qquad [4.3] \\ s.t. \quad & \mathbb{E}_{\boldsymbol{X}^*}\left[\mathrm{KL}\left(g_*(\cdot \mid \boldsymbol{X}^*, \boldsymbol{Y}, \mathbb{X}),\ f(\cdot \mid \boldsymbol{X}^*, \boldsymbol{\beta})\right)\right] \le C. \end{aligned}$$

The Lagrange associated with this problem is

$$\min_{\boldsymbol{\beta}} \max_{\tau>0} \left\{ -\log f(\boldsymbol{Y} \mid \mathbb{X}, \boldsymbol{\beta}) + \tau\ \mathbb{E}_{\boldsymbol{X}^*}\left[\mathrm{KL}\left(g_*(\cdot \mid \boldsymbol{X}^*, \boldsymbol{Y}, \mathbb{X}),\ f(\cdot \mid \boldsymbol{X}^*, \boldsymbol{\beta})\right)\right] - \tau\ C \right\}.$$

If both the objective function and the constraint function are convex and the original optimization problem is strictly feasible, namely, there exists some $\boldsymbol{\beta}^\sharp$ such that

$$\mathbb{E}_{\boldsymbol{X}^*}\left[\mathrm{KL}\left(g_*(\cdot \mid \boldsymbol{X}^*, \boldsymbol{Y}, \mathbb{X}),\ f(\cdot \mid \boldsymbol{X}^*, \boldsymbol{\beta}^\sharp)\right)\right] < C, \qquad [4.4]$$

then the optimization is equivalent to its dual problem: there exists some $\tau_C > 0$ depending on C such that the minimum of the original problem is the minimum of the following problem

$$\min_{\boldsymbol{\beta}} \{ -\log f(\boldsymbol{Y} \mid \mathbb{X}, \boldsymbol{\beta}) + \tau_C\ \mathbb{E}_{\boldsymbol{X}^*}\left[\mathrm{KL}\left(g_*(\cdot \mid \boldsymbol{X}^*, \boldsymbol{Y}, \mathbb{X}),\ f(\cdot \mid \boldsymbol{X}^*, \boldsymbol{\beta})\right)\right] - \tau_C\ C \},$$

where we have used the KL divergence formula Eq. (4.1). Note that the convexity of the objective function and the constraint function is guaranteed when the working model is a GLM with the canonical link.

Ignoring the terms that do no involve $\boldsymbol{\beta}$, the last optimization is equivalent to Eq. (4.2). Therefore, we have shown the equivalence of the posterior mode under the population catalytic prior and the optimization problem in Eq. (4.3).

B. Effect of the synthetic covariate Generating Distribution and the Connection to L_1 Estimators. The equivalent form of the posterior mode under a catalytic prior in Eq. (4.3) suggests us to study how the synthetic covariate generating distribution $Q(\boldsymbol{x}^*)$ affects the posterior mode, which would provide additional insight on choosing the synthetic-covariate generating distribution. This section considers finding the optimal design for sampling $\boldsymbol{X}^*$ so that the posterior mode has desirable frequentist properties.

Note that only the constraint on the expected KL divergence in Eq. (4.3) relies on $Q(\boldsymbol{x}^*)$. To understand how $Q(\boldsymbol{x}^*)$ affects the constraint, we first define the projected parameter of the synthetic-data generating distribution as the $\boldsymbol{\beta}$ in the predictive model that minimizes the expected KL divergence

$$\tilde{\boldsymbol{\beta}}_0 = \operatorname*{argmin}_{\boldsymbol{\beta}} \mathbb{E}_{\boldsymbol{X}^*}\left\{\mathrm{KL}\left[g_*(\cdot \mid \boldsymbol{X}^*, \boldsymbol{Y}, \mathbb{X}),\ f(\cdot \mid \boldsymbol{X}^*, \boldsymbol{\beta})\right]\right\}. \qquad [4.5]$$

For the selection of $Q(\boldsymbol{x}^*)$, let us consider the Fisher information, which is a measure of the amount of information contained in the data about the parameters in a model, and a choice of $Q(\boldsymbol{x}^*)$ is preferable if it maximizes the Fisher information matrix at the true regression coefficient $\boldsymbol{\beta}^\dagger$ (or the best projected parameter in the case of model misspecification). Since $Q(\boldsymbol{x}^*)$ should be selected at the stage of specifying a prior, we instead consider maximizing the "prior" Fisher information matrix that depends only on the synthetic dataset.

One precise and convenient way to capture this idea is to minimize the trace of inverse "prior" Fisher information matrix. This leads to the following optimization

$$\min_{Q} \mathrm{Tr}\left\{ \left(\tau \mathbb{E}_{\boldsymbol{X}^*\sim Q}\mathbb{E}_{\boldsymbol{Y}^*}\left[-\nabla^2 \log f(Y^* | \boldsymbol{X}^*, \boldsymbol{\beta}^\dagger)\right]\right)^{-1} \right\}, \qquad [4.6]$$

which is known as *A-optimality* in the optimal design literature. To integrate such an optimization and the optimization in Eq. (4.3), we consider the minimization that combines the likelihood function, the constraint on the expected KL-divergence from the simplified model and the trace of the inverse prior Fisher information matrix simultaneously as follows

$$\min_{\boldsymbol{\beta}, Q}\left\{ -\log f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}) + \tau \mathbb{E}_{\boldsymbol{X}^*\sim Q}\left(\mathbb{E}_{\boldsymbol{Y}^*}(-\log f(Y^*|\boldsymbol{X}^*, \boldsymbol{\beta})\right) + \lambda \mathrm{Tr}\left\{\left(\tau \mathbb{E}_{\boldsymbol{X}^*\sim Q}\mathbb{E}_{\boldsymbol{Y}^*} - \nabla^2 \log f(Y^*|\boldsymbol{X}^*, \boldsymbol{\beta}^\dagger)\right)^{-1}\right\} \right\}. \qquad [4.7]$$

Since $\boldsymbol{\beta}^\dagger$ is unknown, such a optimization is unsolvable in general. However, in the special case of linear regression with known noise level, the Hessian of the log likelihood does not depend on the parameter and the optimization becomes possible, as we will discuss soon in Example 1. In addition, optimizing over all possible sampling schemes can be too ambitious, so in Example 1 we restrict ourselves to only the sampling distributions that affinely transform data from a fixed distribution $Q_0(\boldsymbol{x}^*)$. By simplifying the problem, we obtain an interesting connection between catalytic prior and other estimators in the regression literature.

Example 1. Let us consider the linear regression in Section 1. When the simplified model is a sub-model, say, $Y^* \mid \boldsymbol{X}^* \overset{g_*}{\sim} N(\tilde{\boldsymbol{\beta}}_0^\top \boldsymbol{X}^*, \sigma^2)$, we have

$$\begin{aligned}\mathbb{E}_{Y^*\sim g_*} \log\left(\frac{g_*(Y^* \mid \boldsymbol{X}^*, \boldsymbol{Y}, \mathbb{X})}{f(Y^* \mid \boldsymbol{X}^*, \boldsymbol{\beta})}\right) &= \mathbb{E}_{Y^*\sim g_*}\left(-\frac{1}{2\sigma^2}(Y^* - \tilde{\boldsymbol{\beta}}_0^\top \boldsymbol{X}^*)^2 + \frac{1}{2\sigma^2}(Y^* - \boldsymbol{\beta}^\top \boldsymbol{X}^*)^2\right) \\ &= \frac{1}{2\sigma^2}[(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_0)^\top \boldsymbol{X}^*]^2,\end{aligned}$$

which implies that the expected KL divergence is

$$\mathbb{E}_{\boldsymbol{X}^*}\left\{\frac{1}{2\sigma^2}[(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_0)^\top \boldsymbol{X}^*]^2\right\} = \frac{1}{2\sigma^2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_0)^\top \Sigma_Q (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_0), \qquad [4.8]$$

where $\Sigma_Q = \mathbb{E}_{\boldsymbol{X}^*}[\boldsymbol{X}^*(\boldsymbol{X}^*)^\top]$ is the covariance matrix under $Q(\boldsymbol{x}^*)$. From Eq. (4.8), it is clear that a large variability in $\boldsymbol{X}^*$ (mathematically, it means the eigenvalues Σ_Q are large) leads to a restrictive constraint on the KL divergence.

Now consider the following class of sampling distribution

$$\boldsymbol{X}^* = \boldsymbol{A}\boldsymbol{X}_0^*, \quad \boldsymbol{X}_0^* \sim Q_0(\cdot) \qquad [4.9]$$

where $Q_0(\boldsymbol{x}^*)$ is the the independent resampling distribution and A belongs to the set of all $p \times p$ non-singular matrices. Some elementary calculation deduces Eq. (4.7) to

$$\min_{\boldsymbol{\beta}, A}\left\{\frac{1}{2}\cdot\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\tau}{2}\cdot(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_0)^\top(A^\top \boldsymbol{D}_X A)(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_0) + \frac{\lambda\sigma^2}{\tau}\mathrm{Tr}(A^\top \boldsymbol{D}_X A)^{-1}\right\}, \qquad [4.10]$$

where $\boldsymbol{D}_X = \mathrm{diag}(1, \widehat{\sigma^2_{X,1}}, \ldots, \widehat{\sigma^2_{X,p-1}})$.

- If we further assume A to be positively diagonal, i.e., $A = \mathrm{diag}(1, \sqrt{s_1}, \ldots, \sqrt{s_{p-1}})$, we are equivalently scaling the synthetic covariate components separately. Then the optimization reduces to

  $$\min_{\boldsymbol{\beta}, s_{1:(p-1)}}\left\{\frac{1}{2}\cdot\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\tau}{2}\cdot(\beta_0 - \tilde{\beta}_{0,0})^2 + \frac{\tau}{2}\cdot\sum_{j=1}^{p-1} s_j\widehat{\sigma^2_{X,j}}(\beta_j - \tilde{\beta}_{0,j})^2 + \frac{\lambda\sigma^2}{\tau}\sum_{j=1}^{p-1}\frac{1}{s_j\widehat{\sigma^2_{X,j}}}\right\}. \qquad [4.11]$$

  Minimizing w.r.t. s_j's is straightforward and the minimizer is

  $$s_j = \begin{cases} \frac{\sqrt{2\lambda\sigma^2}}{\widehat{\sigma^2_{X,j}}\tau|(\beta_j - \tilde{\beta}_{0,j})|}, & \text{if} \quad (\beta_j - \tilde{\beta}_{0,j}) \neq 0 \\ \infty & \text{if} \quad (\beta_j - \tilde{\beta}_{0,j}) = 0 \end{cases}$$

  Thus the optimization further reduces to

  $$\min_{\boldsymbol{\beta}}\left\{\frac{1}{2}\cdot\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\tau}{2}\cdot(\beta_0 - \tilde{\beta}_{0,0})^2 + \sqrt{2\lambda\sigma^2}\cdot\sum_{j=1}^{p-1}|\beta_j - \tilde{\beta}_{0,j}|\right\}. \qquad [4.12]$$

  If the simplified model is the intercept-only model, then $\tilde{\beta}_{0,j} = 0$ and $\tilde{\beta}_{0,0} = \bar{Y}$, and the optimization is the same as the formulation for the LASSO estimator (9).

- More generally, assume A to be a non-singular linear transformation. Suppose the spectral decomposition of $A^\top \boldsymbol{D}_X A$ is $U^\top \mathrm{diag}(1, s_1, \ldots, s_{p-1})U$ where U is an orthonormal matrix with its first column U_1 parallel to $\mathbf{1}$. Optimizing Eq. (4.7) w.r.t. A reduces the problem to

  $$\min_{\boldsymbol{\beta}}\left\{\frac{1}{2}\cdot\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\tau}{2}\cdot(\beta_0 - \tilde{\boldsymbol{\beta}}_{0,0})^2 + \sqrt{2\lambda\sigma^2}\|\boldsymbol{\beta}_{-0}\|\right\}. \qquad [4.13]$$

  Note that the penalty on $\boldsymbol{\beta}_{-0}$ is in the ℓ_2 norm rather than its square, and is thus different from the classic Ridge regression. The same penalty is also used to define an estimator in Ref. (10) (see Eq (33) therein).

□

For a general model, the Hessian of the log density may change with the parameter, so optimizing Eq. (4.7) may be infeasible. One possible solution is to replace $\boldsymbol{\beta}^\dagger$ by an estimate $\hat{\boldsymbol{\beta}}^0$ and iteratively update $\hat{\boldsymbol{\beta}}$ and $Q(\boldsymbol{x}^*)$. We left this idea for future investigation.

5. Theoretical Properties

We begin with some notations. In Section A and Section B, we use the generic u for a value of the response variable, $\boldsymbol{W}$ for a covariate random vector, $\mathbb{W}$ for a covariate random matrix, and m for the number of data points. In Section C and Section D, we denoted by X^* / $\boldsymbol{X}^*$ / $\mathbb{X}^*$ a synthetic covariate variable/vector/matrix, Y^* / $\boldsymbol{y}^*$ a synthetic response variable/vector, and M the synthetic-sample size. Throughout, p is the number of parameters including both the covariates and the intercept term.

We adapt some notations and terminologies for exponential family from the classical reference (11). Let ν be a fixed σ-finite measure on Borel sets of $\mathbb{R}$, and $\mathcal{Y}$ be the interior of the convex hull of the support set of ν. Assume $\mathcal{Y}$ is nonempty and open. For any $\theta \in \mathbb{R}$, define $b(\theta) = \log \int e^{y\theta} d\nu(y)$, and $\Theta = \{\theta : b(\theta) < \infty\}$. Assume Θ is nonempty and open. It can be shown that Θ is convex and $b'(\theta) \in \mathcal{Y}$, for any $\theta \in \Theta$. The exponential family is defined by

$$d\mathbf{P}_\theta(y) = e^{y\theta - b(\theta)} d\nu(y). \tag{5.1}$$

Without loss of generality, we assume the sample mean and the sample variance of each observed covariate are 0 and 1, i.e., the observed covariates are standardized. Please note that without loss of generality, we assume the sufficient statistic $t(y) = y$ in the GLM formula $d\mathbf{P}_\theta(y) = e^{\theta t(y) - b(\theta)} d\nu(y)$ throughout this section; otherwise we can redefine the response as $Y' = t(Y)$ and proceed.

The structure of this theoretical section is as follows. Section A quantifies the tail integral for catalytic priors under certain conditions, and Section B shows these conditions are satisfied if the synthetic covariates are drawn from the the independent resampling distribution. The results in these two sections will be used in Section C to establish bounds on the integrals of the catalytic priors. These bounds will be used to show the properness of catalytic priors. Section D quantifies two types of divergence between the catalytic prior and the population catalytic prior.

