Received: 14 December 2020 | Accepted: 13 July 2021

DOI: 10.1111/rssb.12466

ROYAL STATISTICAL SOCIETY
Series B Statistical Methodology

ORIGINAL ARTICLE

Approximate Laplace approximations for scalable model selection

David Rossell[1,2] | **Oriol Abril**[1] | **Anirban Bhattacharya**[3]

[1]Universitat Pompeu Fabra, Barcelona, Spain

[2]Data Science Center, Barcelona Graduate School of Economics, Barcelona, Spain

[3]Statistics, Texas A&M Univesrsity, Texas, USA

Correspondence
David Rossell, Universitat Pompeu Fabra, Barcelona, Spain.
Email: rosselldavid@gmail.com

Funding information
Spanish Government grants Europa Excelencia, Grant/Award Number: EUR2020-112096, RYC-2015-18544 and PGC2018-101643-B-I00; NIH, Grant/Award Number: R01 CA158113DMS-01

Abstract

We propose the approximate Laplace approximation (ALA) to evaluate integrated likelihoods, a bottleneck in Bayesian model selection. The Laplace approximation (LA) is a popular tool that speeds up such computation and equips strong model selection properties. However, when the sample size is large or one considers many models the cost of the required optimizations becomes impractical. ALA reduces the cost to that of solving a least-squares problem for each model. Further, it enables efficient computation across models such as sharing pre-computed sufficient statistics and certain operations in matrix decompositions. We prove that in generalized (possibly non-linear) models ALA achieves a strong form of model selection consistency for a suitably-defined optimal model, at the same functional rates as exact computation. We consider fixed- and high-dimensional problems, group and hierarchical constraints, and the possibility that all models are misspecified. We also obtain ALA rates for Gaussian regression under non-local priors, an important example where the LA can be costly and does not consistently estimate the integrated likelihood. Our examples

include non-linear regression, logistic, Poisson and survival models. We implement the methodology in the R package `mombf`.

KEYWORDS

approximate inference, hierarchical constraints, group constraints, model misspecification, model selection, non-local priors, non-parametric regression

A main computational bottleneck in Bayesian model selection is evaluating integrated likelihoods, either when the sample size n is large or there are many models to consider. If said integrals can be obtained quickly, one can often use relatively simple algorithms to explore effectively the model space. For example, one may rely on the fast convergence of Metropolis–Hastings moves when posterior model probabilities concentrate (Yang et al., 2016), sequential Monte Carlo methods to lower the cost of model search (Schäfer & Chopin, 2013), tempering strategies to explore model spaces with strong multi-modalities (Zanella & Roberts, 2019), or adaptive Markov Chain Monte Carlo to reduce the effort in exploring low posterior probability models (Griffin et al., 2020). Unfortunately, except for very specific settings such as Gaussian regression under conjugate priors, the integrated likelihood has no closed-form, which seriously hampers scaling computations to even moderate dimensions.

Our main contribution is proposing a simple yet powerful approximate inference technique, the Approximate Laplace Approximation (ALA). Analogously to the classical Laplace approximation (LA) to an integral, ALA uses a second-order Taylor expansion, the difference being that the expansion is done at a point that simplifies calculations. Also, there is a particular version of the ALA for which, within the exponential family, one may pre-compute statistics to obtain the ALA in all models. After said pre-computation, the computational cost does not depend on n.

We outline the idea. Let $y = (y_1, ..., y_n)$ be an observed outcome of interest and suppose that one considers several models $\gamma \in \Gamma$ within some set of models Γ. Given prior model probabilities $p(\gamma)$, Bayesian model selection assigns posterior probabilities $p(\gamma|y) = p(y|\gamma)p(\gamma)/p(y)$, where

$$p(y \mid \gamma) = \int p(y \mid \eta_\gamma, \gamma) p(\eta_\gamma \mid \gamma) d\eta_\gamma \tag{1}$$

is the integrated likelihood, $p(y \mid \eta_\gamma, \gamma)$ the likelihood-function under model γ, $\eta_\gamma \in \mathbb{R}^{p_\gamma}$ the model parameters, $p(\eta_\gamma \mid \gamma)$ their prior density, and $p(y) = \sum_{\gamma \in \Gamma} p(y \mid \gamma) p(\gamma)$. The LA provides an approximation $\widehat{p}(y \mid \gamma)$ using a Taylor expansion of the log-integrand in Equation (1) at the posterior mode $\widehat{\eta}_\gamma$, giving

$$\widehat{p}(y \mid \gamma) = p(y \mid \widehat{\eta}_\gamma, \gamma) p(\widehat{\eta}_\gamma \mid \gamma) (2\pi)^{p_\gamma/2} |\widehat{H}_\gamma|^{-\frac{1}{2}}, \tag{2}$$

where $\widehat{H}_\gamma$ is the log-integrand's negative hessian at $\widehat{\eta}_\gamma$. Although $\widehat{p}(y \mid \gamma)$ is typically accurate, the optimization to obtain $\widehat{\eta}_\gamma$ can be costly when $p_\gamma = \dim(\eta_\gamma)$ is large, especially when one repeats such a calculation for many models. It is also costly when the sample size n is large, since for most common models evaluating the likelihood and derivatives has a linear cost in n, and sometimes higher (e.g. high-dimensional models where p_γ grows with n).

ALA avoids the need to obtain $\hat{\eta}_\gamma$ by expanding the log-likelihood at a suitably-chosen initial value $\eta_{\gamma 0}$, saving the associated optimization time to compute $\hat{\eta}_\gamma$. Let $\tilde{\eta}_\gamma = \eta_{\gamma 0} - H_{\gamma 0}^{-1} g_{\gamma 0}$ be a guess at $\hat{\eta}_\gamma$ given by a Newton–Raphson iteration from $\eta_{\gamma 0}$, where $g_{\gamma 0}$ and $H_{\gamma 0}$ are the gradient and hessian of the negative log-likelihood at $\eta_{\gamma 0}$. A quadratic log-likelihood expansion at $\eta_{\gamma 0}$ (see Section S1) gives the ALA to the integrated likelihood

$$\tilde{p}(y \mid \gamma) = p(y \mid \eta_{\gamma 0}, \gamma) p(\tilde{\eta}_\gamma \mid \gamma) (2\pi)^{p_\gamma/2} \mid H_{\gamma 0} \mid^{-\frac{1}{2}} \exp\{\frac{1}{2} g_{\gamma 0}^T H_{\gamma 0}^{-1} g_{\gamma 0}\}, \tag{3}$$

leading to ALA posterior probabilities $\tilde{p}(\gamma \mid y) = \tilde{p}(y \mid \gamma) p(\gamma) / \sum_{\gamma' \in \Gamma} \tilde{p}(y \mid \gamma') p(\gamma')$. Figure 1 offers a simple illustration in a univariate logistic regression example. See Section S1.1 for an alternative ALA based on expanding the full integrand, which attains the same rates as (3) under mild conditions and performed similarly in our examples.

We focus attention in regression problems where setting coefficients in $\eta_{\gamma 0}$ to 0 results in further simplifications, particularly in exponential family models where it allows pre-computing sufficient statistics. The computational savings are substantial, see Figure 2 and Figure S1 for logistic and Poisson regression examples. Even when sufficient statistics are not available the savings from avoiding the optimization exercise can still be significant, see our survival model examples in Table 1.

A caveat is that, unlike LA, in general $\tilde{p}(y \mid \gamma)$ does not consistently estimate $p(y|\gamma)$ as $n \to \infty$. For concave log-likelihoods the LA has relative error converging to 1 in probability, under the minimal condition that $\widehat{H}_\gamma / n$ converges in probability to a positive-definite matrix (Rossell and Rubio (2019), Proposition 8). Under further conditions, the LA estimates Bayes factors with relative error of order $1/n^2$ (Kass et al., 1990), see also Ruli et al. (2016) for higher-order approximations to high-dimensional integrals. The ALA does not equip such properties. Figure 1 (right) illustrates a situation where, due to the posterior distribution concentrating far from

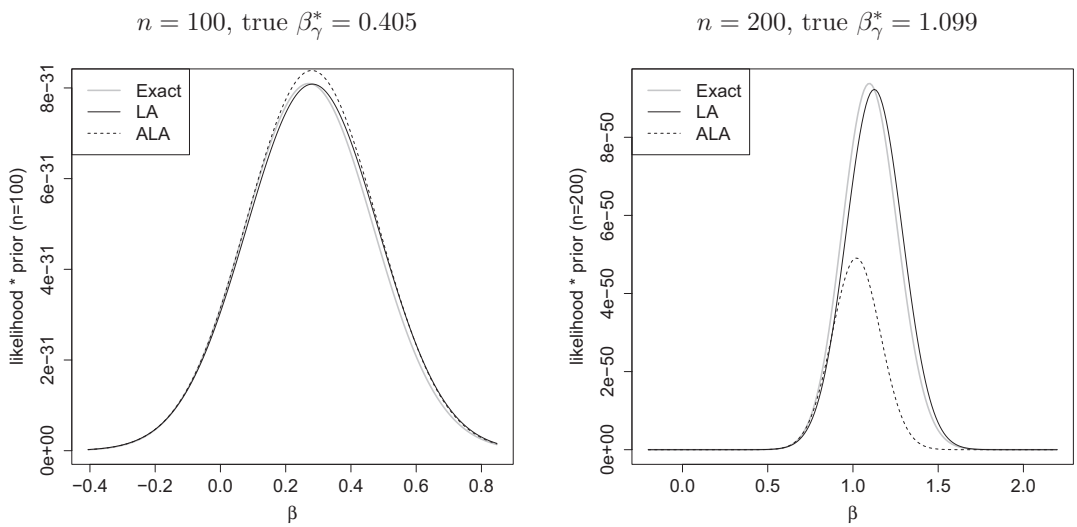

FIGURE 1 Logistic regression simulation in one dimension with a standard Gaussian prior on the coefficient. The likelihood multiplied by the prior $p(y \mid \beta_\gamma, \gamma) N(\beta_\gamma; 0, 1)$ is plotted in grey (Exact). The solid black line (LA) plots an approximation by replacing the log-likelihood with a second-order Taylor expansion at the MLE. The dashed line (ALA) does the same with a quadratic expansion around zero

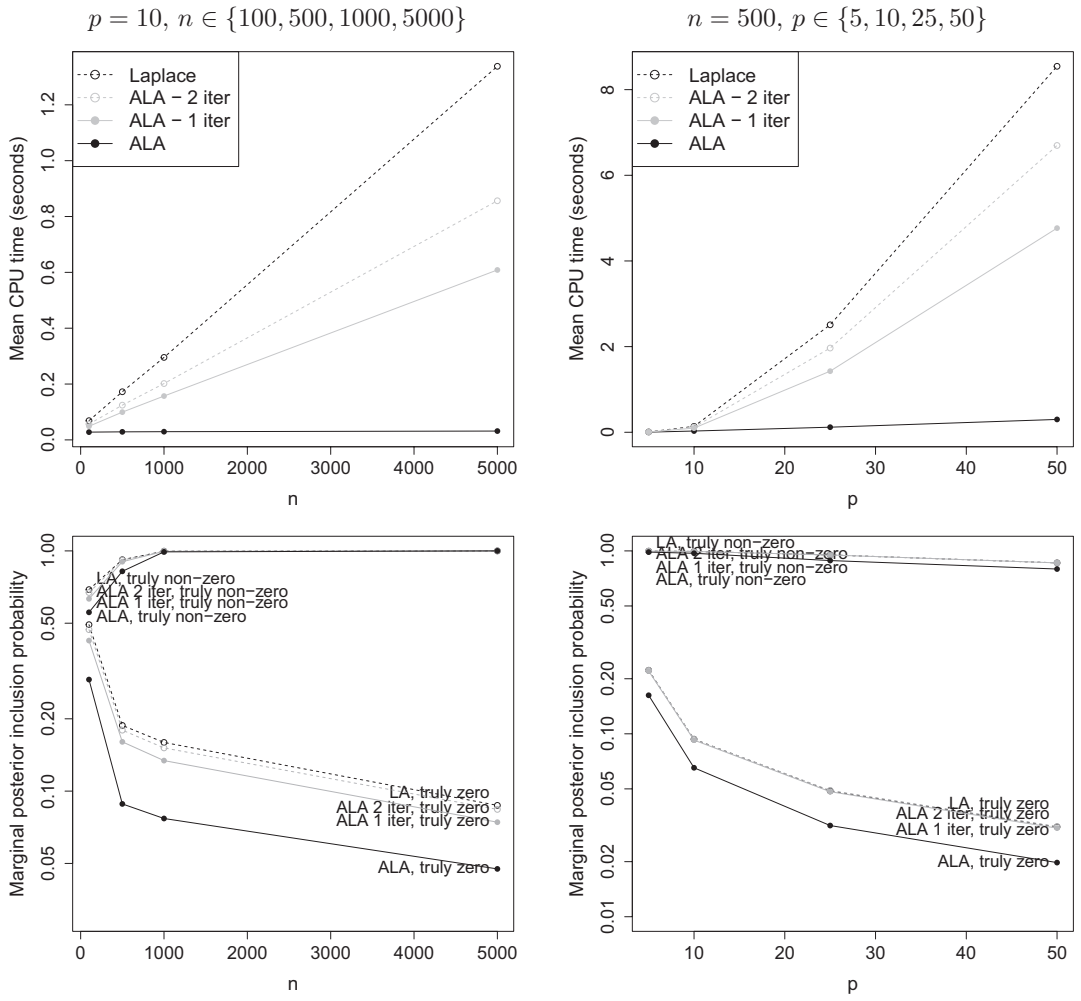

FIGURE 2 Logistic regression simulation. Top: average run time (seconds) in a single-core i7 processor running Ubuntu 20.04. Bottom: average posterior inclusion probabilities for truly active and inactive variables