A. Upper Bounds on the Integrals of Tails. We will derive upper bounds on the integrals of the tails (i.e., for $\|\boldsymbol{\beta}\| > K$) of the unnormalized density function for both the catalytic prior and the population catalytic prior, provided a condition called *norm-recovery* holds.

We begin with an elementary lemma that bounds the integral of a multivariate tail for $\exp(-a\|\boldsymbol{x}\|)$.

Lemma 5.1. *For $K > 0, a > 0$,*

$$\int_{\|\boldsymbol{x}\|>K} exp(-a\|\boldsymbol{x}\|)d\boldsymbol{x} \leq \frac{\Gamma(p)s_{p-1}}{a^p} \min\left(1, e^{p-aK}(\frac{aK}{p})^p\right),$$

where the constant $s_{p-1} = \frac{2\pi^{p/2}}{\Gamma(p/2)}$ is the surface area of a $(p-1)$-dimension sphere. Furthermore, $\frac{\Gamma(p)s_{p-1}}{a^p} \leq C_{Stirling}\frac{(2\pi p/e)^{p/2}}{a^p}$, where $C_{Stirling}$ is a universal constant. Thus we also have

$$\int_{\|\boldsymbol{x}\|>K} exp(-a\|\boldsymbol{x}\|)d\boldsymbol{x} \leq C_{Stirling} \min\left(\frac{(2\pi p/e)^{p/2}}{a^p}, e^{p-aK}(\frac{\sqrt{2\pi}K}{\sqrt{pe}})^p\right). \tag{5.2}$$

Proof. Using the p-dimensional spherical coordinates, the integral is equal to

$$\int_{\|\boldsymbol{x}\|>K} \exp(-a\|\boldsymbol{x}\|)d\boldsymbol{x} = \int_{r>K} r^{p-1}s_{p-1}\exp(-ar)dr = \frac{\Gamma(p)s_{p-1}}{a^p}\mathbf{P}(G > aK),$$

where $G \sim Gamma(p)$. By the Chernoff inequality and the moment generating function of G, for any $t > 0$, we have

$$\mathbf{P}(G > aK) \leq \exp(-taK)\mathbb{E}e^{tG} = \exp(-taK)/(1-t)^p.$$

Minimizing the right-hand side of the last inequality with respect to $t > 0$, we obtain

$$\mathbf{P}(G > aK) \leq \min(1, \exp(p - aK + p\log(aK/p))),$$

which is the first part of the lemma. By Stirling's formula, $\Gamma(z) = \sqrt{\frac{2\pi}{z}}(\frac{z}{e})^z(1+O(\frac{1}{z}))$, and

$$\frac{\Gamma(p)}{\Gamma(p/2)} = \frac{\sqrt{\frac{2\pi}{p}}(\frac{p}{e})^p(1+O(\frac{1}{p}))}{\sqrt{\frac{2\pi}{p/2}}(\frac{p/2}{e})^{p/2}(1+O(\frac{1}{p}))} = \frac{2^{p/2-1/2}p^{p/2}}{e^{p/2}}\frac{1+O(1/p)}{1+O(1/p)} \leq C_{Stirling}\frac{2^{p/2-1}p^{p/2}}{e^{p/2}},$$

where $C_{Stirling}$ is a universal constant. Hence we conclude the second part. The third part follows by combining the first and second parts. □

The following two theorems bound the tail integrals for catalytic priors under two conditions. These results will be used later in Section C.

The first condition says that all responses are at least δ away from the boundary of the sample space. The second condition is referred to as *norm-recoverability* and will be studied in Section B.

Theorem 5.2. *Let $u_1, \cdots, u_m \in \mathcal{Y}, w_1, \cdots, w_m \in \mathbb{R}^p$. Suppose*

(1) there exist $u_-, u_+ \in \mathcal{Y}, \delta > 0$, such that $u_- \leq u_i - \delta < u_i + \delta \leq u_+$, for $i = 1, \cdots, m$;

(2) there is a positive constant c_1 such that $\frac{1}{m}\sum_{i=1}^m |\phi(w_i^\top \boldsymbol{\beta})| \geq c_1\|\boldsymbol{\beta}\|$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$.

Then there exists a constant C that only depends on u_-, u_+ and the exponential family Eq. (5.1)*, such that*

(a) $\sup_{\boldsymbol{\beta}\in\mathbb{R}^p}\max_{1\leq i\leq m}\left(u_i\phi(w_i^\top\boldsymbol{\beta}) - b(\phi(w_i^\top\boldsymbol{\beta}))\right) \leq \log C$;

(b) there exists a universal constant $C_{Stirling}$, such that for any $K > p/(\alpha\delta c_1)$

$$\int_{\|\boldsymbol{\beta}\|>K} exp\left(\frac{\alpha}{m}\sum_{i=1}^m(u_i\phi(w_i^\top\boldsymbol{\beta}) - b(\phi(w_i^\top\boldsymbol{\beta}))\right)d\boldsymbol{\beta} \leq C_{Stirling}C^\alpha exp(p - c_1\alpha\delta K)(\frac{\sqrt{2\pi}K}{\sqrt{pe}})^p;$$

(c)

$$\int_{\|\boldsymbol{\beta}\|\in\mathbb{R}^p} exp\left(\frac{\alpha}{m}\sum_{i=1}^m(u_i\phi(w_i^\top\boldsymbol{\beta}) - b(\phi(w_i^\top\boldsymbol{\beta}))\right)d\boldsymbol{\beta} \leq C_{Stirling}C^\alpha\frac{(2\pi p)^{p/2}}{(c_1\alpha\delta)^p}.$$

Proof. By Condition (1) and Eq.(2.4) in Ref. (11), there are two compact convex subsets A_- and A_+ of $\mathbb{R}$ and $u_{A_+}, u_{A_-} \in \mathcal{Y}$ such that

$$u_{A_-} \leq u_i - \delta, \quad u_{A_+} \geq u_i + \delta, \tag{5.3}$$

and

$$e^{-b(\theta)} \leq (\mu(A))^{-1}e^{-\theta u_A}, A \in \{A_-, A_+\}. \tag{5.4}$$

Let $C = \max(\mu(A_-)^{-1}, \mu(A_+)^{-1})$. For a given $\boldsymbol{\beta}$, denote $\theta_i = \phi(w_i^\top\boldsymbol{\beta})$ and let S_i be the sign of θ_i, either $+$ or $-$. By Eq. (5.3), it holds that $(u_i\theta_i - u_{A_{S_i}}\theta_i) \leq -\delta|\theta_i|$ regardless of S_i. Together with Eq. (5.4), we have $u_i\theta_i - b(\theta_i) \leq \log C - \delta|\theta_i|$, which yields the result of part (a). Now the integral can be bounded as

$$\begin{aligned}&\int_{\|\boldsymbol{\beta}\|>K}\exp\left(\frac{\alpha}{m}\sum_{i=1}^m(u_i\theta_i - b(\theta_i)\right)d\boldsymbol{\beta}\\ \leq &\int_{\|\boldsymbol{\beta}\|>K}\exp\left(\frac{\alpha}{m}\sum_{i=1}^m(\log C - \delta|\theta_i|)\right)d\boldsymbol{\beta} \leq C^\alpha\int_{\|\boldsymbol{\beta}\|>K}\exp\left(-c_1\alpha\delta\|\boldsymbol{\beta}\|\right)d\boldsymbol{\beta},\end{aligned} \tag{5.5}$$

where the second inequality uses Condition (2). By Lemma 5.1, the last inequality can be bounded from above by $C_{Stirling}C^\alpha\exp(p - c_1\alpha\delta K)(\frac{\sqrt{2\pi}K}{\sqrt{pe}})^p$, which is part (b). Note that Eq. (5.5) actually also holds for $K = 0$ and can be bounded from above by $C_{Stirling}C^\alpha\frac{(2\pi p/e)^{p/2}}{(c_1\alpha\delta)^p}$ using Lemma 5.1, which gives part (c). □

Theorem 5.3. *Suppose* $(\boldsymbol{W}, U)$ *are jointly random, and*

(1) there exist $u_-, u_+ \in \mathcal{Y}, \delta > 0$, *such that* $u_- \leq \mathbb{E}(U \mid \boldsymbol{W}) - \delta < \mathbb{E}(U \mid \boldsymbol{W}) + \delta \leq u_+$*;*

(2) there is a positive constant c_0 *such that* $\mathbb{E}|\phi(\boldsymbol{W}^\top \boldsymbol{\beta})| \geq c_0 \|\boldsymbol{\beta}\|$, *for all* $\boldsymbol{\beta} \in \mathbb{R}^p$*;*

then there exists some constant C *that only depends on* u_-, u_+ *and the exponential family, such that*

(a) $\mathbb{E}[U\phi(\boldsymbol{W}^\top \boldsymbol{\beta}) - b(\phi(\boldsymbol{W}^\top \boldsymbol{\beta}))] \leq \log C$

(b) There exists a universal constant $C_{Stirling}$, *such that for any* $K > p/(\alpha\delta c_0)$, *we have*

$$\int_{\|\boldsymbol{\beta}\|>K} exp\left\{\alpha\mathbb{E}[U\phi(\boldsymbol{W}^\top \boldsymbol{\beta}) - b(\phi(\boldsymbol{W}^\top \boldsymbol{\beta}))]\right\} d\boldsymbol{\beta} \leq C_{Stirling} C^\alpha exp(p - c_0\alpha\delta K)(\frac{\sqrt{2\pi}K}{\sqrt{pe}})^p;$$

(c)

$$\int_{\|\boldsymbol{\beta}\|\in\mathbb{R}^p} exp\left\{\alpha\mathbb{E}[U\phi(\boldsymbol{W}^\top \boldsymbol{\beta}) - b(\phi(\boldsymbol{W}^\top \boldsymbol{\beta}))]\right\} d\boldsymbol{\beta} \leq C_{Stirling} C^\alpha \frac{(2\pi p)^{p/2}}{(c_0\alpha\delta)^p}.$$

Proof. The proof follows the same argument as that of Theorem 5.2. □

B. Synthetic-Covariate Generation with Norm-Recoverability. This section focuses on the synthetic-covariate generating distribution. Both Theorem 5.2 and Theorem 5.3 can be used to bound the tail integrals as long as we have the following two lower bounds

$$\forall \boldsymbol{\beta} \in \mathbb{R}^p, \quad \mathbb{E}|\phi(\boldsymbol{W}^\top \boldsymbol{\beta})| \geq c_0 \|\boldsymbol{\beta}\|, \quad \frac{1}{m}\sum_{i=1}^{m} |\phi(\boldsymbol{W}_i^\top \boldsymbol{\beta})| \geq c_1 \|\boldsymbol{\beta}\|. \tag*{[5.6]}$$

Note that the first inequality is deterministic while the second one is stochastic.

Definition 5.4. If Eq. (5.6) holds with high probability for a synthetic-covariate generating distribution, then we call this distribution *norm-recoverable.*

We will focus on the case where the θ-link function is the identity, that is $\phi(\eta) = \eta$, because, with the condition in the main text that

$$\inf_{\eta\neq 0} |\phi(\eta)/\eta| > 0,$$

if Eq. (5.6) holds for the identity link, then it holds for $\phi(\cdot)$.

A sufficient condition for norm-recoverability is given below.

Condition 5.5. *The random vector* $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)$ *satisfies (1)* $X_1 \equiv 1$*; (2)* $X_2, \cdots, X_p$ *are independent; (3)* $\mathbb{E}X_j = 0, Var(X_j) = 1$ *for* $j = 2, \cdots, p$*; (4)* $|X_j| \leq B_1, a.s.$ *for* $j = 2, \cdots, p$.

Remark 5.6. $X_1 \equiv 1$ corresponds to the intercept (constant) term in a GLM. □

Theorem 5.7. *If* $\boldsymbol{X}$ *satisfies Condition 5.5, then there exist positive constants* ρ_0, η_0, c *and* C *that only depend on* B_1 *such that*

(a) $\mathbf{P}\left(|\boldsymbol{X}^\top \boldsymbol{\beta}| > \eta_0\right) \geq \rho_0$ *for any* $\boldsymbol{\beta}$ *with* $\|\boldsymbol{\beta}\| = 1$

(b) $\mathbb{E}(|\boldsymbol{X}^\top \boldsymbol{\beta}|) \geq \eta_0\rho_0$, *for any* $\boldsymbol{\beta}$ *with* $\|\boldsymbol{\beta}\| = 1$

(c) *if* $\{\boldsymbol{X}_i\}_{i=1}^M$ *are i.i.d. copies of* $\boldsymbol{X}$, *then with probability at least*

$$1 - e^{-cM} - exp\left(-\frac{M\rho_0^2}{2} + p\log(1 + \frac{8C}{\eta_0\rho_0})\right),$$

$\mathbb{X}$ *has full column rank and*

$$\inf_{\|\boldsymbol{\beta}\|=1} \frac{1}{M}\sum_{i=1}^{M} |\boldsymbol{X}_i^\top \boldsymbol{\beta}| \geq \frac{\eta_0\rho_0}{4}.$$

We will establish several results that serve as the basis for Theorem 5.7, and the proof of Theorem 5.7 will be deferred to the end of this section. The first result shows that *the small ball probability*, $\mathbf{P}(|\boldsymbol{X}^\top\boldsymbol{\beta}| > \eta)$, being bounded away from 0 is a sufficient condition for $\boldsymbol{X}$ to be norm-recoverable.