the expansion point, ALA significantly underestimates the integral. Nevertheless ALA attains a strong type of model selection consistency, even when (inevitably) models are misspecified, that is the data are truly generated by a distribution F^* outside the considered models. Specifically, we prove that the ALA posterior probability $\tilde{p}(\tilde{\gamma}^* \mid y)$ converges to 1 in the L_1 sense for a suitably-defined optimal model $\tilde{\gamma}^*$. Said $\tilde{\gamma}^*$ is in general different from the model γ^* asymptotically recovered by exact calculations. Under misspecification, neither γ^* nor $\tilde{\gamma}^*$ in general recover the set of covariates associated to the mean of y under the data-generating F^*, but the optimal covariates under two different implicit loss functions. However we show via examples that $\tilde{\gamma}^*$ and γ^* often coincide, and provide a sufficient condition for $\tilde{\gamma}^*$ to discard truly spurious parameters. Intuitively, the reason why situations like that in Figure 1 (right) need not be problematic is that Bayes factors target ratios of integrated likelihoods. Despite the integral being under-estimated, it is still very large relative to the likelihood at 0 is very large, signaling that the parameter should be included.

TABLE 1 Mean run times for poverty data, survival analysis in truly AFT simulations, truly PH simulations, and colon cancer data. Laptop with Ubuntu 18.04, Intel i7 1.8GHz processor, 15.4 Gb RAM, 1 core

Poverty data ($n = 89{,}755$)		
	Main effects ($p = 60$)	Interactions ($p = 1469$)
gZellner ALA	19.5 s	5.2 min
gZellner LA	10.1 days	17.3 days
Simulation under true AFT model ($J = 100$, $p = 300$)		
	$n = 100$	$n = 500$
gMOM ALA	4.0 s	7.5 s
gZellner ALA	9.6 s	13.2 s
gZellner LA	188.9 s	251.1 s
Simulation under true PH model ($J = 100$, $p = 300$)		
	$n = 100$	$n = 500$
gMOM ALA	2.0 s	4.1 s
gZellner ALA	2.3 s	10.7 s
gZellner LA	152.4 s	81.1 s
Colon cancer data ($n = 260$, $p = 175$)		
	Uniform $p(\gamma)$	BetaBin $p(\gamma)$
gMOM ALA	3.3 min	39.3 s
gZellner ALA	11.1 min	48.8 s
gZellner LA	55.0 h	15.1 min

Relative to approximate inference methods primarily designed for estimation or prediction such as variational Bayes (Jordan et al., 1999) or expectation propagation (Minka, 2001), ALA focuses on model selection problems where the goal is structural learning; see, however, Carbonetto and Stephens (2012) and Huang et al. (2016) for variational Bayes approaches to variable selection in Gaussian regression with conjugate priors. We focus our study on a wide model class within the exponential family, which includes generalized linear, generalized additive models and other (possibly non-additive) generalized structured regression. We also illustrate the use of ALA with concave log-likelihoods outside the exponential family, via a non-linear additive accelerated failure time model (Rossell & Rubio, 2019). We incorporate two aspects where state-of-the-art methods encounter difficulties. First, we consider that the model selection exercise may combine group and hierarchical constraints, a case where penalized likelihood and shrinkage prior methods can face difficulties in terms of computational complexity. Said constraints are relevant when one considers categorical covariates, interaction terms, and semi- and non-parametric covariate effects, for example. Second, we use ALA to facilitate computation for non-local priors (Johnson & Rossell, 2010, 2012). Non-local priors attain some of the strongest theoretical properties among Bayesian methods in high dimensions, see Shin et al. (2018) and Rossell (2018). However, exact calculations are unfeasible, the LA is costly and it does not consistently estimate $p(y|\gamma)$ for any model γ that includes truly spurious parameters (Rossell & Telesca, 2017). Interestingly, although we generally view the ALA as fast approximate inference that may perform slightly worse than the LA, for non-local priors the ALA often attains better inference.

The paper is structured as follows. Section 1 reviews exponential family models and discusses computational savings associated to the ALA. Although the ALA applies to a wide set of priors, for concreteness Section 2 outlines specific local and a non-local priors on parameters that we use in our examples, and a group hierarchical prior on models $p(\gamma)$. The development of the non-local prior is in fact a secondary contribution of this paper: it is a novel class combining additive penalties (Johnson & Rossell, 2010) suitable for group constraints with product-type penalties required for high-dimensional consistency (Johnson & Rossell, 2012). Section 3 gives specific ALA expressions for local priors, and subsequently for the more challenging non-local prior case. Section 4 gives model selection consistency results for ALA, specifically rates that hold for fixed p under minimal conditions and high-dimensional rates where p grows with n, under slightly stronger conditions. We distinguish cases where the exponential family has a known dispersion parameter (e.g. logistic and Poisson regression) and cases where it is unknown. In particular our high-dimensional theory focuses on the known case, to alleviate the technical exposition, but our results also apply to Gaussian outcomes with unknown error variance. Section 5 shows examples assessing the numerical accuracy of ALA, the computational time, and the quality of its associated model selection. We consider logistic, Poisson and survival examples, as well as non-linear Gaussian regression under non-local priors where $p(y|\gamma)$ are hard to approximate. We also briefly illustrate the use of ALA in combination with importance sampling, variable screening and adding optimization iterations to improve the expansion point $\eta_{\gamma 0}$ in Equation (3). Section 6 concludes. The supplementary material contains proofs, derivations and supplementary results. R code and data to reproduce our examples are available at https://github.com/davidrusi/paper_examples/tree/main/2020_Rossell_Abril_Bhattacharya_ALA.

1 | LIKELIHOOD

We lay out notation. Let $x_i \in \mathcal{X}$ be covariates taking values in some domain $\mathcal{X}$. Consider a generalized structured regression with predictor

$$h(E(y_i \,|\, x_i)) = \sum_{j=1}^{J} z_{ij}^T \beta_j, \tag{4}$$

where $h()$ is the canonical link function and $z_i = (z_{i1}^T, \ldots, z_{iJ}^T)^T$ a basis for the effect of x_i with coefficients $\beta = (\beta_1^T, \ldots, \beta_J^T)^T$. For example a standard generalized linear model corresponds to $z_i = x_i$. We also consider situations where each $z_{ij} \in \mathbb{R}^{p_j}$ defines a group with p_j elements, e.g. multiple binary indicators for a categorical covariate or a non-linear basis expansion for a continuous covariate. That is, (4) includes additive regression on functions of x_i, non-linear interactions between elements of x_i, for example. Let $p = \sum_{j=1}^{J} p_j$ be the total number of parameters and $Z = (z_1^T, \ldots, z_n^T)^T$ the $n \times p$ design matrix.

Our goal is to determine which $\beta_j \in \mathbb{R}^{p_j}$ should be set to zero. Let $\gamma_j = I(\beta_j \neq 0)$ for $j = 1, ..., J$ be group inclusion indicators, so that $\gamma = (\gamma_1, ..., \gamma_J)$ indexes the model. We denote by Z_γ the $n \times p_\gamma$ submatrix of Z with (blocks of) columns selected by γ where $p_\gamma = \sum_{j:\gamma_j=1} p_j$, and by $z_{\gamma i}$ its i^{th} row. For any given model γ, the distribution of y is assumed to be in the exponential family with canonical link and likelihood function

$$p(y \,|\, \beta, \phi, \gamma) = \exp\left\{ [y^T Z_\gamma \beta_\gamma - \sum_{i=1}^{n} b(z_{\gamma i}^T \beta_\gamma)]/\phi + \sum_{i=1}^{n} c(y_i, \phi) \right\}, \tag{5}$$

where $\phi > 0$ is an optional dispersion parameter and $b()$ an infinitely differentiable function.

The gradient and hessian of the negative log-likelihood $-\log p(y|\beta, \phi, \gamma)$ are

$$g_\gamma(\beta_\gamma, \phi) = -\frac{1}{\phi}\begin{pmatrix} Z_\gamma^T y - \sum_{i=1}^n b'(z_{\gamma i}^T \beta_\gamma) z_{\gamma i} \\ -[y^T Z_\gamma \beta_\gamma - \sum_{i=1}^n b(z_{\gamma i}^T \beta_\gamma)]/\phi + \phi \sum_{i=1}^n \nabla_\phi c(y_i, \phi) \end{pmatrix}$$

$$H_\gamma(\beta_\gamma, \phi) = \frac{1}{\phi}\begin{pmatrix} Z_\gamma^T D_\gamma Z_\gamma & g_\beta(\beta_\gamma, \phi) \\ g_\beta(\beta_\gamma, \phi)^T & -2[y^T Z_\gamma \beta_\gamma - \sum_{i=1}^n b(z_{\gamma i}^T \beta_\gamma)]/\phi^2 - \phi \sum_{i=1}^n \nabla_{\phi\phi}^2 c(y_i, \phi) \end{pmatrix}$$

where D_γ is an $n \times n$ diagonal matrix with diagonal entry $b''(z_{\gamma i}^T \beta_\gamma) > 0$. For completeness, Section S2 provides expressions for logistic and Poisson models.

A computationally-convenient choice for the ALA in Equation (3) is to set a global $\beta_0 \in \mathbb{R}^p$ and let $\eta_{\gamma 0} = (\beta_{\gamma 0}, \phi_0)$, where $\beta_{\gamma 0}$ contains the entries of β_0 selected by γ and, if ϕ is a unknown parameter,

$$\phi_0 = \operatorname{argmax}_\phi p(y \mid \beta = \beta_0, \phi) \tag{6}$$

is the maximum likelihood estimator conditional on $\beta = \beta_0$. Since ϕ_0 does not depend on γ it can be computed upfront and shared across all models. By basing ALA on such a global choice, one avoids the model-specific optimization costs that would be required by a LA.