Lemma 5.8. *Suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ are i.i.d. copies of $\boldsymbol{X}$ and $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m)^\top$. If there are positive constants η and ρ such that*

$$\mathbf{P}(|\boldsymbol{X}^\top\boldsymbol{\beta}| > \eta) \geq \rho, \quad \forall \|\boldsymbol{\beta}\| = 1, \tag{5.7}$$

then

- ***(a)*** $\mathbb{E}(|\boldsymbol{X}^\top\boldsymbol{\beta}|) \geq \eta\rho > 0$ *for any $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\| = 1$*
- ***(b)*** $\mathbf{P}(\frac{1}{m}\sum_{i=1}^m |\boldsymbol{X}_i^\top\boldsymbol{\beta}| \geq \frac{1}{2}\eta\rho) \geq 1 - exp\left(-\frac{m\rho^2}{2}\right)$ *for any $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\| = 1$*
- ***(c)*** *Let $\|\mathbb{X}\|$ denote the operator norm of the matrix $\mathbb{X}$. If $\mathbf{P}(\|\mathbb{X}\| > C\sqrt{m}) \leq e^{-cm}$ for some constants c and C, then with probability at least*

  $$1 - e^{-cm} - exp\left(-\frac{m\rho^2}{2} + p\log(1 + \frac{8C}{\rho})\right),$$

  it holds that $\mathbb{X}$ has full column rank and

  $$\inf_{\|\boldsymbol{\beta}\|=1} \frac{1}{m}\sum_{i=1}^m |\boldsymbol{X}_i^\top\boldsymbol{\beta}| \geq \frac{\eta\rho}{4}.$$

Proof. (a) is trivial. For (b), let $\xi_i = \mathbf{1}_{|\boldsymbol{X}_i^\top\boldsymbol{\beta}|>\eta}$. By Hoeffding's inequality, $\mathbf{P}(\sum_{i=1}^m \xi_i \leq \frac{m\rho}{2}) \leq \exp\left(-\frac{m\rho^2}{2}\right)$. Note that the event $\{\sum_{i=1}^m \xi_i > \frac{m\rho}{2}\}$ implies that $\sum_{i=1}^m |\boldsymbol{X}_i^\top\boldsymbol{\beta}| > \frac{m\rho}{2}\eta$. Thus,

$$\mathbf{P}(\frac{1}{m}\sum_{i=1}^m |\boldsymbol{X}_i^\top\boldsymbol{\beta}| \leq \frac{1}{2}\eta\rho) \leq \mathbf{P}(\sum_{i=1}^m \xi_i \leq \frac{m\rho}{2}) \leq \exp\left(-\frac{m\rho^2}{2}\right).$$

For (c), we fixed a $\frac{\eta\rho}{4C}$-net $\mathcal{N}$ to cover the unit sphere S^{p-1}. By a volume argument, $|\mathcal{N}| \leq (1 + \frac{8C}{\eta\rho})^p$. Under the event $\{\|\mathbb{X}\| \leq C\sqrt{m}\}$ and the event

$$\{\sum_{i=1}^m |\boldsymbol{X}_i^\top\boldsymbol{\beta}_k| \geq \frac{m\rho}{2}\eta \text{ for all } \boldsymbol{\beta}_k \in \mathcal{N}\},$$

for any $\|\boldsymbol{\beta}\| = 1$, we can pick $\boldsymbol{\beta}_1 \in \mathcal{N}$ such that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_1\| \leq \frac{\eta\rho}{4C}$. Thus,

$$\begin{aligned}&\frac{1}{m}\left(\sum_{i=1}^m |\boldsymbol{X}_i^\top\boldsymbol{\beta}| - \sum_{i=1}^m |\boldsymbol{X}_i^\top\boldsymbol{\beta}_1|\right)\\ \leq &\frac{1}{m}\sum_{i=1}^m |\boldsymbol{X}_i^\top(\boldsymbol{\beta} - \boldsymbol{\beta}_1)| \leq \frac{1}{\sqrt{m}}\|\mathbb{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_1)\| \leq \frac{\eta\rho}{4C\sqrt{m}}\|\mathbb{X}\| \leq \frac{\eta\rho}{4},\end{aligned}$$

where the first inequality is due to the triangle inequality, the second is due to the generalized mean inequality, and the third is from the definition of the operator norm and $\|\boldsymbol{\beta} - \boldsymbol{\beta}_1\| \leq \frac{\eta\rho}{4C}$. It follows that under these two events, we have $\inf_{\|\boldsymbol{\beta}\|=1} \frac{1}{m}\sum_{i=1}^m |\boldsymbol{X}_i^\top\boldsymbol{\beta}| \geq \eta\rho/4$. This also implies that $\mathbb{X}\boldsymbol{\beta}$ is a non-zero vector for any $\boldsymbol{\beta}$, and thus $\text{rank}(\mathbb{X}) = p$. Since the union bound on the exception probability of these two events is $(1 + \frac{8C}{\eta\rho})^p\exp\left(-\frac{m\rho^2}{2}\right) + e^{-cm}$, we proved (c). □

In order to use Lemma 5.8 to prove Theorem 5.7, we need to establish a lower bound on the small ball probability and an upper bound on the operator norm of the synthetic covariate matrix.

B.1. Lower Bounds on Small Ball Probability. We first show the condition Eq. (5.7) in Lemma 5.8 is implied by Condition 5.5. The proof makes use of two classic results: the first lemma is a direct implication of Hoeffding's inequality, and the second lemma is standard in the literature of the Littlewood–Offord problem.

Lemma 5.9. *If $\boldsymbol{X}$ satisfies Condition 5.5, then for any $\boldsymbol{y} \in \mathbb{R}^p$ with $\|\boldsymbol{y}\| = 1$ and $|y_1| < 1$, and $t > 0$*

$$\max\left(\mathbf{P}(\sum_{i=2}^{p} y_i X_i > t), \mathbf{P}(\sum_{i=2}^{p} y_i X_i < -t)\right) \leq \mathit{exp}\left(-\frac{t^2}{2B_1^2(1-y_1^2)}\right)$$

Lemma 5.10 (Lemma 3.1 in Ref. (12))**.** *Let $2 < r \leq 3$ and $\mu \geq 1$. Suppose $\xi_1, \cdots, \xi_q$ are independent centered r.v. with $\mathbb{E}|\xi_i|^r \leq \mu^r$ for all $i = 1, \cdots, q$. Let $\boldsymbol{y} \in \mathbb{R}^q$ and $\|\boldsymbol{y}\| = 1$. Then for every $\lambda \geq 0$*

$$\mathbf{P}(|\sum_i \xi_i y_i| > \lambda) \geq \left(\frac{(\mathbb{E}\sum_i \xi_i^2 y_i^2 - \lambda^2)_+}{8\mu^2}\right)^{r/(r-2)}.$$

We are now ready for the main result on the small ball probability.

Proposition 5.11. *Assume $\boldsymbol{X}$ satisfy Condition 5.5 then there exist positive constants η_0 and ρ_0 that only depend on B_1, such that*

$$\mathbf{P}(|\boldsymbol{\beta}^\top \boldsymbol{X}| > \eta_0) \geq \rho_0, \quad \forall \|\boldsymbol{\beta}\| = 1. \tag*{[5.8]}$$

Proof. For any $\rho, \eta \in [0, 1)$, define a function $f(\eta, \rho) = \frac{\eta + \sqrt{(1-2B_1^2\log(1-\rho)-\eta^2)(-2B_1^2\log(1-\rho))}}{1-2B_1^2\log(1-\rho)}$; $f(\eta, \rho)$ is the larger root in $[0, 1)$ of the following equation

$$x - \sqrt{(1-x^2)2B_1^2\log(\frac{1}{1-\rho})} = \eta. \tag*{[5.9]}$$

For any $\eta, \rho \in (0, 1)$, we separately consider $|\boldsymbol{\beta}_1| \geq f(\eta, \rho)$ and $|\boldsymbol{\beta}_1| < f(\eta, \rho)$.

- If $|\beta_1| \geq f(\eta, \rho)$, without loss of generality, assume $\beta_1 > 0$. Let $t^2 = 2B_1^2(1-\beta_1^2)\log(\frac{1}{1-\rho})$. Lemma 5.9 reads

  $$\mathbf{P}\left(\sum_{i=2}^{p} \beta_i X_i \geq -t\right) \geq \rho \tag*{[5.10]}$$

  Since $f(\eta, \rho)$ is the larger root of Eq. (5.9) and $\boldsymbol{\beta}_1 \geq f(\eta, \rho)$, one can check $\boldsymbol{\beta}_1 - t = \boldsymbol{\beta}_1 - \sqrt{(1-\boldsymbol{\beta}_1^2)2B_1^2\log(\frac{1}{1-\rho})} \geq \eta$. Thus, $\mathbf{P}(\beta_1 + \sum_{i=2}^{p} \beta_i X_i \geq \eta) \geq \rho$.

- If $|\beta_1| < f(\eta, \rho)$, Lemma 5.10 with $r = 3$, $q = p - 1$, and $\boldsymbol{y} = (\beta_2, \ldots, \beta_p)/\sqrt{1-\beta_1^2}$ implies that for any $s > 0$,

  $$\mathbf{P}\left(|\sum_{i=2}^{p} \beta_i X_i| \geq s\right) \geq \left(\frac{1-\beta_1^2-s^2}{(1-\beta_1^2)8B_1^2}\right)^3 \geq \left(\frac{1-f^2(\eta,\rho)-s^2}{(1-\beta_1^2)8B_1^2}\right)^3.$$

  For any $s \geq f(\eta, \rho) + \eta$, with $|\beta_1| \leq f(\eta, \rho)$, we have

  $$\mathbf{P}(|\beta_1 + \sum_{i=2}^{p} \beta_i X_i| \geq \eta) \geq \left(\frac{1-f^2(\eta,\rho)-s^2}{(1-\beta_1^2)8B_1^2}\right)^3. \tag*{[5.11]}$$

  Note that $f(\eta, \rho)$ is continuous on $[0, 1] \times [0, 1/2]$ and $f(0, 0) = 0$. The same holds for $(f(\eta, \rho) + \eta)^2 + f(\eta, \rho)^2$. Therefore, there is some constant $C > 0$ only depends on B_1, such that for all $\eta, \rho \in (0, C]$,

  $$(f(\eta, \rho) + \eta)^2 + f(\eta, \rho)^2 < 1.$$

  Now take $s = f(\eta, \rho) + \eta$ and substitute it into Eq. (5.11). We see that the right-hand side of Eq. (5.11) is positive for any $\eta, \rho \in (0, C]$

Finally, fix any $0 < \eta_0 \leq C$ and choose $\rho_0 = \min(C, (\frac{1-f^2(\eta_0,C)-(f(\eta_0,C)+\eta_0)^2}{8B_1^2})^3)$. Combining the above two cases with $\eta = \eta_0$ and $\rho = C$, we showed

$$\mathbf{P}(|\beta_1 + \sum_{i=2}^{p} \beta_i X_i| \geq \eta_0) \geq \min(C, \left(\frac{1-f^2(\eta_0,C)-(f(\eta_0,C)+\eta_0)^2}{8B_1^2}\right)^3) = \rho_0 > 0.$$

□

B.2. Upper Bounds for Operator Norm. We now show that the condition in Part **(c)** of Lemma 5.8 is guaranteed by Condition 5.5, by deriving an upper bound on the operator norm of the synthetic covariate matrix.

Proposition 5.12. *If i.i.d. covariates $\boldsymbol{X}_i$'s satisfies Condition 5.5 for $i = 1, \dots, M$, and $M \geq p$, then*

$$\mathbf{P}\left(\|\mathbb{X}\| > t\sqrt{2M}\right) \leq e^{-c_0 t^2 M}, \quad \forall t \geq C_0$$

where the constants $C_0, c_0 > 0$ depend only on B_1, the constant in Condition 5.5.

Before the proof, we recall a classic result, Lemma 5.13, about the largest singular value of a random matrix. This lemma holds for general independent subgaussian random variables. A r.v. ξ is *subgaussian* if its tail is dominated by that of a normal random variable: there exists $B_\xi > 0$ such that

$$\mathbf{P}(|\xi| > t) \leq 2\exp(-\frac{t^2}{B_\xi^2}), \quad \forall t > 0. \qquad [5.12]$$

The subgaussian moment of ξ is the minimal B_ξ such that this inequality holds. Note that by Hoeffding's inequality, a bounded and centered r.v. $\xi \in [a, b]$ is subgaussian with subgaussian moment $\frac{b-a}{2}$. Therefore, if $\boldsymbol{X}$ satisfies Condition 5.5 then its coordinates are independent subgaussian with subgaussian moment B_1.

Lemma 5.13 (Proposition 2.3 in Ref. (13))**.** *Let $\mathbb{W}$ be an $m \times p$ random matrix, $m \geq p$, whose elements are independent subgaussian random variables with uniformly bounded subgaussian moments. Then*

$$\mathbf{P}\left(\|\mathbb{W}\| > t\sqrt{m}\right) \leq e^{-c_0 t^2 m}, \quad \forall t \geq C_0,$$

where $C_0 > 0$ and $c_0 > 0$ depend only on the subgaussian moment B.

Proof of Proposition 5.12. Let $\mathbb{W}$ be the sub-matrix of $\mathbb{X}$ without the first column (recall the first column corresponds to the constant term). By Lemma 5.13, we have

$$\mathbf{P}\left(\|\mathbb{W}\| > t\sqrt{M}\right) \leq e^{-c_0 t^2 M}, \quad \forall t \geq C_0.$$

For any $t \geq \max(1, C_0)$, on the event $\{\|\mathbb{W}\| \leq t\sqrt{M}\}$, we have that for any $\boldsymbol{\beta} \in \mathbb{R}^p$

$$\begin{aligned}\|\mathbb{X}\boldsymbol{\beta}\|^2 = &\ \beta_1^2 M + \boldsymbol{\beta}_{-1}^\top \mathbb{W}^\top \mathbb{W}\boldsymbol{\beta}_{-1} + 2\beta_1 \mathbf{1}_M^\top \mathbb{W}\boldsymbol{\beta}_{-1} \\ \leq &\ 2(\beta_1^2 M + \boldsymbol{\beta}_{-1}^\top \mathbb{W}^\top \mathbb{W}\boldsymbol{\beta}_{-1}) \leq\ 2(\beta_1^2 M + t^2 M(1-\beta_1^2)) \leq 2t^2 M.\end{aligned}$$

Thus, for any $t \geq \max(1, C_0)$,

$$\mathbf{P}\left(\|\mathbb{X}\| > t\sqrt{2M}\right) \leq \mathbf{P}\left(\|\mathbb{W}\| > t\sqrt{M}\right) \leq e^{-c_0 t^2 M}.$$

□

B.3. Synthesis.

Proof of Theorem 5.7. By Proposition 5.11, Condition 5.5 implies that there are η_0 and ρ_0 depending on B_1 such that Eq. (5.7) holds for $\boldsymbol{W} = \boldsymbol{X}$. This proves the result of Part **(a)**. In addition, Part **(a)** of Lemma 5.8 implies $\mathbb{E}(|\boldsymbol{X}^\top \boldsymbol{\beta}|) \geq \eta_0\rho_0$ for any $\boldsymbol{\beta}$ with norm 1. Therefore, we obtain the result of Part **(b)**. By Proposition 5.12, Condition 5.5 implies that the condition in Part **(c)** of Lemma 5.8 holds. Then Lemma 5.8, together with Eq. (5.7), implies the result of Part **(c)**. □

C. Properness of Catalytic Priors.

C.1. Catalytic Priors with Fixed Prior Weight. We begin with Theorem 5.14 and Corollary 5.15, which show the properness of catalytic priors with finite M. These two results are combined into one presented as Theorem 1 in the main text.