The choice $\beta_0 = 0$ gives further computational simplifications (one may also set the intercept to a non-zero value at essentially no cost). To ease notation let $\tilde{y} = (y - b'(0)\mathbb{1})/b''(0)$ denote a shifted and scaled version of y, $\mathbb{1} = (1, \ldots, 1)^T$ being the $n \times 1$ unit vector. The gradient and hessian at $(\beta_{\gamma 0}, \phi) = (0, \phi_0)$ are

$$\begin{aligned} g_{\gamma 0} &= -\frac{b''(0)}{\phi_0}\begin{pmatrix} Z_\gamma^T \tilde{y} \\ 0 \end{pmatrix} \\ H_{\gamma 0} &= \frac{b''(0)}{\phi_0}\begin{pmatrix} Z_\gamma^T Z_\gamma & -Z_\gamma^T \tilde{y}/\phi_0 \\ -\tilde{y}^T Z_\gamma/\phi_0 & s(\phi_0) \end{pmatrix}, \end{aligned} \tag{7}$$

where $s(\phi_0) = [2nb(0)/\phi_0^2 + \phi_0 \sum_{i=1}^n \nabla_{\phi\phi}^2 c(y_i, \phi_0)]/b''(0)$. To interpret these expressions, the exponential family predicted variance for $\beta_{\gamma 0} = 0$ is $V(y_i \,|\, z_{\gamma i}, \beta_{\gamma 0} = 0, \phi) = \phi b''(0)$, hence $V(\tilde{y}_i \,|\, z_{\gamma i}, \beta_{\gamma 0} = 0, \phi) = \phi/b''(0)$. Thus, $(g_{\gamma 0}, H_{\gamma 0})$ are analogous to the gradient and hessian in a least-squares regression of $\tilde{y}$ on Z, with model-based variance $\phi/b''(0)$.

Sections 3 and 4 show that $\eta_{\gamma 0} = (0, \phi_0)^T$ leads to desirable model selection rates. The ALA then basically requires least-squares type computations where $(Z^T\tilde{y}, Z^TZ)$ play the role of sufficient statistics that can be computed upfront and shared across all models. To further save memory and computational requirements, in our implementation we store Z^{TZ} in a sparse matrix that is incrementally filled the first time that any given entry is required. That is, when searching models typically many elements in Z^{TZ} are never used, hence there is no need to compute nor to allocate them to memory beforehand. One may also consider alternative $\eta_{\gamma 0}$, say obtained after a few Newton–Raphson iterations, in an attempt to obtain an ALA that is closer to the LA in Equation (2). See Figure 2 and Sections 5.1 and S12.2 for some examples. Such alternatives can

lead to improved inference, at a higher computational cost. Their theoretical study requires a separate treatment, however, and is left for future work.

2 | PRIOR

Most of our results apply to a wide class of priors $p(\gamma)$. For concreteness we outline a structure that assigns the same probability to all models with the same number of active groups $|\gamma| = \sum_{j=1}^{J} \gamma_j$, and an arbitrary distribution $p(|\gamma|)$ on $|\gamma|$.

Although unnecessary in canonical regression problems, we also consider that in certain situations one may want to impose hierarchical constraints, in the sense that $\beta_l = 0$ implies $\beta_j = 0$ for some $l \neq j$. For instance, one may exclude interaction terms unless the corresponding main effects are present, or decompose non-linear effects as a linear plus a non-linear term, and only include the latter if the linear term is present (Rossell & Rubio, 2019; Scheipl et al., 2012). Such (optional) constraints can be added as follows.

Let $C \subseteq \Gamma$ be the models satisfying the constraints. These are easily incorporated by assigning $\pi(\gamma) = 0$ to any $\gamma \notin C$. Specifically,

$$p(\gamma) = \begin{cases} Kp(|\gamma|)\binom{J}{|\gamma|}^{-1}, & \text{if } \gamma \in C \text{ and } |\gamma| \leq \bar{J} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where $\bar{J}$ is the maximum model size one wishes to consider and K a prior normalization constant that does not need to be evaluated explicitly. The formulation allows both a number of parameters $p \gg n$ and groups $J \gg n$, but restricts the model space to using combinations of at most $\bar{J}$ groups. Given that models with $p_\gamma \geq n$ parameters result in data interpolation, typically one sets both $\bar{J} \ll n$ and $p_\gamma \ll n$ for any allowed $\gamma \in C$, see Section 4.2 for further discussion. In a standard generalized linear model without groups nor constraints; $J = p$, the constraint $\gamma \in C$ is removed, and $K = 1$.

In Section 4 we provide pairwise Bayes factor rates for general $p(|\gamma|)$, whereas to ease exposition for posterior model probability rates we focus on

$$p(|\gamma|) \propto p^{-c|\gamma|},$$

where $c \geq 0$ is a user-specified constant and $\propto$ denotes "proportional to". For $c = 0$ one obtains uniform $p(|\gamma|) = 1/(\bar{J} + 1)$, which generalizes the Beta-Binomial(1,1) distribution advocated by Scott and Berger (2006) to a setting where there may be groups and hierarchical constraints. In our experience $c = 0$ strikes a good balance between sparsity and retaining power to detect truly non-zero coefficients, hence in all our examples we used $c = 0$. One may also set $c > 0$, which is motivated by the so-called Complexity priors of Castillo et al. (2015). These set a stronger prior penalty on the model size that leads to faster rates to discard spurious parameters, at the cost of slower rates to detect active parameters. See Section 4 for further details.

We remark that adding hierarchical constraints to penalized likelihood and Bayesian shrinkage frameworks lead to computational difficulties. For instance, hierarchical constraints for LASSO penalties lead to a challenging optimization problem, and while one can devise relaxed constraints (Bien et al., 2013), our examples indicate the computation can be prohibitive.

2.1 | Group product priors

Regarding the prior on parameters, our examples use Normal priors and a novel group moment (gMOM) prior family factorizing over groups

$$\begin{aligned} p^L(\beta_\gamma|\phi,\gamma) &= \prod_{\gamma_j=1} N\left(\beta_j; 0, \frac{\phi g_L n}{p_j}(Z_j^T Z_j)^{-1}\right) \\ p^N(\beta_\gamma|\phi,\gamma) &= \prod_{\gamma_j=1} \frac{\beta_j^T Z_j^T Z_j \beta_j}{\phi g_N n p_j/(p_j+2)} N\left(\beta_j; 0, \frac{\phi g_N n}{p_j+2}(Z_j^T Z_j)^{-1}\right) \end{aligned} \tag{9}$$

and, for models where ϕ is unknown, we set $p(\phi) = \text{IG}(\phi;\, a,\, b)$ for given prior parameters $g_L, g_N, a, b > 0$. All other parameters (β_j such that $\gamma_j = 0$) are zero with probability 1. Both priors feature a Normal kernel with a group-Zellner precision matrix given by $Z_j^T Z_j$. Other covariances may be used, but our choice leads to inference that is robust to affine within-group reparameterizations of β_j (for example, changing the reference category for discrete predictors), and to simple default parameter values $g_L = g_N = 1$ (Section 2.2).

The group Zellner $p^L()$ is a local prior, in the nomenclature of Johnson and Rossell (2010), whereas the gMOM $p^N()$ is a non-local prior. The defining property of non-local priors is that the density vanishes as β_γ approaches any value that lies in the parameter space of a submodel of γ, i.e. $\beta_j = 0$ in our setting. Their interest is that they help discard spurious parameters, by inducing a data-dependent penalty that has little asymptotic effect on power (Rossell & Telesca, 2017). Earlier proposals (Johnson & Rossell, 2010, 2012) did not account for group structure, however. The intuition is simple, the gMOM penalizes groups with small contributions $\beta_j^T Z_j^T Z_j \beta_j$ relative to its size $p_j = \dim(\beta_j)$, which helps induce sparsity.

2.2 | Prior elicitation

Although our focus is computational and our theory applies to any prior parameters (g_L, g_N) (under minimal conditions), we outline a simple strategy to obtain default (g_L, g_N) that we used in our examples. The strategy builds upon the unit information prior, a popular default leading to the Bayesian information criterion (Schwarz, 1978), the difference being that we account for the presence of groups in Z_γ.

Suppose that there were no groups in Z_γ. The unit information prior can be interpreted as containing as much information as a single observation. Another (perhaps more natural) interpretation is its specifying the prior belief that $E(\beta_\gamma^T Z_\gamma^T Z_\gamma \beta_\gamma/[n\phi]) = p_\gamma$. The expected contribution $\beta_\gamma^T Z_\gamma^T Z_\gamma \beta_\gamma/n$ relative to the dispersion ϕ, which is a measure of the predictive ability contained in Z_γ, is given by the number of variables p_γ.

Suppose now that a variable defines a group of columns in Z_γ, for instance a non-linear basis expansion. Then p_γ depends on the basis dimension, which is often chosen arbitrarily. The unit information prior would imply the belief that the predictive power of Z_γ increases with the arbitrary basis dimension, rather than the number of variables $\sum_{j=1}^J \gamma_j$. Instead, we set prior parameters such that the prior expected predictive power depends on $\sum_{j=1}^J \gamma_j$ and is unaffected by the basis dimension. That is, we set (g_L, g_N) such that

$$E\left(\frac{\beta_\gamma^T Z_\gamma^T Z_\gamma \beta_\gamma}{n\phi}\right) = E\left(\sum_{\gamma_j=1} \frac{\beta_j^T Z_j^T Z_j \beta_j}{n\phi}\right) = \sum_{j=1}^{J} \gamma_j.$$

For the Normal and gMOM priors in Equation (9) this rule gives $g_L = g_N = 1$, see Section S3.

3 | APPROXIMATE LAPLACE APPROXIMATION

We first discuss the ALA under a local prior $p^L()$, and subsequently that for the gMOM prior $p^N()$. Recall that for the latter the integrated likelihood has no computationally-convenient closed-form, even in Gaussian regression.

3.1 | Local priors

The ALA to the Bayes factor between any pair of models (γ, γ') is

$$\tilde{B}^L_{\gamma\gamma'} = \frac{\tilde{p}^L(y|\gamma)}{\tilde{p}^L(y|\gamma')} = \exp\left\{\frac{1}{2}(g^T_{\gamma 0} H^{-1}_{\gamma 0} g_{\gamma 0} - g^T_{\gamma' 0} H^{-1}_{\gamma' 0} g_{\gamma' 0})\right\} (2\pi)^{\frac{p_\gamma - p_{\gamma'}}{2}} \frac{|H_{\gamma' 0}|^{\frac{1}{2}} p^L(\tilde{\eta}_\gamma|\gamma)}{|H_{\gamma 0}|^{\frac{1}{2}} p^L(\tilde{\eta}_{\gamma'}|\gamma')}. \quad (10)$$

Expression (10) can be used beyond the exponential family, provided the log-likelihood is concave. As an example, Section 5.4 illustrates the Gaussian accelerated failure time model; see Section S2.3 for the corresponding log-likelihood, gradient and derivatives, and the conditions for log-likelihood concavity.

We now provide specific expressions for exponential family models (5), and discuss a curvature adjustment designed to improve finite n performance. Consider first the case where ϕ is a known constant, as in logistic and Poisson models. Taking $\eta_{\gamma 0} = \beta_{\gamma 0} = 0$ gives

$$\tilde{B}^L_{\gamma\gamma'} = \exp\left\{\frac{b''(0)}{2\phi}(\tilde{\beta}^T_\gamma Z^T_\gamma Z_\gamma \tilde{\beta}_\gamma - \tilde{\beta}^T_{\gamma'} Z^T_{\gamma'} Z_{\gamma'} \tilde{\beta}_{\gamma'})\right\} \left(\frac{2\pi\phi}{b''(0)}\right)^{\frac{p_\gamma - p_{\gamma'}}{2}} \frac{|Z^T_{\gamma'} Z_{\gamma'}|^{\frac{1}{2}} p^L(\tilde{\beta}_\gamma|\phi, \gamma)}{|Z^T_\gamma Z_\gamma|^{\frac{1}{2}} p^L(\tilde{\beta}_{\gamma'}|\phi, \gamma')} \quad (11)$$

where $\tilde{\beta}_\gamma = (Z^T_\gamma Z_\gamma)^{-1} Z^T_\gamma \tilde{y}$ and $\tilde{y} = (y - b'(0)\mathbb{1})/b''(0)$. See Section S1.2 for the derivation.

Consider now the case where ϕ is an unknown model parameter. Then, taking $\eta_{\gamma 0} = (0, \phi_0)$ as in Section 1, one obtains

$$\tilde{B}^L_{\gamma,\gamma'} = \exp\left\{\frac{b''(0)}{2\phi_0}[t_\gamma \tilde{\beta}^T_\gamma Z^T_\gamma Z_\gamma \tilde{\beta}_\gamma - t_{\gamma'} \tilde{\beta}^T_{\gamma'} Z^T_{\gamma'} Z_{\gamma'} \tilde{\beta}_{\gamma'}]\right\} (2\pi)^{\frac{p_\gamma - p_{\gamma'}}{2}} \frac{|H_{\gamma' 0}|^{\frac{1}{2}} p^L(\tilde{\beta}_\gamma, \tilde{\phi}_\gamma|\gamma)}{|H_{\gamma 0}|^{\frac{1}{2}} p^L(\tilde{\beta}_{\gamma'}, \tilde{\phi}_{\gamma'}|\gamma')} \quad (12)$$

where

$$t_\gamma = 1 + \frac{\tilde{\beta}^T_\gamma Z^T_\gamma Z_\gamma \tilde{\beta}_\gamma}{\phi_0^2(s(\phi_0) - \tilde{\beta}^T_\gamma Z^T_\gamma Z_\gamma \tilde{\beta}_\gamma)},$$

and $s(\phi_0)$ is as in Equation (7), see Section S1.3 for the derivation.

3.2 | Curvature adjustment

The Bayes factor in Equation (11) for models where ϕ is known attains desirable theoretical properties as $n \to \infty$, see Section 4. There is however an important practical remark, which makes us recommend a curvature-adjusted ALA to improve finite n performance. We outline the idea and refer the reader to Section S1.4 for a full description. Expression (10) can be given in terms of the model-predicted covariance $\text{Cov}(y \mid Z_\gamma, \beta = \beta_{\gamma 0}, \phi) = E((y-\mu_{\gamma 0})^T(y-\mu_{\gamma 0}) \mid Z_\gamma, \beta = \beta_{\gamma 0}, \phi)$, where $\mu_{\gamma 0} = E(y \mid Z_\gamma, \beta = \beta_{\gamma 0}, \phi)$. Even when the data are truly generated from a distribution F^* included in the assumed model (5) for some β_γ^*, there is a mismatch between said covariance and $E_{F^*}((y-\mu_{\gamma 0})^T(y-\mu_{\gamma 0}) \mid Z_\gamma)$, due to $\mu_{\gamma 0}$ being different from the true mean $E(y \mid Z_\gamma, \beta = \beta_\gamma^*, \phi)$. That is, the data may be either over- or under-dispersed relative to the model prediction at $\beta = \beta_{\gamma 0}$, which can adversely affect inference.

The curvature-adjusted ALA is obtained by replacing $\text{Cov}(y \mid Z_\gamma, \beta = \beta_{\gamma 0}, \phi)$ for $\widehat{\rho}\text{Cov}(y \mid Z_\gamma, \beta = \beta_{\gamma 0}, \phi) = \widehat{\rho}\phi\text{diag}(b''(Z_\gamma \beta_{\gamma 0}))$, where $\widehat{\rho} = \sum_{i=1}^n (y_i - \overline{y})^2 / [\phi b''(h(\overline{y}))(n-1)]$ is a Pearson residual-based estimate of over-dispersion, $h()$ is the link function in Equation (4) and $\overline{y} = \sum_{i=1}^n y_i / n$ the sample mean. The curvature-adjusted Bayes factor is

$$\tilde{B}_{\gamma\gamma'} = \frac{p(\tilde{\beta}_\gamma \mid \phi, \gamma) \mid Z_{\gamma'}^T Z_{\gamma'} \mid^{1/2}}{p(\tilde{\beta}_{\gamma'} \mid \phi, \gamma') \mid Z_\gamma^T Z_\gamma \mid^{1/2}} \left(\frac{2\pi\phi}{\widehat{\rho} b''(h(\overline{y}))} \right)^{(p_\gamma - p_{\gamma'})/2} e^{\frac{b''(h(\overline{y}))}{2\widehat{\rho}\phi} [\tilde{\beta}_\gamma^T (Z_\gamma^T Z_\gamma) \tilde{\beta}_\gamma - \tilde{\beta}_{\gamma'}^T (Z_{\gamma'}^T Z_{\gamma'}) \tilde{\beta}_{\gamma'}]}, \quad (13)$$

where $\tilde{\beta}_\gamma = (Z_\gamma^T Z_\gamma)^{-1} Z_\gamma^T [y - b'(h(\overline{y}))\mathbb{1}] / b''(h(\overline{y}))$. Note that $\tilde{\beta}_\gamma$ here denotes the parameter estimate after one Newton–Raphson iteration from the maximum likelihood estimator under the intercept-only model. It is possible to use alternative over-dispersion estimators that are specific for each model, at a slightly higher computational cost, see Section S1.4 for a discussion. In all our logistic and Poisson regression examples we used $\widehat{\rho}$ as outlined above, since in our experience this simple choice performs fairly well in practice. See Section 5.1 and Figure S1 for a Poisson example where the curvature adjustment significantly improves inference.

3.3 | Non-local priors

The ALA Bayes factors $\tilde{B}_{\gamma\gamma'}^N = \tilde{p}^N(y \mid \gamma) / \tilde{p}^N(y \mid \gamma')$ for the gMOM prior in Equation (9) require an alternative strategy. Let $\pi(\beta_\gamma, \phi \mid \gamma) = \mathcal{N}(\beta_\gamma; 0, \phi V_j^{-1})$ where $V_j = Z_j^T Z_j (p_j + 2) / (n p_j g_N)$, so that the gMOM prior equals

$$p^N(\beta_\gamma \mid \phi, \gamma) = \pi(\beta_\gamma \mid \phi, \gamma) \prod_{\gamma_j = 1} \frac{\beta_j^T V_j \beta_j}{\phi p_j}.$$

Denote by $\pi(\beta_\gamma, \phi \mid y, \gamma)$ the posterior density and $\pi(y|\gamma)$ the integrated likelihood associated to the prior $\pi(\beta_\gamma, \phi \mid \gamma)$. By Proposition 1 in Rossell and Telesca (2017), the identity

$$p^N(y|\gamma) = \pi(y|\gamma) \int \left(\prod_{\gamma_j = 1} \frac{\beta_j^T V_j \beta_j}{\phi p_j} \right) \pi(\beta_\gamma, \phi | y, \gamma) d\beta_\gamma d\phi \quad (14)$$

holds exactly. $\pi(y|\gamma)$ is the integrated likelihood under a local prior, hence one may obtain an ALA $\tilde{\pi}(y\,|\,\gamma)$ as described in Section 3.1. The second term in Equation (14) is the posterior expectation of a product, and its computation requires a number of operations that grow exponentially with the model dimension p_γ. As an alternative, in Equation (14) we replace $\pi(\beta_\gamma\,|\,\phi, y)$ by its ALA-based normal approximation and we also replace the integral by a product of expectations. Specifically,

$$\tilde{p}^N(y\,|\,\gamma) = \tilde{\pi}(y\,|\,\gamma)\left[\prod_{\gamma_j=1}\int \frac{\beta_j^T V_j \beta_j}{\phi p_j}\mathcal{N}(\beta_j; m_j, \phi S_j)\tilde{\pi}(\phi|y,\gamma)d\beta_j d\phi\right] \tag{15}$$

where m_j and S_j are the sub-vector of $\tilde{\beta}_\gamma = (Z_\gamma^T Z_\gamma)^{-1}Z_\gamma^T\tilde{y}$ and sub-matrix of $(Z_\gamma^T D_\gamma Z_\gamma)^{-1}$ associated to β_j, and recall that D_γ is diagonal with (i, i) entry $z_{i\gamma}^T\tilde{\beta}_\gamma$. The following lemma is useful.

Lemma 1 *Let* A *and* S *be* $\mathrm{l}\times\mathrm{l}$ *full-rank matrices and* a, b > 0 *be constants. Then*

$$\int \frac{\xi^T A\xi}{\phi}\mathcal{N}(\xi; m, \phi S)d\xi = tr(AS) + \frac{m^T Am}{\phi}$$
$$\int\int \frac{\xi^T A\xi}{\phi}\mathcal{N}(\xi; m, \phi S)\mathrm{IG}(\phi; a, b)d\xi d\phi = tr(AS) + \frac{a}{b}m^T Am.$$

By Lemma 1 when ϕ is known the integral in Equation (15) has the simple expression

$$\tilde{p}^N(y|\phi,\gamma) = \tilde{\pi}(y|\phi,\gamma)\left[\prod_{\gamma_j=1}\mathrm{tr}(V_j S_j)/p_j + \frac{m_j^T V_j m_j}{\phi p_j}\right]. \tag{16}$$

As a remark, in linear regression with known error variance ϕ where the groups β_j are independent a posteriori ($Z^T Z$ is block-diagonal), then $p^N(y\,|\,\phi,\,\gamma) = \tilde{p}^N(y\,|\,\phi,\,\gamma)$, i.e. Expression (16) is exact. In contrast, Laplace approximations $\hat{p}^N(y\,|\,\phi,\,\gamma)$ are not consistent even in this simplest setting. See Section 5.2.1 for examples.

Lemma 1 is also useful in Gaussian regression with unknown error variance ϕ. Suppose that one sets the prior $\phi \sim \mathrm{IG}(\phi; a', b')$, then

$$\tilde{p}^N(y\,|\,\phi,\gamma) = \pi(y\,|\,\phi,\gamma)\left[\prod_{\gamma_j=1}\mathrm{tr}(V_j S_j)/p_j + m_j^T V_j m_j E(1/\phi|y,\gamma)/p_j\right],$$

where $E(1/\phi\,|\,y,\,\gamma) = (a' + n)/(b' + y^T y - \tilde{\beta}_\gamma^T Z_\gamma^T Z_\gamma\tilde{\beta}_\gamma)$ and $\pi(y|\phi,\,\gamma)$ has closed-form expression. Finally, for non-Gaussian regression case and unknown ϕ we propose

$$\tilde{p}^N(y\,|\,\gamma) = \tilde{\pi}(y\,|\,\gamma)\left[\prod_{\gamma_j=1}\mathrm{tr}(V_j S_j)/p_j + \frac{m_j^T V_j m_j}{\tilde{\phi}_\gamma p_j}\right]. \tag{17}$$

4 | THEORY

We consider a general setting where (y, Z) arise from a data-generating F^* that may be outside the assumed model class (5). We prove that as n grows the ALA-based $\tilde{p}(\gamma \mid y)$ assign probability increasing to 1 to an optimal $\tilde{\gamma}^*$. For the particular choice $\eta_{\gamma 0} = (0, \phi_0)$, such $\tilde{\gamma}^*$ is the smallest model minimizing a mean squared loss associated to linear projections. We first explain that exact $p(\gamma|y)$ asymptotically select a γ^*, in general different from $\tilde{\gamma}^*$, defined by a log-likelihood loss and Kullback-Leibler (KL) projections to F^*. We then provide Theorem 1, characterizing certain situations where γ^* coincides with $\tilde{\gamma}^*$. Subsequently, in Section 4.1 we consider fixed p settings where one may characterize $\tilde{p}(\tilde{\gamma}^* \mid y)$ via the rate at which pairwise Bayes factors $\tilde{B}_{\gamma,\tilde{\gamma}^*} = \tilde{p}(y \mid \gamma)/\tilde{p}(y \mid \tilde{\gamma}^*)$ converge to 0 in probability. Section 4.2 considers high-dimensional settings where p may grow with n.

For any model γ, denote the KL-optimal η_γ under F^* by

$$\eta_\gamma^* = \arg\max_{\eta_\gamma} E_{F^*}[\log p(y \mid \eta_\gamma, \gamma)], \tag{18}$$

and by η^* that under the full model $p(y|\eta)$ including all parameters. Multiple models may attain the global maximum $E_{F^*}(\log p(y \mid \eta_\gamma^*, \gamma)) = E_{F^*}(\log p(y \mid \eta^*))$, and we define the optimal γ^* as that with smallest dimension. If the model is well-specified, that is F^* is truly contained in Equation (18) for some η^*, then γ^* drops any parameters such that $\eta_j^* = 0$. Under misspecification, then γ^* is such that adding any parameter to γ^* cannot improve the fit, as measured by the expected log-likelihood in Equation (18).