Theorem 5.14. *Assume the synthetic covariate matrix $\mathbb{X}^*$ has full column rank and each synthetic response Y_i^* lies in $\mathcal{Y}$. Suppose $c_\phi := \inf_{\eta \neq 0} |\phi(\eta)/\eta| > 0$. Then the catalytic prior with finite M is proper for any $\tau > 0$.*

Proof of Theorem 5.14. For all $\boldsymbol{\beta} \in \mathbb{R}^p$, it holds that

$$\frac{1}{M}\sum_{i=1}^{M} |\phi((\boldsymbol{X}_i{}^*)^\top \boldsymbol{\beta})| \geq c_\phi \frac{1}{M}\sum_{i=1}^{M} |(\boldsymbol{X}_i{}^*)^\top \boldsymbol{\beta}| \geq c_\phi \frac{1}{M}\sqrt{\sum_{i=1}^{M}((\boldsymbol{X}_i{}^*)^\top \boldsymbol{\beta})^2} \geq c_\phi \sqrt{\frac{\sigma_{\min}}{M}}\|\boldsymbol{\beta}\|^2,$$

where $\sigma_{\min}$ is the smallest eigenvalue of $(\mathbb{X}^*)^\top \mathbb{X}^*/M$ (which is positive because $\mathbb{X}^*$ has full column rank). This means Condition (2) in Theorem 5.2 holds.

Furthermore, since each Y_i^*, $i = 1, \ldots, M$, lies in $\mathcal{Y}$, which is an open set, there exist $u_-, u_+ \in \mathcal{Y}, \delta > 0$, such that $u_- \leq Y_i^* - \delta < Y_i^* + \delta \leq u_+$, for all $i = 1, \cdots, M$. Thus, Condition (1) in Theorem 5.2 holds. Theorem 5.2 (with $u_i = Y_i^*$, $w_i = \boldsymbol{X}_i^*$, $\alpha = \tau$, $c_1 = \sqrt{\sigma_{\min}/M}$) implies that the integral of the unnormalized density function is finite, which means the catalytic prior with finite M is proper. □

Theorem 5.14 requires every synthetic response to lie in $\mathcal{Y}$, which is guaranteed when the synthetic response is taken to be the predictive mean of the sufficient statistic as in the case of exponential families, as described in Section *Catalytic Prior for GLM* in the main text. Corollary 5.15 relaxes this condition and allows a synthetic response to be on the boundary of $\mathcal{Y}$. The condition 2 in Corollary 5.15 can be easily satisfied when the stratified synthetic data generation is used.

Corollary 5.15. *Let $\{(\boldsymbol{X}_i^*, Y_i^*) : 1 \leq i \leq M\}$ be the synthetic dataset. Suppose*

1. *$c_\phi := \inf_{\eta \neq 0} |\phi(\eta)/\eta| > 0$,*
2. *there exists a set of linearly independent covariate vectors $\{\boldsymbol{x}_j^{(0)}\}_{j=1}^p$ such that for each $1 \leq j \leq p$*

$$\frac{1}{\#\{i : \boldsymbol{X}_i^* = \boldsymbol{x}_j^{(0)}\}} \sum_{i:\boldsymbol{X}_i^* = \boldsymbol{x}_j^{(0)}} Y_i^* \quad \in \mathcal{Y}, \qquad [5.13]$$

3. *for each $1 \leq i \leq M$, $\sup_\theta (Y_i^*\theta - b(\theta)) \leq \log C$ for a constant C,*

then the catalytic prior based on $\{(\boldsymbol{X}_i^*, Y_i^*) : 1 \leq i \leq M\}$ is proper for any $\tau > 0$.

Proof. We define another set of synthetic data points and weights:

$$\tilde{\boldsymbol{X}}_j^* = \boldsymbol{x}_j^{(0)}, \quad \tilde{Y}_j^* = \frac{1}{\#\{i : \boldsymbol{X}_i^* = \boldsymbol{x}_j^{(0)}\}} \sum_{i:\boldsymbol{X}_i^* = \boldsymbol{x}_j^{(0)}} Y_i^*, \quad \tilde{w}_j = \frac{\tau}{M}\#\{i : \boldsymbol{X}_i^* = \boldsymbol{x}_j^{(0)}\}, \quad 1 \leq j \leq p \qquad [5.14]$$

Denote by $\tilde{\pi}$ the catalytic prior corresponding to the synthetic dataset $\{(\tilde{\boldsymbol{X}}_j^*, \tilde{Y}_j^*) : 1 \leq j \leq p\}$ and weights $\tilde{w}_j$. Using the proof of Theorem 5.14, we see that $\tilde{\pi}$ is proper. It follows that

$$\begin{aligned}
&\int_{\mathbb{R}^p} \exp\left(\frac{\tau}{M}\sum_{i=1}^{M}(Y_i^*\phi((\boldsymbol{X}_i{}^*)^\top \boldsymbol{\beta}) - b(\phi((\boldsymbol{X}_i{}^*)^\top \boldsymbol{\beta}))\right) d\boldsymbol{\beta} \\
&\leq C^{\frac{\tau}{M}(M - \sum_{j=1}^{p}\tilde{w}_j)} \int_{\mathbb{R}^p} \exp\left(\frac{\tau}{M}\sum_{j=1}^{p}\#\{i : \boldsymbol{X}_i^* = \boldsymbol{x}_j^{(0)}\} \times (\tilde{Y}_j^*\phi((\tilde{\boldsymbol{X}}_j^*)^\top \boldsymbol{\beta}) - b(\phi((\tilde{\boldsymbol{X}}_j^*)^\top \boldsymbol{\beta}))\right) d\boldsymbol{\beta} < \infty,
\end{aligned}$$

where the first inequality comes from rearranging the terms and the third condition of the corollary, and the second inequality is due to the properness of $\tilde{\pi}$. This completes the proof. □

Theorem 5.14 and Corollary 5.15 show that if the synthetic covariate matrix has full column rank, then the catalytic prior with finite M is proper for any $\tau > 0$. The following theorem relaxes the full column rank assumption and shows that the properness of catalytic priors can also be guaranteed with high probability if the synthetic-data generating distribution satisfies some mild conditions. It also provides upper bounds on the tail integrals of the catalytic priors, which are used in Section D to study the convergence of catalytic priors.

Theorem 5.16. *Suppose (i) there exists a compact subset $\mathcal{Y}^{com}$ of $\mathcal{Y}$ such that every synthetic response is in $\mathcal{Y}^{com}$ with probability 1, (ii) the synthetic-covariate generating distribution satisfies Condition 5.5, and (iii) $c_\phi := \inf_{\eta \neq 0} |\phi(\eta)/\eta| > 0$. Then there exist constants C_* and δ that only depend on $\mathcal{Y}^{com}$ and the exponential family, and constants ρ_0, η_0, c and C that only depend on B_1, such that with probability at least*

$$1 - e^{-cM} - exp\left(-\frac{M\rho_0^2}{2} + p\log(1 + \frac{8C}{\eta_0\rho_0})\right),$$

the following holds

(a) $\mathbb{X}^*$ has full column rank, the catalytic prior is proper for any $\tau > 0$, and

$$\int_{\|\boldsymbol{\beta}\| \in \mathbb{R}^p} exp\left\{\frac{\tau}{M}\sum_{i=1}^{M}\left(Y_i^*\phi((\boldsymbol{X}_i{}^*)^\top\boldsymbol{\beta}) - b(\phi((\boldsymbol{X}_i{}^*)^\top\boldsymbol{\beta}))\right)\right\} d\boldsymbol{\beta} \leq C_{Stirling}\frac{C^\tau(32\pi p)^{p/2}}{(\tau c_\phi \eta_0\rho_0\delta)^p}.$$

(b) $\sup_{\boldsymbol{\beta}\in\mathbb{R}^p}\max_{i\leq M}\left(Y_i^*\phi((\boldsymbol{X}_i{}^*)^\top\boldsymbol{\beta}) - b(\phi((\boldsymbol{X}_i{}^*)^\top\boldsymbol{\beta}))\right) \leq \log C_*$

(c) For any $K > 4p/(\tau\delta c_\phi\eta_0\rho_0)$, we have

$$\int_{\|\boldsymbol{\beta}\| > K} exp\left\{\frac{\tau}{M}\sum_{i=1}^{M}\left(Y_i^*\phi((\boldsymbol{X}_i{}^*)^\top\boldsymbol{\beta}) - b(\phi((\boldsymbol{X}_i{}^*)^\top\boldsymbol{\beta}))\right)\right\} d\boldsymbol{\beta}$$
$$\leq C_{Stirling}C_*^\tau\, exp(p - \tau c_\phi\eta_0\rho_0\delta K/4)(\frac{\sqrt{2\pi}K}{\sqrt{pe}})^p.$$

Before proving Theorem 5.16, we state an analogous theorem for population catalytic priors. The proof for both theorems relies on the same idea and uses Theorem 5.7.

Theorem 5.17. *Under the same conditions as in Theorem 5.16, the following holds*

(a) the population catalytic prior is proper for any $\tau > 0$ and

$$\int_{\|\boldsymbol{\beta}\| \in \mathbb{R}^p} exp\left\{\tau\mathbb{E}[Y^*\phi((\boldsymbol{X}^*)^\top\boldsymbol{\beta}) - b(\phi((\boldsymbol{X}^*)^\top\boldsymbol{\beta}))]\right\} d\boldsymbol{\beta} \leq C_{Stirling}\frac{C^\tau(2\pi p)^{p/2}}{(\tau c_\phi\eta_0\rho_0\delta)^p}.$$

(b) $\mathbb{E}[Y^*\phi((\boldsymbol{X}^*)^\top\boldsymbol{\beta}) - b(\phi((\boldsymbol{X}^*)^\top\boldsymbol{\beta}))] \leq \log C_*$

(c) There exists a universal constant $C_{Stirling}$, such that for any $K > p/(\tau\delta c_\phi\eta_0\rho_0)$, we have

$$\int_{\|\boldsymbol{\beta}\| > K} exp\left\{\tau\mathbb{E}[Y^*\phi((\boldsymbol{X}^*)^\top\boldsymbol{\beta}) - b(\phi((\boldsymbol{X}^*)^\top\boldsymbol{\beta}))]\right\} d\boldsymbol{\beta}$$
$$\leq C_{Stirling}C_*^\tau\, exp(p - \tau c_\phi\eta_0\rho_0\delta K)(\frac{\sqrt{2\pi}K}{\sqrt{pe}})^p.$$

Proof of Theorem 5.16 and Theorem 5.17. Since $\mathcal{Y}^{\text{com}}$ is a compact subset of $\mathcal{Y}$, using an open covering argument, there exist some u_-, $u_+ \in \mathcal{Y}$ and $\delta > 0$ such that for any $u \in \mathcal{Y}^{\text{com}}$, it holds that $u_- + \delta \leq u \leq u_+ - \delta$. Since the synthetic response is in $\mathcal{Y}^{\text{com}}$, so is the conditional expectation given the synthetic covariates. Thus, Condition (1) in Theorem 5.3 and Condition (1) in Theorem 5.2 hold.

For Theorem 5.17

Under Condition 5.5, Theorem 5.7 implies that for any $\boldsymbol{\beta} \in \mathbb{R}^p$

$$\mathbb{E}|(\boldsymbol{X}^*)^\top \boldsymbol{\beta}| \geq \eta_0 \rho_0 \|\boldsymbol{\beta}\|.$$

Therefore, for all $\boldsymbol{\beta} \in \mathbb{R}^p$, it holds that

$$\mathbb{E}|\phi((\boldsymbol{X}^*)^\top \boldsymbol{\beta})| \geq c_\phi \mathbb{E}|(\boldsymbol{X}^*)^\top \boldsymbol{\beta}| \geq c_\phi \eta_0 \rho_0 \|\boldsymbol{\beta}\|.$$

That is, Condition (2) in Theorem 5.3 holds with $c_0 = c_\phi \eta_0 \rho_0$. By Theorem 5.3 (with $U = Y^*$, $\boldsymbol{W} = \boldsymbol{X}^*$ and $\alpha = \tau$), we conclude Theorem 5.17.

For Theorem 5.16

Under Condition 5.5, Theorem 5.7 implies that with probability at least $1 - e^{-cM} - \exp\left(-\frac{M\rho_0^2}{2} + p\log(1+\frac{8C}{\eta_0\rho_0})\right)$, $\mathbb{X}^*$ has full column rank and it holds that

$$\inf_{\|\boldsymbol{\beta}\|=1} \frac{1}{M}\sum_{i=1}^{M} |\phi((\boldsymbol{X}_i{}^*)^\top \boldsymbol{\beta})| \geq \frac{c_\phi \eta_0 \rho_0}{4}.$$

Thus, Condition (2) in Theorem 5.2 holds with $c_1 = c_\phi \eta_0 \rho_0 / 4$. By Theorem 5.2 (with $u_i = Y_i^*$, $w_i = \boldsymbol{X}_i^*$ and $\alpha = \tau$), we conclude Theorem 5.16. □

C.2. Properness of the Joint Priors for (τ, β).

Theorem 5.18. *If $\Gamma_{\alpha,\gamma}(\tau)$ is taken as* Eq. (16) *of the main text for linear regression or as* Eq. (17) *of the main text for other generalized linear models, where both α and γ are positive, then under the same conditions in either Theorem 5.14 or Corollary 5.15, the following holds:*

(1) The joint prior on $(\tau, \boldsymbol{\beta})$ is proper;

(2) For any $\alpha' \in (0, \alpha)$, the α'-th moment of $\boldsymbol{\beta}$ exists;

(3) Denoting by $h_{\alpha,\gamma}(\tau)$ the marginal prior on τ. If the MLE based on the synthetic data exists, then $\lim_{\tau\to\infty} \frac{1}{\tau}\log h_{\alpha,\gamma}(\tau) = -1/\gamma < 0$.

Remark 5.19. The conclusions (2) and (3) indicate how the hyper-parameters α and γ affect the joint prior. Specifically, α controls the tail behavior of $\boldsymbol{\beta}$: the larger α, the lighter the tail $\boldsymbol{\beta}$ has; γ controls the tail behavior of τ: the larger γ, the heavier the tail τ has. □

Proof. Denote by $\ell(\boldsymbol{\beta})$ the log likelihood based on the synthetic data:

$$\ell(\boldsymbol{\beta}) = \frac{1}{M}\sum_{i=1}^{M} \left(Y_i^* \phi((\boldsymbol{X}_i{}^*)^\top \boldsymbol{\beta}) - b(\phi((\boldsymbol{X}_i{}^*)^\top \boldsymbol{\beta}))\right).$$

Since Conclusion (2) implies (1), we only need to prove (2) and (3).