In contrast, the ALA optimal model $\tilde{\gamma}^*$ is based on mean squared error or, equivalently, on linear projections of $E_{F^*}(\tilde{y} \mid Z)$, where $\tilde{y} = (y - b'(0)\mathbb{1})/b''(0)$ as in Equation (7). Let

$$\tilde{\beta}_\gamma^* = \arg\min_{\beta_\gamma} E_{F^*} \| \tilde{y} - Z_\gamma \beta_\gamma \|_2^2 = [E_{F^*}(Z_\gamma^T Z_\gamma)]^{-1} E_{F^*}[Z_\gamma^T \tilde{y}] \tag{19}$$

be the parameters giving the linear projection of $E_{F^*}(\tilde{y} \mid Z)$ on Z_γ, where we assume $E_{F^*}(Z_\gamma^T Z_\gamma)$ to be a finite positive-definite matrix for all $\gamma \in \Gamma$ (see Condition (C1) below). Then $\tilde{\gamma}^*$ is the smallest model minimizing mean squared error, that is $\tilde{\gamma}^* = \arg\min_{\gamma \in \tilde{\Gamma}^*} p_\gamma$ where

$$\tilde{\Gamma}^* = \{\gamma : E_{F^*} \| \tilde{y} - Z_\gamma \tilde{\beta}_\gamma^* \|_2^2 = \min_\beta E_{F^*} \| \tilde{y} - Z\beta \|_2^2\}$$

For simplicity we assume $\tilde{\gamma}^*$ to be unique, but our results generalize when there are multiple such models by defining $\tilde{\gamma}^*$ to be their union.

It is important to note that when (5) is misspecified neither γ^* nor $\tilde{\gamma}^*$ recover the truth, but the simplest model according to their implicit loss functions. That said, Theorem 1 below delineates an interesting robustness property of the linear projection $\tilde{\beta}^* = [E_{F^*}(Z^T Z)]^{-1} E_{F^*}(Z^T \tilde{y})$, under which terms that do not affect $E_{F^*}(y \mid Z)$ are discarded by $\tilde{\gamma}^*$.

Theorem 1 *Suppose* $(y_i, z_i) \overset{\text{i.i.d.}}{\sim} F^*$ *for* i = *1, ...,* n*, where* F^* *is a probability distribution on* $\mathbb{R} \otimes \mathbb{R}^p$ *with a finite positive-definite covariance matrix. Also, assume* $E_{F^*}(z_i) = 0$*, i.e., the covariate distribution is centered. Let* $\delta \subseteq \{1, ..., p\}$ *denote the true regression model so that* $E_{F^*}(y_i \mid z_i) = f(z_{i\delta})$ *almost-everywhere* F^**, where* $f: \mathbb{R}^{p_\delta} \to \mathbb{R}$ *is a measurable function. Letting* $\upsilon = \{1, ..., p\} \setminus \delta$ *indicate the truly inactive parameters, assume that*

$$E_{F^*}(z_{iv} \mid z_{i\delta}) = Az_{i\delta} \text{ almost-everywhere } F^*, \tag{20}$$

where $A \in \mathbb{R}^{(p-p_\delta)\times p_\delta}$. Then, $\tilde{\beta}^* = (\tilde{\beta}^*_\delta; 0)$, where recall that $\tilde{\beta}^*_\delta = [E_{F^*}(Z_\delta^T Z_\delta)]^{-1} E_{F^*}[Z_\delta^T \tilde{y}]$. In particular, $\tilde{\beta}^*_j = 0$ whenever $j \notin \delta$.

The assumption in Equation (20) states that the conditional mean of truly spurious variables is linear in the truly active $z_{i\delta}$. For example, the assumption is satisfied when the components of z_i are independent, or when their marginal distribution under F^* follows a (centered) elliptical distribution, such as the multivariate Gaussian or T family. Since we assumed that z_i has a finite positive-definite covariance, standard elliptical results (see, e.g., Chapter 1 of Muirhead (2009)) give that $E_{F^*}(z_{iv} \mid z_{i\delta})$ is linear in $z_{i\delta}$.

Under the assumptions of Theorem 1, the ALA-optimal model $\tilde{\gamma}^*$ is contained in the true model δ. In fact, $\tilde{\gamma}^* = \delta$ whenever all the entries of $\tilde{\beta}^*_\delta$ are non-zero. More generally, since $\tilde{\gamma}^*$ is defined by zeroes in $\tilde{\beta}^*$, we have that $\tilde{\gamma}^*$ includes any term j conditionally uncorrelated with the true regression function $f(z_{i\delta})$, that is satisfying $\mathrm{Cov}_{F^*}(z_{ij}, f(z_{i\delta}) \mid z_{i\tilde{\gamma}^*}) = 0$. Observe that $f(z_{i\delta})$ need not be linear for the theorem to hold. For example, for a truly generalized linear model with $E_{F^*}(y_i \mid z_i) = f(\sum_{j\in\delta} \beta^*_j z_{ij})$, we have $f = h^{-1}$ from Equation (4). More flexible models such as mixtures of generalized linear models are also permitted. For example, consider a two-component mixture with $E(y_i \mid z_i) = \pi f(\sum_{j\in\delta_1} \beta^*_{1j} z_{ij}) + (1-\pi) f(\sum_{k\in\delta_2} \beta^*_{2k} z_{ik})$, $\pi \in [0, 1]$. Then, $\tilde{\gamma}^*$ discards any variable outside the true model $\delta = \delta_1 \cup \delta_2$ under the theorem assumptions. We remark that there are simple examples where the ALA asymptotic model $\tilde{\gamma}^* \neq \delta$. For instance, for $E_{F^*}(y_i \mid z_i) = z_{i1}^2$ we have that truly $\delta = \{1\}$ but $\tilde{\beta}^*_j = 0$, see also the Poisson example in Section S12.2. In practice, however, in most of our examples we observed that the ALA-based $\tilde{\gamma}^*$ largely coincides with the model-based γ^*.

We next prove that ALA asymptotically recovers $\tilde{\gamma}^*$, and give the associated rates.

4.1 | Finite-dimensional problems

The assumptions to obtain ALA Bayes factor rates are minimal. For any model $\gamma \in \Gamma$ and $\tilde{\beta}^*_\gamma$ in Equation (19), we assume the following conditions.

1. $(y_i, z_i) \sim F^*$ independently for $i = 1, \ldots, n$, with finite positive-definite $\Sigma_{z\gamma} = \mathrm{Cov}_{F^*}(z_{i\gamma})$ and finite $\Sigma_{y|z,\gamma} = \mathrm{diag}(\mathrm{Cov}_{F^*}(y_1 \mid Z_\gamma), \ldots, \mathrm{Cov}_{F^*}(y_n \mid Z_\gamma))$.
2. The matrix $L_\gamma = E_{F^*}(z_{i\gamma} z_{i\gamma}^T [E_{F^*}(y_i \mid z_{i\gamma}) - z_{i\gamma}^T \beta^*_\gamma]^2)$ has finite entries.
3. The prior density $p^L(\beta_\gamma \mid \phi, \gamma)$ is continuous and strictly positive at $\tilde{\beta}^*_\gamma$.
4. The equations $\phi^2 \sum_{i=1}^n c(y_i, \phi)/n = -b(0)$ and $\phi^2 E_{F^*}[\nabla_\phi c(y_i, \phi)] = -b(0)$ have unique roots ϕ_0 and ϕ_0^* respectively, where $E_{F^*}[\nabla_\phi c(y_i, \phi)] < \infty$.

Conditions (C1)–(C2) require that (y_i, z_i) have finite full-rank second-order moments. Condition (C3) states that $p^L(\beta_\gamma \mid \phi, \gamma)$ is a local prior assigning positive density to $\tilde{\beta}^*_\gamma$, Theorem 2 and Corollary 1 below give Bayes factor rates for such prior and also for the non-local $p^N(\beta_\gamma \mid \phi, \gamma)$ in Equation (9). Condition (C4) states that ϕ_0 and ϕ_0^* are the unique maximizers of the observed and expected log-likelihood under F^*, conditional on $\beta = 0$. (C4) is only used in Corollary 1 where ϕ is unknown to show that $\phi_0 \xrightarrow{P} \phi_0^*$, and can be easily replaced, should (ϕ, ϕ_0^*) not be unique. One may instead assume that $\log p(y|\beta = 0, \phi)$ defines a Glivenko-Cantelli class, a sufficient condition being that the log-likelihood is dominated by an integrable function under F^* (van der Vaart (1998), Theorems 5.7, 5.9 and Lemma 5.10).

Theorem 2 states that for any model γ adding spurious parameters (in the linear projection sense) to $\tilde{\gamma}^*$, then $\tilde{B}^L_{\gamma\tilde{\gamma}^*} = O_p(n^{-(p_\gamma - p_{\tilde{\gamma}^*})/2})$, for any local prior $p^L(\beta\,|\,\gamma)$. Hence ALA Bayes factors discard spurious parameters at a polynomial rate in n. The rates for the gMOM-based $\tilde{B}^N_{\gamma\tilde{\gamma}^*}$ are faster, akin to results on exact Bayes factors (Johnson & Rossell, 2010, 2012). In contrast, if γ misses active parameters, then $\tilde{B}^L_{\gamma\tilde{\gamma}^*}$ and $\tilde{B}^N_{\gamma\tilde{\gamma}^*}$ decrease exponentially in n. Corollary 1 extends Theorem 2 to the unknown ϕ case.

Theorem 2 *Assume Conditions (C1)–(C3). Let $\tilde{B}^L_{\gamma,\tilde{\gamma}^*}$ and $\tilde{B}^N_{\gamma,\tilde{\gamma}^*} = \tilde{p}^N(y\,|\,\phi,\gamma)/\tilde{p}^N(y\,|\,\phi,\tilde{\gamma}^*)$ be the ALA Bayes factor in Equations (11) and (16) when ϕ is known.*
(i) Suppose that $\tilde{\gamma}^* \subset \gamma$. Then $\tilde{B}^L_{\gamma\tilde{\gamma}^*} = n^{(p_{\tilde{\gamma}^*} - p_\gamma)/2} O_p(1)$ and $\tilde{B}^N_{\gamma\tilde{\gamma}^*} = n^{3(p_{\tilde{\gamma}^*} - p_\gamma)/2} O_p(1)$ as $n \to \infty$.
(ii) Suppose that $\tilde{\gamma}^* \not\subseteq \gamma$. Then $\frac{1}{n}\log\tilde{B}^L_{\gamma\tilde{\gamma}^*} < o_p(1) + W/n$ and $\frac{1}{n}\log\tilde{B}^N_{\gamma\tilde{\gamma}^*} < o_p(1) + W/n$ for a random variable W satisfying $W/n \xrightarrow{P} c > 0$, as $n \to \infty$.

Corollary 1 *Assume Conditions (C1)–(C4). Let $\tilde{B}^L_{\gamma,\tilde{\gamma}^*}$ and $\tilde{B}^N_{\gamma,\tilde{\gamma}^*}$ be the ALA Bayes factor corresponding to the local and non-local priors in Equations (12) and (17) when ϕ is unknown. Then the statements in Theorem 2 (i)–(ii) remain valid.*

The rates in Theorem 2 are of the same form, as a function of n, as those for standard (Dawid, 1999; Johnson & Rossell, 2010) and miss-specified Bayes factors (Rossell & Rubio, 2018, 2019). The main difference with such standard rates is in Part (ii). The leading term in $\tilde{B}_{\gamma,\tilde{\gamma}^*}$ is given by a random variable W that converges to a chi-square distribution with non-centrality parameter $l(\tilde{\gamma}^*, \gamma) = n(\tilde{b}^*)^T \tilde{S} \tilde{b}^* > 0$, where $\tilde{S}$ is a positive-definite matrix and $\tilde{b}^* \neq 0$ are asymptotic partial regression coefficients for columns in $\tilde{\gamma}^* \setminus \gamma$ (see the proof for details). In contrast, the leading term in exact B_{γ,γ^*} has a different non-centrality parameter $n(b^*)^T S b^*$, where (b^*,S) now depend on KL projections. That is, although $p(\gamma|y)$ and $\tilde{p}(\gamma\,|\,y)$ are both exponentially fast in n at discarding models γ that miss truly active parameters in γ^* and $\tilde{\gamma}^*$ respectively, the coefficients governing these rates may change. For instance, if the exponential family model (5) is well-specified, even when $\tilde{\gamma}^* = \gamma^*$ one expects $\tilde{B}_{\gamma,\tilde{\gamma}^*}$ to have lower statistical power than $B_{\gamma\gamma^*}$ to detect active parameters. In contrast, if (5) is misspecified and $E_{F^*}(y_i\,|\,z_i)$ is better approximated by a linear function of z_i than by (4), then one expects $\tilde{B}_{\gamma,\tilde{\gamma}^*}$ to attain higher asymptotic power.