Part 1. We first prove Conclusion (2). The proof is adapted from the proof of Theorem 3.1 in Ref. (14).

1(a). Suppose $\Gamma_{\alpha,\gamma}(\tau)$ is taken as Eq. (17) of the main text for GLMs. By Tonelli's theorem, for any $\alpha' \in (0, \alpha)$,

$$\int_{\tau>0}\int_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{\beta}\|^{\alpha'} \tau^{p+\alpha-1} \exp\left(-(\kappa + \frac{1}{\gamma})\tau + \tau\ell(\boldsymbol{\beta})\right) d\boldsymbol{\beta} d\tau = \int_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{\beta}\|^{\alpha'} \frac{\Gamma(p+\alpha)}{(\kappa + \gamma^{-1} - \ell(\boldsymbol{\beta}))^{p+\alpha}} d\boldsymbol{\beta}.$$

From the proof of Theorem 5.14 and the proof of Corollary 5.15, we know that there exist positive constants c_1 and C_1 such that

$$\exp(\ell(\boldsymbol{\beta})) \leq C_1 \exp(-c_1\|\boldsymbol{\beta}\|), \quad \forall \boldsymbol{\beta} \neq 0. \qquad [5.15]$$

Split the integral in Eq. (5.15) into two: $\int_{\ell(\boldsymbol{\beta})\leq \kappa - c_1\|\boldsymbol{\beta}\|/2}$ and $\int_{\ell(\boldsymbol{\beta})>\kappa - c_1\|\boldsymbol{\beta}\|/2}$. We will separately bound the two integrals (without the constant term $\Gamma(p+\alpha)$ there).

For the first integral, we have

$$\begin{aligned}
&\int_{\ell(\boldsymbol{\beta})\leq\kappa-c_1\|\boldsymbol{\beta}\|/2}\|\boldsymbol{\beta}\|^{\alpha'}\frac{1}{(\kappa+1/\gamma-\ell(\boldsymbol{\beta}))^{p+\alpha}}d\boldsymbol{\beta}\\
\leq\ &\int_{\ell(\boldsymbol{\beta})\leq\kappa-c_1\|\boldsymbol{\beta}\|/2}\|\boldsymbol{\beta}\|^{\alpha'}\frac{1}{(1/\gamma+c_1\|\boldsymbol{\beta}\|/2)^{p+\alpha}}d\boldsymbol{\beta}\leq\int_{\mathbb{R}^p}\frac{\|\boldsymbol{\beta}\|^{\alpha'}}{(1/\gamma+c_1\|\boldsymbol{\beta}\|/2)^{p+\alpha}}d\boldsymbol{\beta},
\end{aligned}$$

where the last integral is finite by elementary calculus using the fact that $\alpha'\in(0,\alpha)$.

For the second integral, we have

$$\begin{aligned}
&\int_{\ell(\boldsymbol{\beta})>\kappa-c_1\|\boldsymbol{\beta}\|/2}\|\boldsymbol{\beta}\|^{\alpha'}\frac{1}{(\kappa+1/\gamma-\ell(\boldsymbol{\beta}))^{p+\alpha}}d\boldsymbol{\beta}\\
\leq\ &\int_{\ell(\boldsymbol{\beta})>\kappa-c_1\|\boldsymbol{\beta}\|/2}\|\boldsymbol{\beta}\|^{\alpha'}\frac{1}{(1/\gamma)^{p+\alpha}}d\boldsymbol{\beta}\\
\leq\ &\int_{\ell(\boldsymbol{\beta})>\kappa-c_1\|\boldsymbol{\beta}\|/2}\|\boldsymbol{\beta}\|^{\alpha'}\frac{1}{(1/\gamma)^{p+\alpha}}\exp(\ell(\boldsymbol{\beta})+c_1\|\boldsymbol{\beta}\|/2-\kappa)d\boldsymbol{\beta}\\
\leq\ &C_1\int_{\ell(\boldsymbol{\beta})>\kappa-c_1\|\boldsymbol{\beta}\|/2}\gamma^{p+\alpha}e^{-\kappa}\|\boldsymbol{\beta}\|^{\alpha'}\exp(-c_1\|\boldsymbol{\beta}\|+c_1\|\boldsymbol{\beta}\|/2)d\boldsymbol{\beta}\\
\leq\ &C_1\int_{\boldsymbol{\beta}\in\mathbb{R}^p}\gamma^{p+\alpha}e^{-\kappa}\|\boldsymbol{\beta}\|^{\alpha'}\exp(-c_1\|\boldsymbol{\beta}\|/2)d\boldsymbol{\beta},
\end{aligned}$$

where the first inequality is because by definition of κ it is no less than $\ell(\boldsymbol{\beta})$, the second inequality is due to the fact that $\ell(\boldsymbol{\beta})+c_1\|\boldsymbol{\beta}\|/2-\kappa\geq 0$ in the domain of the integral, and the third inequality is due to Eq. (5.15). The last expression is finite due to its exponential tail. Therefore, we prove the conclusion of (2) for GLMs.

1(b). Suppose $\Gamma_{\alpha,\gamma}(\tau)$ is taken as Eq. (16) of the main text for linear regression. By Tonelli's theorem, for any $\alpha'\in(0,\alpha)$, we have

$$\int_{\tau>0}\int_{\boldsymbol{\beta}\in\mathbb{R}^p}\|\boldsymbol{\beta}\|^{\alpha'}\tau^{(p+\alpha)/2-1}\exp\left(-(\kappa+\frac{1}{\gamma})\tau+\tau\ell(\boldsymbol{\beta})\right)d\boldsymbol{\beta}d\tau=\int_{\boldsymbol{\beta}\in\mathbb{R}^p}\|\boldsymbol{\beta}\|^{\alpha'}\frac{\Gamma((p+\alpha)/2)}{(\kappa+1/\gamma-\ell(\boldsymbol{\beta}))^{(p+\alpha)/2}}d\boldsymbol{\beta}. \quad [5.16]$$

Eq. (1.1) in Section 1 implies that there exists a positive constant c_2 only depending on the noise variance and the smallest singular value of the synthetic covariate matrix such that $\ell(\boldsymbol{\beta})\leq\kappa-c_2\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}_0\|^2$. Splitting the integral in Eq. (5.16) into two: $\int_{\ell(\boldsymbol{\beta})\leq\kappa-c_2\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}_0\|^2/2}$ and $\int_{\ell(\boldsymbol{\beta})>\kappa-c_2\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}_0\|^2/2}$. We can bound Eq. (5.16) similarly as before (ignoring the constant $\Gamma((p+\alpha)/2)$ in the numerator):

$$\begin{aligned}
&\int_{\ell(\boldsymbol{\beta})\leq\kappa-c_2\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}_0\|^2/2}\|\boldsymbol{\beta}\|^{\alpha'}\frac{1}{(\kappa+1/\gamma-\ell(\boldsymbol{\beta}))^{(p+\alpha)/2}}d\boldsymbol{\beta}\\
\leq\ &\int_{\ell(\boldsymbol{\beta})\leq\kappa-c_2\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\|^2/2}\|\boldsymbol{\beta}\|^{\alpha'}\frac{1}{(1/\gamma+c_2\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}_0\|^2/2)^{(p+\alpha)/2}}d\boldsymbol{\beta}\\
\leq\ &\int_{\boldsymbol{\beta}\in\mathbb{R}^p}\|\boldsymbol{\beta}\|^{\alpha'}\frac{1}{(1/\gamma+c_2\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}_0\|^2/2)^{(p+\alpha)/2}}d\boldsymbol{\beta}<\infty,
\end{aligned}$$

and

$$\begin{aligned}
&\int_{\ell(\boldsymbol{\beta})>\kappa-c_2\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}_0\|^2/2}\|\boldsymbol{\beta}\|^{\alpha'}\frac{1}{(\kappa+1/\gamma-\ell(\boldsymbol{\beta}))^{(p+\alpha)/2}}d\boldsymbol{\beta}\\
\leq\ &\int_{\ell(\boldsymbol{\beta})>\kappa-c_2\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}_0\|^2/2}\|\boldsymbol{\beta}\|^{\alpha'}\frac{1}{(1/\gamma+0)^{(p+\alpha)/2}}\exp\left(\ell(\boldsymbol{\beta})+c_2\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}_0\|^2/2-\kappa\right)d\boldsymbol{\beta}\\
\leq\ &\int_{\ell(\boldsymbol{\beta})>\kappa-c_2\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}_0\|^2/2}\|\boldsymbol{\beta}\|^{\alpha'}\gamma^{(p+\alpha)/2}\exp\left(-c_2\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}_0\|^2+c_2\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}_0\|^2/2\right)d\boldsymbol{\beta}\\
\leq\ &\int_{\boldsymbol{\beta}\in\mathbb{R}^p}\|\boldsymbol{\beta}\|^{\alpha'}\gamma^{(p+\alpha)/2}\exp\left(-c_2\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}_0\|^2/2\right)d\boldsymbol{\beta}<\infty.
\end{aligned}$$

Combining the two cases, we conclude (1) and (2).

Part 2. We now prove Conclusion (3). Write the marginal prior for τ as

$$h(\tau) \propto \tau^{p+\alpha-1} \exp\left(-\tau/\gamma\right) \int_{\boldsymbol{\beta}\in\mathbb{R}^p} e^{\tau(\ell(\boldsymbol{\beta})-\kappa)} d\boldsymbol{\beta}.$$

By L'Hôpital's rule,

$$\lim_{\tau\to\infty} \frac{\log h(\tau)}{\tau} = -1/\gamma + \lim_{\tau\to\infty} \frac{\int_{\boldsymbol{\beta}\in\mathbb{R}^p} (\ell(\boldsymbol{\beta})-\kappa) e^{\tau(\ell(\boldsymbol{\beta})-\kappa)} d\boldsymbol{\beta}}{\int_{\boldsymbol{\beta}\in\mathbb{R}^p} e^{\tau(\ell(\boldsymbol{\beta})-\kappa)} d\boldsymbol{\beta}}. \tag*{[5.17]}$$

It remains to show

$$\lim_{\tau\to\infty} \frac{\int_{\boldsymbol{\beta}\in\mathbb{R}^p} (\ell(\boldsymbol{\beta})-\kappa) e^{\tau(\ell(\boldsymbol{\beta})-\kappa)} d\boldsymbol{\beta}}{\int_{\boldsymbol{\beta}\in\mathbb{R}^p} e^{\tau(\ell(\boldsymbol{\beta})-\kappa)} d\boldsymbol{\beta}} = 0. \tag*{[5.18]}$$

We state the following lemma.

Lemma 5.20. *Suppose $f(\boldsymbol{x})$ is a continuous function on $\mathbb{R}^p$, and $\boldsymbol{x}_0 \in \mathbb{R}^p$ uniquely minimizes $f(\boldsymbol{x})$ and $f(\boldsymbol{x}_0) = 0$. Furthermore, if there are some constants C and ω such that*

$$f(\boldsymbol{x}) \geq C + \omega \|\boldsymbol{x}\|, \tag*{[5.19]}$$

then

$$\lim_{\tau\to\infty} \frac{\int_{\boldsymbol{x}\in\mathbb{R}^p} f(\boldsymbol{x}) e^{-\tau f(\boldsymbol{x})} d\boldsymbol{x}}{\int_{\boldsymbol{x}\in\mathbb{R}^p} e^{-\tau f(\boldsymbol{x})} d\boldsymbol{x}} = 0.$$

By Theorem 5.14 and Corollary 5.15, the condition for Lemma 5.20 with function $\kappa - \ell(\boldsymbol{\beta})$ holds, and we conclude Eq. (5.18). □

Proof of Lemma 5.20. Without loss of generality, we can assume $\boldsymbol{x}_0 = \boldsymbol{0}$; otherwise, let $\tilde{\boldsymbol{x}} = \boldsymbol{x} - \boldsymbol{x}_0$, $\tilde{C} = (C - \omega\|\boldsymbol{x}_0\|)$, then $\tilde{f}(\tilde{\boldsymbol{x}}) = f(\tilde{\boldsymbol{x}} + \boldsymbol{x}_0) \geq (C - \omega\|\boldsymbol{x}_0\|) + \omega\|\tilde{\boldsymbol{x}}\| = \tilde{C} + \omega\|\tilde{\boldsymbol{x}}\|$, i.e., the condition holds for $\tilde{f}(\tilde{\boldsymbol{x}})$.

Part 1. We first show that the numerator is finite for any $\tau > 0$.