From a technical point of view a contribution of Theorem 2 relative to earlier results is that, by building upon parameter estimation results for concave log-likelihoods in Hjort and Pollard (2011), it requires near-minimal technical conditions.

4.2 | High-dimensional problems

Our main result proves that $\tilde{p}(\tilde{\gamma}^*\,|\,y) \xrightarrow{L_1} 1$ as $n \to \infty$ and provides the associated convergence rates. Recall that $\tilde{\gamma}^*$ is the ALA-optimal model, where $\tilde{\gamma}_j = I(\tilde{\beta}^*_j \neq 0)$ indicates zeroes in the ALA-optimal parameter $\tilde{\beta}^*$ in Equation (19). By definition of L_1 convergence, this is equivalent to $E_{F^*}\sum_{\gamma \neq \tilde{\gamma}^*} \tilde{p}(\gamma\,|\,y)$ converging to 0. L_1 convergence guarantees the asymptotic control of certain frequentist model selection probabilities. Let $\hat{\gamma} = \arg\max \tilde{p}(y\,|\,\gamma)$ be the highest posterior probability model, then $P_{F^*}(\hat{\gamma} \neq \tilde{\gamma}^*) \leq 2(E_{F^*}[\tilde{p}(\tilde{\gamma}^*\,|\,y)] - 1)$ (Rossell (2018), Proposition 1). The same bound applies to the family-wise type I-II error probabilities, and when setting $\hat{\gamma}$ to be the median probability model of Barbieri and Berger (2004) (Rossell (2018), Corollary 1).

Our main assumption is that $\tilde{y}$ has sub-Gaussian tails with variance parameter σ^2 under F^*, see Definition 1 in the supplement. We use the assumption to derive novel bounds on integrated

tail probabilities of sub-Gaussian quadratic forms, see Propositions S1 and S2, which may have some independent interest. The assumption is satisfied for example if F^* has Gaussian tails or $\tilde{y}$ is bounded as in logistic, multinomial or ordinal regression, and in fact allows for dependence in $\tilde{y}$, but is not satisfied when $\tilde{y}$ has thick tails such as the Poisson distribution. One may extend Propositions S1 and S2 to thicker-tailed F^*, then the L_1 rates could be slower than those presented here. For simplicity we focus on the known dispersion parameter ϕ case, non-random design matrix Z and Zellner's prior $p^L(\beta_\gamma \mid \phi, \gamma) = \mathcal{N}(\beta_\gamma; 0, (g_L\phi/n)(Z_\gamma^T Z_\gamma)^{-1})$. Our proofs can be extended to unknown ϕ and other priors, see the proof for a discussion, at the expense of more involved arguments and technical conditions. By default we recommend setting constant g_L, say $g_L = 1$ as in Section 2.2, but our results allow for g_L to change with n. For example, Narisetty and He (2014) proposed letting g_L grow with n to obtain sparser solutions, whereas proceeding analogously to the uniformly most powerful tests of Johnson (2013) one might let g_L decrease with n to improve power.

Theorem 3 below provides a first result on pairwise Bayes factors, specifically on

$$E_{F^*}\left(\left[1 + \tilde{B}_{\tilde{\gamma}^*\gamma}\frac{p(\tilde{\gamma}^*)}{p(\gamma)}\right]^{-1}\right),$$

that is the posterior probability assigned to γ if one only considered the models γ and $\tilde{\gamma}^*$. Bounding this quantity also bounds the rate at which $\tilde{B}_{\gamma\tilde{\gamma}^*} \xrightarrow{P} 0$, hence Theorem 3 extends Theorem 2 to high dimensions. Theorem 4 is our main result characterizing $\tilde{p}(\tilde{\gamma}^* \mid y)$.

We first interpret Theorem 3, subsequently discuss the required technical conditions and finally state the theorem. Part (i) says that models adding spurious parameters to $\tilde{\gamma}^*$ are discarded at the same polynomial rate in n (up to log terms) as in the fixed p case,

$$r_\gamma = \left(ng_L\right)^{\frac{p_\gamma - p_{\tilde{\gamma}^*}}{2}} \frac{p(\tilde{\gamma}^*)}{p(\gamma)}.$$

This rate holds when the model-predicted variance $V(\tilde{y}_i \mid z_i,\ \beta = 0,\ \phi) = \phi/b''(0) > \sigma^2$, that is data under F^* are under-dispersed. Alternatively, if $\phi/b''(0) < \sigma^2$ (over-dispersion) then a (slower) rate r_γ^a is attained, where $a = \phi/[b''(0)\sigma^2] < 1$. The intuition is that, if the model underestimates σ^2 then it becomes easier to add spurious parameters to $\tilde{\gamma}^*$. Part (ii) states that models missing active parameters are discarded at an exponential rate in the non-centrality parameter

$$\lambda_\gamma = (Z_{\tilde{\gamma}^*}\beta^*_{\tilde{\gamma}^*})^T(I - H_\gamma)Z_{\tilde{\gamma}^*}\beta^*_{\tilde{\gamma}^*} \tag{21}$$

where $H_\gamma = Z_\gamma(Z_\gamma^T Z_\gamma)^{-1}Z_\gamma^T$ is the projection matrix onto the column space of Z_γ. For simplicity the result raises the rate at a constant power b arbitrarily close to 1, but one can actually take $b = 1$ and add logarithmic terms, see the proof for details.

The parameter λ_γ has a simple interpretation, it is the reduction in mean squared error when one approximates $E_{F^*}(y \mid Z)$ with $Z_{\tilde{\gamma}^*}\beta^*_{\tilde{\gamma}^*}$, relative to $Z_\gamma\beta^*_\gamma$. A common strategy in high-dimensional model selection theory is to assume conditions on the eigenvalues of Z^{TZ} to lower-bound λ_γ in terms of $\beta^T_{\tilde{\gamma}^*}\beta_{\tilde{\gamma}^*}$. Instead, here we give the result directly in terms of λ_γ, and state near-minimal conditions on λ_γ required for the result to hold. To build intuition, however, in a simplest case

where the columns in $\tilde{\gamma}^* \setminus \gamma$ are uncorrelated with those in γ, then $\lambda = (\beta^*_{\tilde{\gamma}^* \setminus \gamma})^T Z^T_{\tilde{\gamma}^* \setminus \gamma} Z_{\tilde{\gamma}^* \setminus \gamma} \beta^*_{\tilde{\gamma}^* \setminus \gamma}$ and one can roughly think of λ_γ as being linear in n.

The technical conditions required for Theorem 2 and any model $\gamma \in \Gamma$ are below. For two sequences a_n, b_n, $a_n \ll b_n$ denotes that $\lim_{n\to\infty} a_n/b_n = 0$.

D1 There exists a finite $\sigma^2 > 0$ such that $\tilde{y} - Z_{\tilde{\gamma}^*} \tilde{\beta}^*_{\tilde{\gamma}^*} \sim \mathrm{SG}(0, \sigma^2)$, where $\tilde{\beta}^*_{\tilde{\gamma}^*}$ is as in Equation (19).

D2 $Z_\gamma^T Z_\gamma$ is invertible.

D3 For any $\gamma \supset \tilde{\gamma}^*$,

$$\log(n g_L) + \frac{2}{p_\gamma - p_{\tilde{\gamma}^*}} \log\left(\frac{p(\tilde{\gamma}^*)}{p(\gamma)}\right) \gg 1. \tag{22}$$

For any $\gamma \not\supset \tilde{\gamma}^*$ of dimension $p_\gamma \geq p_{\tilde{\gamma}^*}$,

$$(p_{\tilde{\gamma}^*} - p_\gamma)\log\left(n g_L\right) + \log\left(\frac{p(\gamma)}{p(\tilde{\gamma}^*)}\right) + p_\gamma \ll \frac{\lambda_\gamma}{\log(\lambda_\gamma)}. \tag{23}$$

For any $\gamma \not\supset \tilde{\gamma}^*$ of dimension $p_\gamma < p_{\tilde{\gamma}^*}$,

$$(p_{\tilde{\gamma}^*} - p_\gamma)\log\left(n g_L\right) + \log\left(\frac{p(\gamma)}{p(\tilde{\gamma}^*)}\right) + p_{\tilde{\gamma}^*} \ll \frac{\lambda_\gamma}{\log(\lambda_\gamma)}. \tag{24}$$

As discussed earlier, Condition (D1) states that the data-generating F^* satisfies a tail property, specifically that the ALA-optimal errors $\tilde{y} - Z^*_{\tilde{\gamma}} \tilde{\beta}^*_{\tilde{\gamma}^*}$ are no thicker than sub-Gaussian. (D2) can be relaxed but ensures that $p^L(\beta_\gamma \mid \phi, \gamma) = \mathcal{N}(\hat{\beta}_\gamma; 0, (g_L \phi/n)(Z_\gamma^T Z_\gamma)^{-1})$ is proper. (D2) requires that $p_\gamma \leq n$, as discussed in Section 2 one may have $p \gg n$ but only models with up to $p_\gamma < n$ parameters receive positive prior probability $p(\gamma) > 0$. (D3) are minimal conditions on the prior parameters and the signal strength λ_γ. If the number of truly active groups $|\tilde{\gamma}^*| < J/2$, where J is the number of total groups, $p(|\gamma|)$ in Equation (8) is non-increasing in $|\gamma|$, and g_L is non-decreasing in n, then (22) and (23) hold. (D3) states weaker, near-necessary assumptions for pairwise $B_{\gamma\gamma^*}$ convergence. Also, for $p(\gamma)$ in Equation (8), a sufficient condition for Equation (24) is that

$$p_{\tilde{\gamma}^*} \log(g_L n) + |\tilde{\gamma}^*| \log J \ll \frac{\lambda_\gamma}{\log \lambda_\gamma},$$

that is (D3) allows the number of parameters $p_{\tilde{\gamma}^*}$ (and groups $|\tilde{\gamma}^*|$) in $\tilde{\gamma}^*$ to grow near-linearly in λ_γ, and the total number of groups J to grow near-exponentially.

Theorem 3 *Let $p^L(\beta_\gamma \mid \phi, \gamma) = N(\beta_\gamma; 0, (g_L \phi/n)(Z_\gamma^T Z_\gamma)^{-1})$, where $\phi > 0$ is fixed, and assume Conditions D1–D3.*

(i) Suppose that $\tilde{\gamma}^* \subset \gamma$. If $\phi/b''(0) > \sigma^2$, then

$$E_{F^*}\left(\left[1+\tilde{B}^L_{\tilde{\gamma}^*\gamma}\frac{p(\tilde{\gamma}^*)}{p(\gamma)}\right]^{-1}\right) \leq \frac{2\max\left\{[(2/[p_\gamma - p_{\tilde{\gamma}^*}])\log r_\gamma]^{(p_\gamma - p_{\tilde{\gamma}^*})/2}, \log(r_\gamma)\right\}}{r_\gamma},$$

for all $n \geq n_0$, and some fixed n_0. Further, if $\phi/b''(0) \leq \sigma^2$, then

$$E_{F^*}\left(\left[1+\tilde{B}^L_{\tilde{\gamma}^*\gamma}\frac{p(\tilde{\gamma}^*)}{p(\gamma)}\right]^{-1}\right) \leq \frac{2.5\max\left\{\left[(2/[p_\gamma - p_{\tilde{\gamma}^*}])\log(r_\gamma^{\frac{a}{1+\epsilon}})\right]^{(p_\gamma - p_{\tilde{\gamma}^*})/2}, \log\left(r_\gamma^{\frac{a}{1+\epsilon}}\right)\right\}}{r_\gamma^{\frac{a}{(1+\epsilon\sqrt{2}(1+\sqrt{2}))}}}$$

for all $n \geq n_0, \epsilon = 1/\sqrt{a\log r_\gamma}$, $a = \phi/[b''(0)\sigma^2]$.

(ii) Suppose that $\tilde{\gamma}^* \not\subset \gamma$. For any $b < 1$ there exists a finite n_0 such that, for all $n \geq n_0$,

$$E_{F^*}\left(\left[1+\tilde{B}^L_{\tilde{\gamma}^*\gamma}\frac{p(\tilde{\gamma}^*)}{p(\gamma)}\right]^{-1}\right) \leq \left(\frac{p(\gamma)e^{-\frac{\lambda_\gamma}{2\max\{\phi/b''(0),\sigma^2\}\log(\lambda_\gamma)}}}{p(\tilde{\gamma}^*)(ng_L)^{\frac{(p_\gamma - p_{\tilde{\gamma}^*})}{2}}}\right)^b$$

Theorem 4 below states that the posterior probability $\tilde{p}(\tilde{\gamma}^*|y)$ converges to 1. Separate rates are given for the set of models $S = \{\gamma: \tilde{\gamma}^* \subset \gamma\}$ containing spurious parameters and its complement $S^c = \{\gamma: \tilde{\gamma}^* \not\subset \gamma\}$. The result assumes the model prior in Equation (8), with $p(|\gamma|) \propto p^{-c|\gamma|}$, for $c \geq 0$. Recall that in the canonical case with no groups nor hierarchical constraints, $c = 0$ reduces to the Beta-Binomial(1, 1) prior and $c > 0$ to a Complexity prior. Theorem 2 requires a technical condition.

D4 Let $a = 1$ if $\phi/b''(0) > \sigma^2$ and $a < \phi/[b''(0)\sigma^2]$ otherwise, $q = \min\{p_j: \gamma_j^* = 0\}$ be the smallest spurious group size and $q' = \max\{p_j\}$. Then, assume that

$$(|\tilde{\gamma}^*| + 1)(\overline{J} - |\tilde{\gamma}^*|) \ll (ng_L)^{aq/2}p^{ca+a-1}. \tag{25}$$

Further, let $\underline{\lambda} = \min_{|\gamma|\leq|\tilde{\gamma}^*|}\lambda_\gamma/\max\{|\tilde{\gamma}^*| - |\gamma|, 1\}$ and $\overline{\lambda} = \min_{|\gamma|>|\tilde{\gamma}^*|,\gamma\not\subset\tilde{\gamma}^*}\lambda_\gamma$. Then

$$(|\tilde{\gamma}^*| + 2)\log J \ll \overline{\lambda} + c\log p + \frac{q}{2}\log(ng_L) \tag{26}$$

$$\frac{q'}{2}\log(ng_L) + c\log p + (|\tilde{\gamma}^*| + 1)\log J \ll \underline{\lambda} \tag{27}$$

Condition (D4) ensures that (D3) holds uniformly across models $\gamma \in \Gamma$. It is stated in terms of $(\underline{\lambda}, \overline{\lambda})$ bounding the non-centrality parameter λ_γ for models of smaller and larger size than $\tilde{\gamma}^*$ groups, respectively. Expressions (25)–(26) impose mild assumptions on the number of active groups $|\tilde{\gamma}^*|$, total groups J and largest number of allowed groups $\overline{J}$. These expressions guarantee

that models of size $|\gamma| > |\tilde{\gamma}^*|$ receive vanishing probability, and in particular they are satisfied when setting $c > 0$. Expression (27) ensures that models of size $|\gamma| \leq |\tilde{\gamma}^*|$ receive vanishing probability, and imposes an upper-bound on c in terms of the signal strength $\underline{\lambda}$.

Theorem 4 *Let $p^L(\beta_\gamma|\phi,\gamma) = N(\beta_\gamma;0,(g_L\phi/n)(Z_\gamma^T Z_\gamma)^{-1})$, for known $\phi > 0$, and $p(\gamma)$ as in Equation (8) with $p(|\gamma|) = p^{-c|\gamma|}$, $c \geq 0$.*

(i) Suppose that Conditions D1, D2 and D4 hold, and that $\lim_{n\to\infty} |\tilde{\gamma}^*|/[p^c(ng_L)^{q/2}] = 0$. Then there is a finite n_0 such that, for all $n \geq n_0$ and all $\varepsilon > 0$, if $\phi/b''(0) < \sigma^2$ then

$$E_{F_0}\left(\tilde{p}^L(S|y)\right) \leq \frac{(|\tilde{\gamma}^*|+1)(\bar{J}-|\tilde{\gamma}^*|)(\log(ng_L^{q/2}) + [c+1]\log(p)+\epsilon)}{p^c(ng_L)^{q/2}}.$$

Further, if $\phi/b''(0) < \sigma^2$, then

$$E_{F_0}\left(\tilde{p}^L(S|y)\right) \leq \frac{(|\tilde{\gamma}^*|+1)(\bar{J}-|\tilde{\gamma}^*|)[\log\left((ng_L)^{q/2}\right) + (c+1)\log(p)+\epsilon]}{p^{a(c+1)-1}(ng_L)^{aq/2}},$$

where $a < 1$ is a constant smaller than but arbitrarily close to $\phi/[b''(0)\sigma^2]$.

(ii) Suppose that (D1), (D2) and (D4) hold. Then there is a finite n_0 such that, for all $n \geq n_0$, the posterior probability assigned to $S^c = \{\gamma : \tilde{\gamma}^* \not\subseteq \gamma\}$ satisfies

$$E_{F_0}\left(\tilde{p}^L(S^c|y)\right) \leq \frac{(|\tilde{\gamma}^*|+1)e^{(|\tilde{\gamma}^*|+2)\log J}}{[e^{\underline{\lambda}}p^c(ng)^{q/2}]^b} + \frac{e^{|\tilde{\gamma}^*|\log J}}{e^{\underline{\lambda}}} + \frac{e^{|\tilde{\gamma}^*|\log J}}{e^{(|\tilde{\gamma}^*|+1)[\underline{\lambda}-c\log(p)-(q'/2)\log(ng)]}}.$$

The three terms in the bound for $E_{F_0}\left(\tilde{p}^L(S^c|y)\right)$ correspond to the posterior probability assigned models with $|\gamma| > |\tilde{\gamma}^*|, |\gamma| = |\tilde{\gamma}^*|$ and $|\gamma| < |\tilde{\gamma}^*|$, respectively. That is, Theorem 2 does not only characterize the posterior probability $\tilde{p}(\tilde{\gamma}^*|y)$ assigned to the ALA-optimal model $\tilde{\gamma}^*$, but also to other interesting model space subsets: those adding spurious parameters to $\tilde{\gamma}^*$, and those missing parameters with either smaller/larger size than $|\tilde{\gamma}^*|$. These rates reflect that sparse priors, for example Complexity priors ($c > 0$) or diffuse priors (g_L grows with n), are faster at discarding models larger than $|\tilde{\gamma}^*|$. The trade-off is that they attain slower rates for models of size $|\gamma| < |\tilde{\gamma}^*|$, that is they have lower statistical power to discard small models missing truly active parameters.

5 | RESULTS

We illustrate the performance of ALA in terms of numerical accuracy, computation time and quality of the model selection inference in simulated and empirical data. Section 1 shows the computational scalability in logistic and Poisson regression, as either n or p grow. We consider the default ALA at $\beta_0 = 0$ and refined versions where $\beta_{\gamma 0}$ is obtained by 1 and 2 Newton-Raphson iterations starting at zero, respectively. We also consider the combined use of ALA with importance sampling to obtain samples from the exact posterior. Section 5.2 studies the accuracy of ALA under the non-local gMOM prior in Equation (15). For this prior, ALA are faster and often more precise than LA, particularly for models that include spurious covariates. We also compare the gMOM ALA model selection performance in a non-linear regression example to exact

gZellner calculations, the group LASSO (Bakin, 1999), group SCAD (Fan & Li, 2001) and group MCP (Zhang, 2010). In Section 5.3 we analyze a poverty dataset that has a binary outcome and large n and p. Finally, Section 5.4 shows survival examples where the likelihood lies outside the exponential family, but is nevertheless log-concave, making it amenable to the ALA.

In all examples we used the gMOM and gZellner priors with default $g_N = 1$ and $g_L = 1$ parameters and the Beta-Binomial prior on models, truncated to models satisfying the hierarchical constraints when required (Section 2). To ensure that run times between LA and ALA are comparable, we implemented both in C++ in R package `mombf`. For problems with $>10^6$ models where full enumeration was unfeasible, we used the augmented Gibbs sampling algorithm from Rossell and Rubio (2019). We used the software defaults producing 10,000 full Gibbs scans and no parallel computing, hence our run times are a conservative figure relative to those potentially attainable with more advanced model search strategies. Packages `grplasso` and `grpreg` (Breheny & Huang, 2015) were were used to implement group LASSO, group SCAD and group MCP.

5.1 | Simulations for logistic and Poisson regression

We considered simulated examples where the data are truly generated from logistic and Poisson models with linear predictor $\beta^* = (0, \ldots, 0, 0.5, 1)$ and $z_i \sim \mathcal{N}(0, \Sigma)$, $\Sigma_{ii} = 1$, $\Sigma_{ij} = 0.5$, and we set the group Zellner prior in Equation (9). We consider a first setting with fixed $p = 10$ and $n \in \{100, 500, 1000, 5000\}$, and a second setting with fixed $n = 500$ and $p \in \{5, 10, 25, 50\}$.

Figure 2 summarizes the logistic regression results, and Figure S1 those for Poisson regression. ALA at $\beta_0 = 0$ significantly reduced run times for larger n or p. In terms of the resulting inference, ALA and LA attained consistency as n grows (bottom left) and discriminateed truly active versus inactive variables, even for larger p (bottom right). Figure S1 also shows that ALA Bayes factors that do not incorporate the over-dispersion curvature adjustment in Equation (13) led to assigning significantly higher inclusion probabilities to truly spurious parameters, even for fairly large n.

We also applied ALA with $\beta_{\gamma 0}$ given by 1 and 2 Newton-Raphson iterations from zero (Figure 2). This refinement gave a closer approximation to the LA posterior, at a non-negligible computational cost, particularly in scalability as n grows.

Finally, we studied the use of ALA as a tool to identify promising models that can be subsequently refined with exact methods. Specifically, we used importance sampling to re-weight models sampled from the ALA posterior. Section S12.2 offers a full description. Briefly, from our theory, the ALA and LA posteriors in general concentrate on two different models $\tilde{\gamma}^*$ and γ^* respectively, resulting in degenerate importance weights as $n \to \infty$. However, in practice for finite n there are situations where importance sampling is effective; see the logistic regression example in Section S12.2. In other cases, such as the Poisson example in Section S12.2, ALA is useful to screen out certain truly inactive covariates, but cannot be directly combined with importance sampling. Recall that Theorem 1 gives conditions where ALA asymptotically recovers or screens out the correct parameters. More generally, combining ALA with exact strategies is an interesting avenue for future research that deserves a separate treatment elsewhere.

5.2 | Simulations under Gaussian outcomes

Even for Gaussian outcomes the marginal likelihood under the gMOM prior requires a number of operations growing exponentially with model size (Kan, 2008). Section 5.2.1 illustrates that

ALA not only provides faster integrated likelihoods than the LA, but that it can also be more precise. Section 5.2.2 compares ALA-based gMOM model selection versus exact calculations under the group-Zellner prior and three penalized likelihood methods.