Note that $\frac{d}{ds}\left(se^{-\tau s}\right) = (1 - \tau s^2)e^{-\tau s}$ and $\lim_{s\to\infty} se^{-\tau s} = 0$, we have

$$f(\boldsymbol{x})e^{-\tau f(\boldsymbol{x})} = \int_{f(\boldsymbol{x})}^{\infty} (\tau t^2 - 1)e^{-\tau t} dt.$$

By Fubini's theorem, we have

$$\int_{\boldsymbol{x}\in\mathbb{R}^p} f(\boldsymbol{x})e^{-\tau f(\boldsymbol{x})} d\boldsymbol{x} = \int_{\boldsymbol{x}\in\mathbb{R}^p} \int_{f(\boldsymbol{x})}^{\infty} (\tau t^2 - 1)e^{-\tau t} dt d\boldsymbol{x} = \int_0^{\infty} (\tau t^2 - 1)e^{-\tau t} dt \int_{\boldsymbol{x}\in\mathbb{R}^p} \mathbf{1}_{t > f(\boldsymbol{x})} d\boldsymbol{x}. \tag*{[5.20]}$$

By the condition of the lemma, $f(\boldsymbol{x}) < t$ implies $\|\boldsymbol{x}\| < \frac{t-C}{\omega}$. Thus, $\int_{\boldsymbol{x}\in\mathbb{R}^p} \mathbf{1}_{t>f(\boldsymbol{x})} d\boldsymbol{x} \leq C_p(\frac{t-C}{\omega})^p$, where the constant C_p is the volume of a p-dimensional unit ball, and Eq. (5.20) can be bounded from above by

$$\int_{\boldsymbol{x}\in\mathbb{R}^p} f(\boldsymbol{x})e^{-\tau f(\boldsymbol{x})} d\boldsymbol{x} \leq C_p \int_0^{\infty} (\frac{t-C}{\omega})^p (\tau t^2 - 1)e^{-\tau t} dt < \infty.$$

Part 2. For any $\epsilon > 0$, we split the numerator into two parts:

$$\begin{aligned}\int_{\boldsymbol{x}\in\mathbb{R}^p} f(\boldsymbol{x})e^{-\tau f(\boldsymbol{x})} d\boldsymbol{x} &\leq \int_{f(\boldsymbol{x})\leq\epsilon} f(\boldsymbol{x})e^{-\tau f(\boldsymbol{x})} d\boldsymbol{x} + \int_{f(\boldsymbol{x})>\epsilon} f(\boldsymbol{x})e^{-\tau f(\boldsymbol{x})} d\boldsymbol{x} \\ &\leq \epsilon \int_{\boldsymbol{x}\in\mathbb{R}^p} e^{-\tau f(\boldsymbol{x})} d\boldsymbol{x} + \int_{f(\boldsymbol{x})>\epsilon} f(\boldsymbol{x})e^{-\tau f(\boldsymbol{x})} d\boldsymbol{x}\end{aligned} \tag*{[5.21]}$$

Fixed any $\tau_0 > 0$, say $\tau_0 = 1$. For any $\tau > \tau_0$, if $f(\boldsymbol{x}) > \epsilon$, then

$$f(\boldsymbol{x})e^{-\tau f(\boldsymbol{x})} = f(\boldsymbol{x})e^{-\tau_0 f(\boldsymbol{x})} e^{-(\tau-\tau_0) f(\boldsymbol{x})} \leq f(\boldsymbol{x})e^{-\tau_0 f(\boldsymbol{x})} e^{-(\tau-\tau_0)\epsilon}.$$

Therefore, the second integral on the right-hand side of Eq. (5.21) can be bounded by

$$e^{-(\tau-\tau_0)\epsilon}\int_{f(\boldsymbol{x})>\epsilon} f(\boldsymbol{x})e^{-\tau_0 f(\boldsymbol{x})}d\boldsymbol{x}. \qquad [5.22]$$

We next bound the denominator from below. By the continuity at $\boldsymbol{x}_0=\mathbf{0}$, there exists $\delta_\epsilon>0$ such that for any $\|\boldsymbol{x}\|\le\delta_\epsilon$, $f(\boldsymbol{x})\le\frac{1}{2}\epsilon$. Thus,

$$\int_{\boldsymbol{x}\in\mathbb{R}^p} e^{-\tau f(\boldsymbol{x})}d\boldsymbol{x}\ge\int_{\|\boldsymbol{x}\|\le\delta_\epsilon} e^{-\tau f(\boldsymbol{x})}d\boldsymbol{x}\ge e^{-\frac{\tau}{2}\epsilon}C_p\delta_\epsilon^p. \qquad [5.23]$$

Combining Eq. (5.21), Eq. (5.22) and Eq. (5.23), we conclude that

$$\frac{\int_{\boldsymbol{x}\in\mathbb{R}^p} f(\boldsymbol{x})e^{-\tau f(\boldsymbol{x})}d\boldsymbol{x}}{\int_{\boldsymbol{x}\in\mathbb{R}^p} e^{-\tau f(\boldsymbol{x})}d\boldsymbol{x}}\le\epsilon+\frac{e^{-(\tau-\tau_0)\epsilon}\int_{f(\boldsymbol{x})>\epsilon} f(\boldsymbol{x})e^{-\tau_0 f(\boldsymbol{x})}d\boldsymbol{x}}{e^{-\frac{\tau}{2}\epsilon}C_p\delta_\epsilon^p},$$

whose right-hand side converges to $\epsilon+0$ as $\tau\to\infty$. Since $\epsilon>0$ is arbitrary, we conclude that the limit of $(\int_{\boldsymbol{x}\in\mathbb{R}^p} f(\boldsymbol{x})e^{-\tau f(\boldsymbol{x})}d\boldsymbol{x})/(\int_{\boldsymbol{x}\in\mathbb{R}^p} e^{-\tau f(\boldsymbol{x})}d\boldsymbol{x})$ is 0. □

D. Convergence from Catalytic Prior with finite *M* to Population Catalytic Prior. We will establish the convergence of the catalytic priors in this section. Recall that the total variation distance between two distributions with density π_1 and π_2 is defined as $d_{\mathrm{TV}}(\pi_1,\pi_2)=\int|\pi_1(\boldsymbol{\beta})-\pi_2(\boldsymbol{\beta})|d\boldsymbol{\beta}$. Our strategy to show the convergence in total variation is to first split the integral into $\int_{\|\boldsymbol{\beta}\|\le K}$ and $\int_{\|\boldsymbol{\beta}\|>K}$, and then obtain bounds for each in terms of K. By choosing K in an appropriate way (see Section D.2), we can obtain an upper bound on the total variation distance. Since upper bounds on the integrals of catalytic priors on $\|\boldsymbol{\beta}\|>K$ have already been established in Section A, the remaining effort is to quantify the uniform convergence of the log likelihood function on $\|\boldsymbol{\beta}\|\le K$ for a fixed K. Section D.1 focuses on this uniform convergence.

Section D.3 directly computes the KL-divergence between the catalytic prior with finite M and the population catalytic prior in the case of linear regression models and obtains an upper bound. This bound for the linear regression case is of independent interest because it holds for all $\tau>0$, unlike the ones for other GLMs.

D.1. Uniform Convergence On a Compact Set. Throughout this section, we denote

$$\ell(y,\theta)=y\theta-b(\theta).$$

The goal of this section is to find a probabilistic bound on

$$Z_K\;:=\;\sup_{\|\boldsymbol{\beta}\|\le K}\frac{1}{M}\sum_{i=1}^{M}\left(\ell(Y_i^*,(\boldsymbol{X}_i{}^*)^\top\boldsymbol{\beta})-\mathbb{E}\ell(Y_i^*,(\boldsymbol{X}_i{}^*)^\top\boldsymbol{\beta})\right). \qquad [5.24]$$

Once a bound on Z_K is obtained, the integral of the absolute difference in the two prior densities on $\|\boldsymbol{\beta}\|\le K$ can also be bounded.

Let $Z_0\;:=\;\frac{1}{M}\sum_{i=1}^{M}(\ell(Y_i^*,0)-\mathbb{E}\ell(Y_i^*,0))$, which corresponds to taking $\boldsymbol{\beta}=\mathbf{0}$ in the likelihood. We first bound the expectation of Z_K-Z_0.

Lemma 5.21. *Assume $\|\boldsymbol{X}_i^*\|^2\le V_X^2$ and the log likelihood function $\ell(y,\theta)$ is Lipschitz-L in θ, then*

$$\mathbb{E}(Z_K-Z_0)\le\frac{4KLV_X}{\sqrt{M}}.$$

The proof of this lemma is based on two classical lemmas in the literature of empirical processes and concentration of measures. They are presented below for completeness. A random variable ϵ is called *Rademacher* if $\mathbf{P}(\epsilon=1)=\mathbf{P}(\epsilon=-1)=1/2$, i.e., it is a symmetric Bernoulli r.v. on ±1.

Lemma 5.22 (Symmetrization theorem, Theorem A.2 in Ref. (15))**.** *Let $U_1, \dots, U_n$ be independent random variables with values in some space $\mathcal{U}$, and let $\epsilon_1, \dots, \epsilon_n$ be a Rademacher sequence independent of $U_1, \dots, U_n$. Let Γ be a class of real-valued functions on $\mathcal{U}$. Then*

$$\mathbb{E}\left(\sup_{\gamma\in\Gamma}\left|\sum_{i=1}^{n}\{\gamma(U_i) - \mathbb{E}\gamma(U_i)\}\right|\right) \le 2\mathbb{E}\left(\sup_{\gamma\in\Gamma}\left|\sum_{i=1}^{n}\epsilon_i\gamma(U_i)\right|\right) \qquad [5.25]$$

Lemma 5.23 (Contraction theorem, Theorem A.3 in Ref. (15))**.** *Let $x_1, \dots, x_n$ be nonrandom elements in some space $\mathcal{X}$, and $\mathcal{F}$ is a class of real-valued functions on $\mathcal{X}$. Consider Lipschitz function $\gamma_i : \mathbb{R} \mapsto \mathbb{R}$, that is,*

$$|\gamma_i(s) - \gamma_i(\tilde{s})| \le |s - \tilde{s}|, \quad \forall s, \tilde{s} \in \mathbb{R}. \qquad [5.26]$$

Let $\epsilon_1, \dots, \epsilon_n$ be a Rademacher sequence. Then for any function $f_0 : \mathcal{X} \mapsto \mathbb{R}$, we have

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i\{\gamma_i(f(x_i)) - \gamma_i(f_0(x_i))\}\right|\right) \le 2\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i\{f(x_i) - f_0(x_i)\}\right|\right). \qquad [5.27]$$

Proof of Lemma 5.21. Let $\epsilon_1, \dots, \epsilon_M$ be a Rademacher sequence independent with all the synthetic data. We have

$$\begin{aligned}
&\mathbb{E}(Z_K - Z_0)\\
\le\ &\mathbb{E}\sup_{\|\boldsymbol{\beta}\|\le K}|\frac{1}{M}\sum_{i=1}^{M}\left(\ell(Y_i^*, (\boldsymbol{X}_i{}^*)^\top\boldsymbol{\beta}) - \mathbb{E}\ell(Y_i^*, (\boldsymbol{X}_i{}^*)^\top\boldsymbol{\beta}) - [\ell(Y_i^*, 0) - \mathbb{E}\ell(Y_i^*, 0)]\right)|\\
\le\ &2\mathbb{E}\sup_{\|\boldsymbol{\beta}\|\le K}|\frac{1}{M}\sum_{i=1}^{M}\epsilon_i\left(\ell(Y_i^*, (\boldsymbol{X}_i{}^*)^\top\boldsymbol{\beta}) - \ell(Y_i^*, 0)\right)| \qquad \text{(by symmetrization, Lemma 5.22)}\\
\le\ &4L\mathbb{E}\sup_{\|\boldsymbol{\beta}\|\le K}|\frac{1}{M}\sum_{i=1}^{M}\epsilon_i\left((\boldsymbol{X}_i{}^*)^\top\boldsymbol{\beta} - 0\right)| \qquad \text{(by contraction principle, Lemma 5.23)}\\
\le\ &4L\mathbb{E}\sup_{\|\boldsymbol{\beta}\|\le K}\|\frac{1}{M}\sum_{i=1}^{M}\epsilon_i\boldsymbol{X}_i^*\|\|\boldsymbol{\beta}\|\\
\le\ &4LK\frac{1}{M}\sqrt{\mathbb{E}\|\sum_{i=1}^{M}\epsilon_i\boldsymbol{X}_i^*\|^2} \le \frac{4KLV_X}{\sqrt{M}},
\end{aligned}$$

with the following reasons:

(1) the second inequality: apply Lemma 5.22 with $U_i = (Y_i^*, \boldsymbol{X}_i^*)$ and $\Gamma = \{\gamma_{\boldsymbol{\beta}}(U_i) = \ell(Y_i^*, (\boldsymbol{X}_i{}^*)^\top\boldsymbol{\beta}) : \|\boldsymbol{\beta}\| \le K\}$;

(2) the third inequality: we first condition on the synthetic data, and then apply Lemma 5.23 with $x_i = \boldsymbol{X}_i^*$, $\mathcal{F} = \{f_{\boldsymbol{\beta}}(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{\beta} : \|\boldsymbol{\beta}\| \le K\}$ and $\gamma_i(s) = \ell(Y_i^*, s)$;

(3) the fourth and fifth inequalities are due to the Cauchy-Schwarz inequality, and the last inequality is due to the independence between ϵ_i and $\boldsymbol{X}_i^*$.

□

Next we adapt a theorem from Ref. (16) to bound the deviation of $Z_K - Z_0$ from its expectation.

Lemma 5.24. *Assume $\|\boldsymbol{X}_i^*\|^2 \le V_X^2$ and the log likelihood function $\ell(y, \theta)$ is Lipschitz-L in θ for $|\theta| \le KV_X$, then*

$$\mathbf{P}\left(Z_K - Z_0 - \mathbb{E}(Z_K - Z_0) \ge \frac{8LKV_X}{\sqrt{M}}s\right) \le e^{-s\min(s,\sqrt{M})} \qquad [5.28]$$

Proof. Let $W_{i,\boldsymbol{\beta}} = \frac{1}{2LKV_X}\left(\ell(Y_i^*, (\boldsymbol{X}_i{}^*)^\top\boldsymbol{\beta}) - \mathbb{E}\ell(Y_i^*, (\boldsymbol{X}_i{}^*)^\top\boldsymbol{\beta}) - [\ell(Y_i^*, 0) - \mathbb{E}\ell(Y_i^*, 0)]\right)$. Then $\mathbb{E}W_{i,\boldsymbol{\beta}} = 0$ and

$$\sup_{\|\boldsymbol{\beta}\|\le K} |W_{i,\boldsymbol{\beta}}| \le 1,$$

because $\ell(y, \theta)$ is Lipschitz-L in θ for $|\theta| \le KV_X$ and $\sup_{\|\boldsymbol{\beta}\|\le K} |(\boldsymbol{X}_i{}^*)^\top\boldsymbol{\beta}| \le KV_X$.
Let $\Sigma^2 := \mathbb{E}\sup_{\|\boldsymbol{\beta}\|\le K}\sum_{i=1}^M W_{i,\boldsymbol{\beta}}^2$ and $\sigma^2 := \sup_{\|\boldsymbol{\beta}\|\le K}\sum_{i=1}^M \mathbb{E}W_{i,\boldsymbol{\beta}}^2$. Clearly $\Sigma^2 \le M$ and $\sigma^2 \le M$. By Theorem 12.2 in Ref. (16), we have

$$\mathbf{P}\left(\sup_{\|\boldsymbol{\beta}\|\le K}\sum_{i=1}^M W_{i,\boldsymbol{\beta}} - \mathbb{E}\sup_{\|\boldsymbol{\beta}\|\le K}\sum_{i=1}^M W_{i,\boldsymbol{\beta}} \ge t\right) \le \exp\left(-\frac{t^2}{2t+8M}\right),$$

which directly implies that

$$\mathbf{P}\left(Z_K - Z_0 - \mathbb{E}(Z_K - Z_0) \ge \frac{2LKV_X}{M}t\right) \le \exp\left(-\min(\frac{t}{4}, \frac{t^2}{16M})\right).$$

Setting $t = 4s\sqrt{M}$, we obtain the result. □

Combining Lemma 5.21 and Lemma 5.24 together, we have the following theorem.

Theorem 5.25. *Assume $\|\boldsymbol{X}_i^*\|^2 \le V_X^2$ and the log likelihood function $\ell(y, \theta)$ is Lipschitz-L in θ for $|\theta| \le KV_X$, then*

$$\mathbf{P}\left(Z_K \ge \frac{12LKV_X}{\sqrt{M}}s\right) \le e^{-s\min(s,\sqrt{M})} \qquad [5.29]$$

Proof. Note that $\ell(y, 0) = y \cdot 0 - b(0) = -b(0)$, so by definition $Z_0 := \frac{1}{M}\sum_{i=1}^M(\ell(Y_i^*, 0) - \mathbb{E}\ell(Y_i^*, 0)) = 0$. We have

$$\begin{aligned}\mathbf{P}\left(Z_K \ge \frac{12LKV_X}{\sqrt{M}}s\right) &= \mathbf{P}\left(Z_K - Z_0 \ge \frac{12LKV_X}{\sqrt{M}}s\right)\\ &\le \mathbf{P}\left(Z_K - Z_0 \ge \mathbb{E}(Z_K - Z_0) + \frac{8LKV_X}{\sqrt{M}}s\right)\\ &\le e^{-s\min(s,\sqrt{M})},\end{aligned}$$

where the first inequality is due to Lemma 5.21, and the second inequality is due to Lemma 5.24. □

D.2. Convergence in Total Variation. We begin with an elementary lemma, which formalizes the key steps to show the convergence in total variation.

Lemma 5.26. *For two measurable functions f, g on $\mathbb{R}^p$ with integrable exponents, let $I_f := \int_{\boldsymbol{\beta}\in\mathbb{R}^p} e^f d\boldsymbol{\beta} \in (0, \infty)$ and $I_g := \int_{\boldsymbol{\beta}\in\mathbb{R}^p} e^g d\boldsymbol{\beta} \in (0, \infty)$. Suppose ϵ_1, ϵ_2 are finite positive numbers such that*

1. $\sup_{\|\boldsymbol{\beta}\|\le K} |f(\boldsymbol{\beta}) - g(\boldsymbol{\beta})| \le \epsilon_1$.
2. $\int_{\|\boldsymbol{\beta}\|>K} e^f d\boldsymbol{\beta} \le \epsilon_2, \quad \int_{\|\boldsymbol{\beta}\|>K} e^g d\boldsymbol{\beta} \le \epsilon_2$

Then

$$\int_{\boldsymbol{\beta}\in\mathbb{R}^p} |\frac{e^f}{I_f} - \frac{e^g}{I_g}| d\boldsymbol{\beta} \le 2(e^{\epsilon_1} - 1) + \frac{3\epsilon_2}{I_f}$$

Proof. This proof is elementary.

$$\int_{\boldsymbol{\beta}\in\mathbb{R}^p} |\frac{e^f}{I_f} - \frac{e^g}{I_g}| d\boldsymbol{\beta} \le \int_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{|e^f - e^g|}{I_f}|d\boldsymbol{\beta} + \int_{\boldsymbol{\beta}\in\mathbb{R}^p} e^g|\frac{1}{I_f} - \frac{1}{I_g}|d\boldsymbol{\beta} = \frac{1}{I_f}\int_{\boldsymbol{\beta}\in\mathbb{R}^p} |e^f - e^g| d\boldsymbol{\beta} + \frac{1}{I_f}|I_f - I_g|. \qquad [5.30]$$

Note that on $\|\boldsymbol{\beta}\| \le K$, $|e^f - e^g| \le (e^{\epsilon_1} - 1)e^f$. It follows that

$$I_f - I_g = \int_{\|\boldsymbol{\beta}\le K}(e^f - e^g)d\boldsymbol{\beta} + \int_{\|\boldsymbol{\beta}>K} e^f d\boldsymbol{\beta} - \int_{\|\boldsymbol{\beta}>K} e^g d\boldsymbol{\beta} \le (e^{\epsilon_1} - 1)I_f + \epsilon_2.$$

Similarly we have $I_g - I_f \leq (e^{\epsilon_1} - 1)I_f + \epsilon_2$. Thus, $|I_g - I_f| \leq (e^{\epsilon_1} - 1)I_f + \epsilon_2$. Therefore, we can bound the right-hand side of Eq. (5.30) by

$$\frac{1}{I_f}\left(2\epsilon_2 + \int_{\|\boldsymbol{\beta}\|\leq K} e^f (e^{\epsilon_1} - 1)d\boldsymbol{\beta} + (e^{\epsilon_1} - 1)I_f + \epsilon_2\right) \leq 2(e^{\epsilon_1} - 1) + \frac{3\epsilon_2}{I_f}.$$

□

We now use this lemma together with Theorems 5.17 and 5.16 to prove the convergence in total variation.

Theorem 5.27. *Suppose there exists a compact subset $\mathcal{Y}^{com}$ of $\mathcal{Y}$ such that every synthetic response is in $\mathcal{Y}^{com}$ with probability* 1*, and the synthetic covariate sampled satisfies Condition 5.5 and $c_\phi := \inf_{\eta\neq 0} |\phi(\eta)/\eta| > 0$. Suppose also the log likelihood function $\ell(y, \theta)$ is Lipschitz-L in θ. Then there exist constants C_* and δ only depending on $\mathcal{Y}^{com}$ and the exponential family, and constants ρ_0, η_0, c and C that only depend on the constant B_1 in Condition 5.5, such that for any $\epsilon \in (0, 1)$, and $\nu \in (0, 1)$, for any $M > M_0 := \max\left(\frac{24}{\epsilon} L\sqrt{p}B_1 K_+, 1\right)^2 \log(1/\nu)$, where*

$$K_+ = \frac{8}{\tau\delta c_\phi \eta_0\rho_0} \max\left(\log(C_{Stirling}C^\tau) - \log(I_{cat,\infty}\epsilon) + p\log(\frac{\sqrt{128\pi p/e}}{\tau\delta c_\phi\eta_0\rho_0}), \frac{p}{2}\right), \tag{5.31}$$

we have that the total variation distance $d_{TV}(\pi_{cat,M}, \pi_{cat,\infty}) \leq 5\epsilon$ with probability at lest

$$1 - \nu - e^{-cM} - exp\left(-\frac{M\rho_0^2}{2} + p\log(1 + \frac{8C}{\eta_0\rho_0})\right). \tag{5.32}$$

Remark 5.28. Based on this result, we can further study the rate of convergence. We will use the asymptotic notation $a_n \lesssim b_n$ (and $a_n \gtrsim b_n$) to indicate $\sup_n \frac{a_n}{b_n} < \infty$ (and $\inf_n \frac{a_n}{b_n} > 0$) for any positive sequences a_n, b_n. We also use $a_n \asymp b_n$ if $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold simultaneously. If we ignore all constants (such as $\tau, \delta, c_\phi, \eta_0, \rho_0, C$) that do not depend on p or M, then

$$\begin{aligned} K_+ &\asymp p\log p + \log(\frac{1}{I_{cat,\infty}\epsilon}) \\ M_0 &\asymp p(\frac{p\log p + \log(\frac{1}{I_{cat,\infty}\epsilon})}{\epsilon})^2 \log(1/\nu). \end{aligned} \tag{5.33}$$

We can assume $\log(1/\nu) > 1$ and $\log(1/\epsilon) > 1$, since only small values of ν and ϵ matter. Together with $p > 1$, Eq. (5.33) implies $\sqrt{M_0} \gtrsim \frac{1}{\epsilon}$. Therefore $\log(\frac{1}{\epsilon}) \lesssim \log M_0$. Plugging in Eq. (5.33), we have

$$\epsilon \lesssim (p\log p + \log\left(\frac{M_0}{I_{cat,\infty}}\right))\sqrt{\frac{p\log(1/\nu)}{M_0}}. \tag{5.34}$$

This analysis suggests that the total variation $d_{TV}(\pi_{cat,M}, \pi_{cat,\infty})$ decays roughly at the rate of $O(\sqrt{\frac{p^3\log^2(p) + p\log^2(M)}{M}})$.

□

Remark 5.29. For fixed p and τ, one can take C_1 sufficiently large such that for any $\nu \in (0, 1/6)$ and $\epsilon \in (0, 1)$, the following inequalities hold

$$\frac{2}{\rho_0^2}\left(\log(1/\nu) + p\log(1 + \frac{8C}{\eta_0\rho_0})\right) < C_1\log(\frac{1}{3\nu}), \tag{5.35}$$

$$\frac{1}{c}\log(1/\nu) < C_1\log(\frac{1}{3\nu}), \tag{5.36}$$

$$\left(\frac{24}{\epsilon}L\sqrt{p}B_1K_+\right)^2 < C_1\frac{1}{(5\epsilon)^2}\left(1 + \log^2(\frac{1}{5\epsilon})\right), \tag{5.37}$$

$$1 < C_1\frac{1}{(5\epsilon)^2}\left(1 + \log^2(\frac{1}{5\epsilon})\right). \tag{5.38}$$

For any $\epsilon_0 \in (0, 1/2)$ and $\epsilon_1 \in (0, 5)$, if $M \geq C_1 \left(1 + \log^2(\frac{1}{\epsilon_1})\right) \frac{1}{\epsilon_1^2} \log(\frac{1}{\epsilon_0})$, substituting $\nu = \epsilon_0/3$ and $\epsilon = \epsilon_1/5$ in the above inequalities (Eq. (5.35) to Eq. (5.38)), we have

$$M > \frac{2}{\rho_0^2}\left(\log(1/\nu) + p\log(1 + \frac{8C}{\eta_0\rho_0})\right), \quad M > \frac{\log(1/\nu)}{c}, \tag{5.39}$$

and

$$M > \max\left(\frac{24}{\epsilon} L\sqrt{p} B_1 K_+, 1\right)^2 \log(1/\nu). \tag{5.40}$$

Using Eq. (5.39), the probability in Eq. (5.32) is bounded from below by $1 - 3\nu = 1 - \epsilon_0$. Therefore, Theorem 5.27, together with Eq. (5.40), implies that with probability at least $1 - \epsilon_0$, the total variation distance $d_{TV}(\pi_{cat,M}, \pi_{cat,\infty})$ is bounded from above by $5\epsilon = \epsilon_1$. This result corresponds to the first statement in Theorem 4 of the main text. □

Remark 5.30. The assumption on $\mathcal{Y}^{\text{com}}$ is automatically satisfied if the synthetic-data generating model is a sub-model, say $g_*(\cdot \mid \boldsymbol{X}^*, \boldsymbol{Y}, \mathbb{X}) = f(\cdot \mid \hat{\boldsymbol{\beta}}^\top \boldsymbol{X}^*)$, and the synthetic response is replaced by the predictive mean under the synthetic-data generating model. The reason is the following. Under Condition 5.5, $\|\boldsymbol{X}^*\|$ is bounded and $\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}^*$ lies in some compact set $\Theta \subset \mathbb{R}$. This implies that every synthetic response will lie in the image $b'(\Theta)$, which is also compact because $b'(\cdot)$ is continuous. □

Proof of Theorem 5.27. Denote by $\pi^U_{cat,M}$ and $\pi^U_{cat,\infty}$ respectively the unnormalized density functions of the finite-M catalytic prior and the population catalytic prior, and denote by $I_{cat,M}$ and $I_{cat,\infty}$ respectively the integrals of $\pi^U_{cat,M}$ and $\pi^U_{cat,\infty}$.

We directly apply Theorem 5.16 to get all the constants and all the inequalities. Denote by A_1 the event that (a)-(c) in Theorem 5.16 hold. On this event, for any $K > 4p/(\tau\delta c_\phi \eta_0\rho_0)$, we have

$$\begin{aligned}\int_{\|\boldsymbol{\beta}\|>K} \pi^U_{cat,M} d\boldsymbol{\beta} &\leq C_{Stirling} C_*^\tau \exp(p - \tau c_\phi \eta_0\rho_0\delta K/4)(\frac{\sqrt{2\pi}K}{\sqrt{pe}})^p \\ &= C_{Stirling} C_*^\tau (\frac{2\pi}{pe})^{p/2} e^p \left[K^p \exp(-2\tau C_2 K)\right],\end{aligned}$$

where $C_2 = c_\phi\eta_0\rho_0\delta/8$. Using the elementary fact that for all a, b and $x > 0$, it always holds that $a\log x - bx \leq -bx/2 + a\log(2a/b) - a$, we know that the above right-hand-side is bounded by (taking $x = K$, $a = p$ and $b = 2\tau C_2$)

$$\int_{\|\boldsymbol{\beta}\|>K} \pi^U_{cat,M} d\boldsymbol{\beta} \leq C_{Stirling} C_*^\tau (\frac{2\pi p}{e})^{p/2} \frac{\exp(-\tau C_2 K)}{(\tau C_2)^p}. \tag{5.41}$$

A similar argument using (a)-(c) in Theorem 5.17 shows that the right-hand side of Eq. (5.41) is also an upper bound for $\int_{\|\boldsymbol{\beta}\|>K} \pi^U_{cat,\infty} d\boldsymbol{\beta}$.

For any $\epsilon \in (0, 1)$, let $K_0 = \frac{1}{\tau C_2}\left(\log(C_{Stirling} C^\tau) + \log(1/(I_{cat,\infty}\epsilon)) + p\log(\frac{\sqrt{2\pi p/e}}{\tau C_2})\right)$. Let

$$K_+ = \max(K_0, 4p/(\tau\delta c_\phi\eta_0\rho_0)).$$

It is straightforward to check that if $K \geq K_+$, then the right-hand side of Eq. (5.41) is no greater than $I_{cat,\infty}\epsilon$.

Let V_X be $\sqrt{p}B_1$. Condition 5.5 implies $\|\boldsymbol{X}_i^*\|^2 \leq pB_1^2 = V_X^2$. By Theorem 5.25, for any $\nu > 0$ and any $M \geq \log(1/\nu)$, it holds with probability at least $1 - \nu$ that

$$\sup_{\|\boldsymbol{\beta}\| \leq K_+} |\log(\pi^U_{cat,M}) - \log(\pi^U_{cat,\infty})| \leq \frac{12LK_+V_X}{\sqrt{M}}\sqrt{\log(1/\nu)}. \tag{5.42}$$

Theorem 5.16 and Theorem 5.25 show that with probability at least

$$1 - \nu - e^{-cM} - \exp\left(-\frac{M\rho_0^2}{2} + p\log(1 + \frac{8C}{\eta_0\rho_0})\right),$$

both events A_1 and Eq. (5.42) hold, in which case Lemma 5.26 implies that

$$d_{TV}(\pi_{cat,M},\pi_{cat,\infty}) \le 2(\exp\left(\frac{12LK_+V_X}{\sqrt{M}}\sqrt{\log(1/\nu)}\right)-1)+\frac{3}{I_{cat,\infty}}I_{cat,\infty}\epsilon.$$

Furthermore, if $M > (\frac{24LV_XK_+}{\epsilon})^2\log(1/\nu)$, then using the fact that $e^x \le 1+2x$ for $x\in(0,1)$,

$$d_{TV}(\pi_{cat,M},\pi_{cat,\infty}) \le 2(\exp(\epsilon/2)-1)+3\epsilon \le 5\epsilon.$$

□

D.3. Convergence of Catalytic Priors for Linear Regression Models. Theorem 5.27 requires the technical Lipschitz condition of the likelihood, which does not apply to linear regression models. In this section, we study the convergence of catalytic priors for linear regression with Gaussian noise. It is interesting to note that the catalytic prior for linear regression has a convergence rate that does not depend on τ. This also guarantees the use of small value of τ for linear regression, which may not be suitable for other GLMs.

In this section, we use the same notation as in Section 1, where we have shown that for a linear regression model, the catalytic prior $\pi_{cat,M}$ is $N(\tilde{\boldsymbol{\beta}}_0,\sigma^2(\frac{\tau}{M}(\mathbb{X}^*)^\top\mathbb{X}^*)^{-1})$, and the population catalytic prior $\pi_{cat,\infty}$ is $N(\tilde{\boldsymbol{\beta}}_0,\sigma^2(\tau\Sigma_X)^{-1})$, where $\tilde{\boldsymbol{\beta}}_0$ is the estimated parameter under the synthetic-data generating distribution (and the predictive mean under the synthetic-data generating distribution is $\mathbb{E}_{g^*}(Y^*|\boldsymbol{X}^*=\boldsymbol{x})=\boldsymbol{x}^\top\tilde{\boldsymbol{\beta}}_0$).

Theorem 5.31. *Suppose $\boldsymbol{X}_1^*,\ldots,\boldsymbol{X}_M^*$ are i.i.d. drawn from the independent resampling distribution, and $\|\boldsymbol{X}_i^*\|_2 \le V_X$. Let σ_X^2 be $\min_{2\le j\le p}\widehat{\sigma_{X,j}^2}$ (which is positive by assumption). Then for any $\delta>0$ and any $M>\frac{16}{9}(\frac{V_X}{\sigma_X})^2\log(\frac{p}{\delta})$, it holds with probability at least $1-\delta$ that*

$$KL(\pi_{cat,\infty},\pi_{cat,M}) \le 2p\frac{V_X}{\sigma_X}\sqrt{\frac{1}{M}\log(\frac{p}{\delta})}. \qquad [5.43]$$

Remark 5.32. Condition 5.5 implies $\|\boldsymbol{X}_i^*\|^2 \le pB_1^2$, that is, V_X can be taken as $\sqrt{p}B_1$. It follows that the KL-divergence decays at the rate of $O(\sqrt{\frac{p^3\log p}{M}})$. This rate is slightly faster than the rate in Theorem 5.27, which is $O(\sqrt{\frac{p^3\log^2(p)+p\log^2(M)}{M}})$. See Remark 5.28. □

Proof of Theorem 5.31. Let $\boldsymbol{D}_X$ be the diagonal matrix $\mathrm{diag}(1,\widehat{\sigma_{X,2}^2},\ldots,\widehat{\sigma_{X,p}^2})$. Under the independent resampling distribution, the limiting covariance matrix $\Sigma_{\boldsymbol{X}}=\lim_{M\to\infty}\frac{1}{M}(\mathbb{X}^*)^\top\mathbb{X}^*$ is the diagonal matrix $\boldsymbol{D}_X$, i.e., $\Sigma_{\boldsymbol{X}}=\boldsymbol{D}_X$. We first prove that for any $t>0$ and any M, with probability at least $1-p\cdot\exp\left(-\frac{t^2}{2(1+\frac{2t}{3\sqrt{M}}\frac{V_X}{\sigma_X})}\right)$,

$$\|\frac{1}{M}\boldsymbol{D}_X^{-1/2}(\mathbb{X}^*)^\top\mathbb{X}^*\boldsymbol{D}_X^{-1/2}-\boldsymbol{I}_p\| \le \frac{V_X}{\sqrt{M}\sigma_X}t. \qquad [5.44]$$

Define $\boldsymbol{U}_i=\boldsymbol{D}_X^{-1/2}\boldsymbol{X}_i^*$. Under the conditions of the theorem, we have

$$\|\boldsymbol{U}_i\| \le \frac{V_X}{\sigma_X}, \text{ and } \mathbb{E}\boldsymbol{U}_i\boldsymbol{U}_i^\top=\boldsymbol{I}_p \qquad [5.45]$$

Let $\boldsymbol{S}_i=\frac{1}{M}(\boldsymbol{U}_i\boldsymbol{U}_i^\top-\boldsymbol{I}_p)$ and $\boldsymbol{\Delta}=\frac{1}{M}\boldsymbol{D}_X^{-1/2}(\mathbb{X}^*)^\top\mathbb{X}^*\boldsymbol{D}_X^{-1/2}-\boldsymbol{I}_p$. Then $\boldsymbol{\Delta}=\sum_{i=1}^M\boldsymbol{S}_i$.

Note that $\sigma_X^2=\min_j\widehat{\sigma_{X,j}^2}\le V_X^2$. Eq. (5.45) implies

$$\|\boldsymbol{S}_i\| \le \frac{1+V_X^2/\sigma_X^2}{M} \le \frac{2V_X^2/\sigma_X^2}{M}.$$

We use the expression $\boldsymbol{A} \preceq \boldsymbol{B}$ to indicate that $\boldsymbol{B} - \boldsymbol{A}$ is positive semi-definite.

$$\begin{aligned}\mathbb{E}\sum_{i=1}^{M} \boldsymbol{S}_i^2 &= \frac{1}{M^2}\sum_{i=1}^{M}\left(\mathbb{E}\boldsymbol{U}_i\boldsymbol{U}_i^\top\boldsymbol{U}_i\boldsymbol{U}_i^\top - \boldsymbol{I}_p\right)\\ &\preceq \frac{1}{M^2}\sum_{i=1}^{M}\left((\frac{V_X}{\sigma_X})^2\mathbb{E}\boldsymbol{U}_i\boldsymbol{U}_i^\top - \boldsymbol{I}_p\right)\\ &\preceq \frac{(V_X/\sigma_X)^2}{M}\boldsymbol{I}_p,\end{aligned}$$

where the first equality and the third inequality uses $\mathbb{E}\boldsymbol{U}_i\boldsymbol{U}_i^\top = \boldsymbol{I}_p$, and the second inequality is due to $\boldsymbol{U}_i^\top\boldsymbol{U}_i = \|\boldsymbol{U}_i\|^2 \leq (\frac{V_X}{\sigma_X})^2$. Applying the Matrix Bernstein inequality (Theorem 1.4 in Ref. (17)), we have that for all $s \geq 0$,

$$\mathbf{P}(\|\boldsymbol{\Delta}\| \geq s) \leq p \cdot \exp\left(\frac{-s^2/2}{\frac{(V_X/\sigma_X)^2}{M} + \frac{2s(V_X/\sigma_X)^2}{3M}}\right).$$

Substitute $s = t(V_X/\sigma_X)/\sqrt{M}$ into the last right-hand side, we conclude

$$\mathbf{P}(\|\boldsymbol{\Delta}\| \geq \frac{tV_X/\sigma_X}{\sqrt{M}}) \leq p \cdot \exp\left(-\frac{t^2}{2(1+\frac{2t}{3\sqrt{M}}\frac{V_X}{\sigma_X})}\right), \qquad [5.46]$$

which is Eq. (5.44).

Now we compute the KL-divergence between the catalytic prior $\pi_{cat,\infty} \sim N(\tilde{\boldsymbol{\beta}}_0, \sigma^2(\tau(\mathbb{X}^*)^\top\mathbb{X}^*/M)^{-1})$ and the population catalytic prior $\pi_{cat,M} \sim N(\tilde{\boldsymbol{\beta}}_0, \sigma^2(\tau\boldsymbol{D}_X)^{-1})$.

By the definition, the KL-divergence between two multivariate normal distributions $\mathbb{N}(\mu_1, \boldsymbol{\Omega}_1^{-1})$ and $\mathbb{N}(\mu_2, \boldsymbol{\Omega}_2^{-1})$ is

$$\frac{1}{2}\left(\mathrm{Tr}(\boldsymbol{\Omega}_1^{-1}\boldsymbol{\Omega}_2) + \log\det(\boldsymbol{\Omega}_1\boldsymbol{\Omega}_2^{-1}) - p + \mathrm{Tr}(\boldsymbol{\Omega}_2(\mu_2-\mu_1)(\mu_2-\mu_1)^\top\right).$$

Take $\boldsymbol{\Omega}_1 = \tau(\mathbb{X}^*)^\top\mathbb{X}^*/(M\sigma^2)$ and $\boldsymbol{\Omega}_2 = \tau\boldsymbol{D}_X/\sigma^2$. Then $\boldsymbol{\Delta} = \boldsymbol{\Omega}_2^{-1/2}\boldsymbol{\Omega}_1\boldsymbol{\Omega}_2^{-1/2} - \boldsymbol{I}_p$ and

$$\mathrm{KL}(\pi_{cat,\infty}, \pi_{cat,M}) = \frac{1}{2}\left(\mathrm{Tr}(\boldsymbol{\Omega}_1^{-1}\boldsymbol{\Omega}_2) + \log\det(\boldsymbol{\Omega}_1\boldsymbol{\Omega}_2^{-1}) - p\right).$$

By the cyclic property of matrix trace, we have

$$\mathrm{Tr}(\boldsymbol{\Omega}_1^{-1}\boldsymbol{\Omega}_2) - p = \mathrm{Tr}(\boldsymbol{\Omega}_1^{-1}(\boldsymbol{\Omega}_2 - \boldsymbol{\Omega}_1)) = -\mathrm{Tr}(\boldsymbol{\Omega}_2^{1/2}\boldsymbol{\Omega}_1^{-1}\boldsymbol{\Omega}_2^{1/2}\boldsymbol{\Delta}) = -\mathrm{Tr}((\boldsymbol{I}_p + \boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}).$$

Let the eigenvalue decomposition of $\boldsymbol{\Delta}$ be $V\Lambda V^\top$, where Λ is a diagonal matrix whose diagonal entries are $\lambda_1, \ldots, \lambda_p$ and V is an orthonormal matrix. By $V^\top V = \boldsymbol{I}_p$, we have

$$\begin{aligned}\mathrm{Tr}(\boldsymbol{\Omega}_1^{-1}\boldsymbol{\Omega}_2) - p + \log\det(\boldsymbol{\Omega}_1\boldsymbol{\Omega}_2^{-1}) &= \log\det(\boldsymbol{I}_p + \boldsymbol{\Delta}) - \mathrm{Tr}((\boldsymbol{I}_p + \boldsymbol{\Delta})^{-1}\boldsymbol{\Delta})\\ &= \sum_{i=1}^{p}\left(\log(1+\lambda_i) - \frac{\lambda_i}{1+\lambda_i}\right).\end{aligned}$$

Note that the inequality $\log(1+\lambda) - \lambda/(1+\lambda) \leq 2|\lambda|$ holds for any $|\lambda| < 1/2$, so on the event $\|\boldsymbol{\Delta}\| < 1/2$, we have

$$\mathrm{Tr}(\boldsymbol{\Omega}_1^{-1}\boldsymbol{\Omega}_2) - p + \log\det(\boldsymbol{\Omega}_1\boldsymbol{\Omega}_2^{-1}) \leq 2\sum_{i=1}^{p}|\lambda_i| \leq 2p\|\boldsymbol{\Delta}\|,$$

which means that

$$\mathrm{KL}(\pi_{cat,\infty}, \pi_{cat,M}) \leq p\|\boldsymbol{\Delta}\|.$$

For any $\delta > 0$, if

$$M > \frac{16}{9}(\frac{V_X}{\sigma_X})^2\log(\frac{p}{\delta}),$$

then $\frac{2(V_X/\sigma_X)\sqrt{4\log(\frac{p}{\delta})}}{3\sqrt{M}} < 1$, and we can take $t = \sqrt{4\log(\frac{p}{\delta})}$ in Eq. (5.46) to conclude that

$$\mathbf{P}\left(\|\mathbf{\Delta}\| \geq (V_X/\sigma_X)\sqrt{\frac{4\log(\frac{p}{\delta})}{M}}\right) \leq \delta,$$

and

$$\mathbf{P}\left(\mathrm{KL}(\pi_{cat,\infty}, \pi_{cat,M}) \geq p(V_X/\sigma_X)\sqrt{\frac{4\log(\frac{p}{\delta})}{M}}\right) \leq \delta.$$

This concludes the proof of the theorem.

□

Remark 5.33. The last inequality does not depend on the value of τ, which implies that one can use a small value of τ in the catalytic prior for linear regression models. □

Remark 5.34. As a special case, when $V_X = 1$ and $\sigma_X = 1$, for any $\epsilon > 0$, if $M \geq \frac{4p^3}{3\epsilon^2}\log(\frac{p}{\delta})$ then with probability at least $1-\delta$, it holds that $\mathrm{KL}(\pi_{cat,\infty}, \pi_{cat,M}) \leq \epsilon$.

□

References

1. Efron B (2004) The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association* 99(467):619–632.
2. Efron B (1986) How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* 81(394):461–470.
3. Efron B, Hastie T (2016) *Computer age statistical inference.* (Cambridge University Press) Vol. 5.
4. Varoquaux G (2018) Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage* 180:68–77.
5. Ibrahim JG, Chen MH (2000) Power prior distributions for regression models. *Statistical Science* 15(1):46–60.
6. 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.
7. Swets JA (1988) Measuring the accuracy of diagnostic systems. *Science* 240(4857):1285–1293.
8. Simpson D, et al. (2017) Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science* 32(1):1–28.
9. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1):267–288.
10. Efron B (2016) Empirical Bayes deconvolution estimates. *Biometrika* 103(1):1–20.
11. Diaconis P, Ylvisaker D (1979) Conjugate priors for exponential families. *The Annals of Statistics* 7(2):269–281.
12. Litvak A, Rivasplata O (2012) Smallest singular value of sparse random matrices. *Studia Mathematica* 3(212):195–218.
13. Rudelson M, Vershynin R (2009) Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 62(12):1707–1739.
14. 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.
15. Van de Geer SA (2008) High-dimensional generalized linear models and the lasso. *The Annals of Statistics* 36(2):614–645.
16. Boucheron S, Lugosi G, Massart P (2013) *Concentration inequalities: A nonasymptotic theory of independence.* (Oxford university press).
17. Tropp JA (2012) User-friendly tail bounds for sums of random matrices. *Foundations of Computational mathematics* 12(4):389–434.