5.2.1 | Numerical accuracy under non-local priors

Consider an example with $p = 10$, $n \in \{50, 200, 500\}$ and truly $y \sim \mathcal{N}(Z\beta^*, I)$, where $\beta^* = (0.4, 0.6, 1.2, 0.8, 0, \ldots, 0)$. The rows in Z are independent draws $z_i \sim \mathcal{N}(0, I)$ or, alternatively, $z_i \sim \mathcal{N}(0, V)$ for random non-diagonal V. Specifically, V is the correlation matrix associated to W^{TW}, where W is a $p \times p$ matrix with $w_{ij} \sim \mathcal{N}(0, 1)$ independently across i, j. We evaluated the integrated likelihood $p^N(y \mid \gamma)$ under the gMOM prior for a sequence of models including $p_\gamma = 1, 2, \ldots, 10$ covariates. For $p_\gamma \leq 10$ one can evaluate $p^N(y \mid \gamma)$ exactly, and hence the error when estimating $\widehat{p}^N(y \mid \gamma)$. We report average errors across 100 simulations.

Figure 3 summarizes the results for the $n = 50$ case. The left panel shows the mean of $\log(\widehat{p}^N(y \mid \gamma)/p^N(y \mid \gamma))$, which quantifies the relative approximation error. Both LA and ALA provided fairly accurate estimates for models including up to 4 covariates. Note that β^* is such that for $p_\gamma \leq 4$ all included covariates are truly active. For models with > 4 covariates the LA error was significantly higher than for ALA, particularly in the non-diagonal covariance setting. The right panel in Figure 3 illustrates the difficulty of the integration exercise by plotting the contours of the log-integrand $p(y \mid \beta, \phi = 1)p^N(\beta \mid \phi = 1, \gamma)$ versus two truly spurious parameters (β_5, β_6) in a randomly selected dataset. The marked multi-modality, in general, does not disappear even as $n \to \infty$. The results for $n = 200$ and $n = 500$ were largely analogous, see Figure S4.

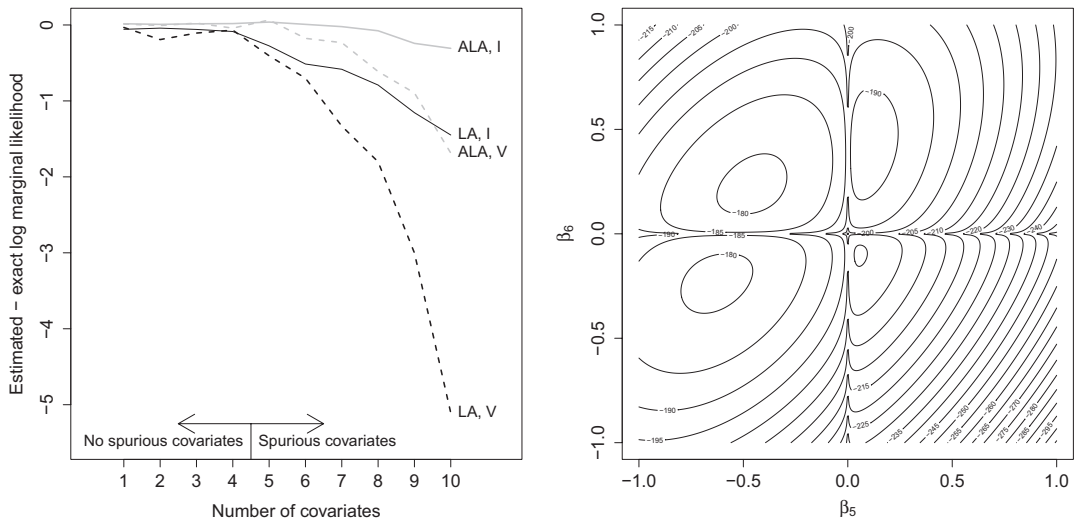

FIGURE 3 Simulated linear regression, gMOM prior ($n = 50$, $p = 10$, $\boldsymbol{\beta} = (0.4, 0.6, 1.2, 0.8, 0, ..., 0)$, $\phi = 1$). Left: mean error $\log\tilde{p}^N(y \mid \gamma) - \log p^N(y \mid \gamma)$ for $x_i \sim \mathcal{N}(0, I)$ and $x_i \sim \mathcal{N}(0, V)$, random non-diagonal V. Right: log-integrand $p(y \mid \beta, \phi)p^N(\beta \mid \gamma)$ contours versus two spurious parameters for a randomly-selected dataset, $x_i \sim \mathcal{N}(0, V)$

5.2.2 | Group constraints in non-linear regression

We present a simulation example where one incorporates group constraints to model non-linear covariate effects. See Section S12.3 for examples where groups are defined by a categorical covariate. We considered the following data-generating truth

$$y_i = \sin(-x_{i1} + 0.1) + \sin(2.5x_{i2}) + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(0, 1)$. We considered scenarios where one observes a total of 5 and 50 covariates (including the two truly active covariates), generated from a multivariate Normal with zero mean, unit variances and 0.5 pairwise correlations.

Suppose that the data analyst poses an additive model that considers non-linear effects but, unaware that the true expectation of y_i depends on sin functions, misspecifies their form. Specifically, the assumed model decomposes covariate effects into a linear plus a deviation-from-linearity component via a 5-dimensional cubic splines, following Rossell and Rubio (2019). That is, each covariate is coded into the design matrix via one column for its linear effect and a five-variable group capturing deviations from linearity. Hence the scenarios with 5 and 50 covariates result in $J = 10$ and $J = 100$ groups (respectively) and in $p = 30$ and $p = 300$ total parameters (respectively).

Figure 4 reports the proportion of correct model selections across 150 simulations. The regularization parameter for grLASSO, grMCP and grSCAD was set via 10-fold cross-validation. In all settings the proportion of correct model selections were highest for either the exact gZellner calculation (particularly for smaller n) or the ALA-based gMOM prior (for larger n), showing that the latter leads to high-quality approximate inference.

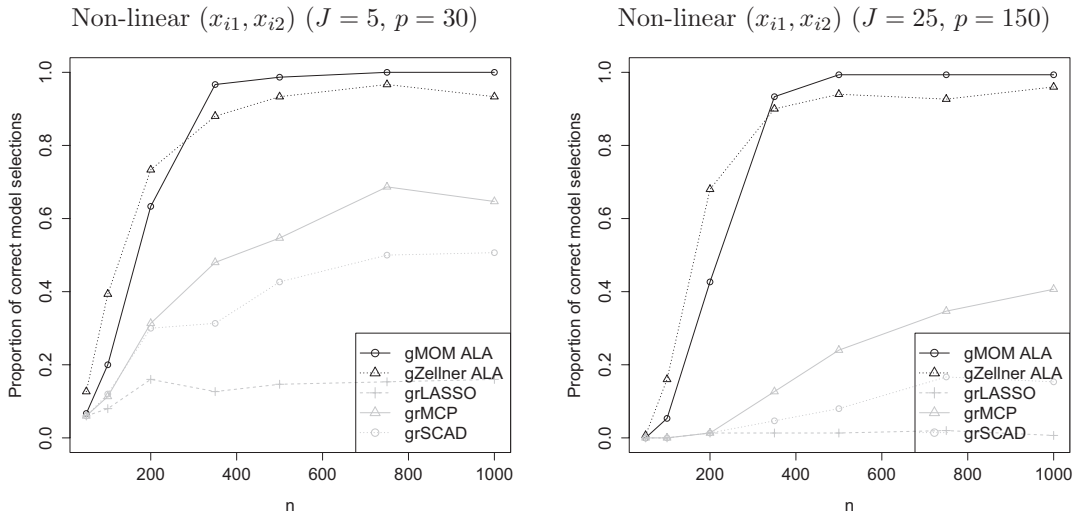

FIGURE 4 Proportion of correct model selections in Gaussian simulations with non-linear effects for two truly active covariates and a total of $J = 5$ and $J = 25$ covariates ($p = 30$ and 150, respectively)

5.3 | Poverty line

We studied what factors are associated to individuals working full-time being below the poverty line in the USA. We used a large dataset from the Current Population Survey (Flood et al., 2020) conducted in 2010 and 2019 for single-race individuals aged 18–65 years who were non-military employed for 35–40 h/week.

The response is a binary indicator $y_i \in \{0, 1\}$ for individual i being below the poverty line. The covariates x_i include gender and hispanic origin indicators, race, marital status, level of education, citizenship status, nativity status, occupation sector, size of the firm employing the individual, the presence of impairment/difficulties, type of employment, moving to another state from within or outside the USA, and the weekly hours worked (35–40). Many of these variables are categorical, see Section S12.4 for a description. The data has $n = 89{,}755$ individuals. In a first exercise we considered a logistic regression analysis with only main effects, where $p = 60$. Subsequently we considered pairwise interactions between all covariates, then the corresponding design matrix Z has $p = 1469$ columns.

We first discuss the computation times. In the main effects analysis LA took > 10 h to run and ALA took 19.5 s. For comparison, GLASSO took 42.7 s. When adding interactions LA took 17.3 days, ALA 5.2 min and grLASSO 10.7 min.

In the main effects analysis ALA and LA selected the same model, with virtually identical marginal posterior inclusion probabilities (the correlation between $\widehat{p}(\gamma \,|\, y)$ and $\tilde{p}(\gamma \,|\, y)$ was > 0.999). All main effects had posterior inclusion probability > 0.95, except for nativity status, with posterior probability < 0.02. To help interpret the results, Table S3 provides the estimated coefficients. Briefly, higher poverty odds were estimated for females, hispanics, blacks and native Americans, individuals with difficulties, lower education levels, non-citizens, working in small firms, having moved from outside the US, and working in sectors such as farming or maintenance. The grLASSO results were similar, except that it selected all main effects, including nativity status (both when setting the penalization parameter via cross-validation or to minimize the BIC). For comparison, the P-value obtained from a maximum likelihood fit under the full model was 0.1959 for nativity, and < 0.0001 for all other main effects.

Regarding the analysis with interactions, LA and ALA selected the same 13 main effects (inclusion probability > 0.5, Table S4). LA selected 5 interaction terms and ALA selected 8. Both selected the interaction of gender vs. marital status, education level vs. hispanic origin, and hispanic origin vs. marital status. LA also selected education vs firm size, whereas ALA selected education vs marital status, and moving state vs hispanic origin, marital status and education level. The ALA- and LA-estimated marginal posterior probabilities $\tilde{p}(\gamma \,|\, y)$ and $\widehat{p}(\gamma \,|\, y)$ had 0.827 correlation.

Table S5 displays parameter estimates. We refrain from making any causal interpretation of these findings, but they suggest interesting future research to better understand poverty. For instance, gender and hispanic origin were associated with poverty, but the differences between hispanic and non-hispanic males was larger (estimated odds-ratio=1.63) than between hispanic and non-hispanic females (odds-ratio= 1.36). As another example, the odds of poverty were similar for non-hispanic married males and divorced males (odds-ratio=0.92), but married females had significantly lower odds than divorced females (odds-ratio=0.281) and married hispanics had higher odds than divorced hispanics (odds-ratio= 1.46)

For comparison, the grLASSO selected 4 main effects and 16 interaction terms. For only 1 out of the 16 interactions the two corresponding main effects were also selected, illustrating the need to explicitly enforce hierarchical restrictions.

5.4 | Cancer survival

We illustrate the ALA in models outside the exponential family via the non-linear additive survival model from Rossell and Rubio (2019), a spline-based log-normal accelerated failure time (AFT) model. If one re-parameterizes the model by dividing the regression parameters by the error standard deviation, then the log-likelihood is concave (Silvapulle & Burridge, 1986), hence amenable to be analyzed via ALA. We present a simulation study, and analyze a colon cancer dataset in Section S12.5.

We compared the ALA results obtained under gMOM and gZellner priors to the LA under the gZellner prior, the semi-parametric AFT model with LASSO penalties of Rahaman-Khan and Shaw (2019) (AFT-LASSO), and to the Cox model with LASSO penalties of Simon et al. (2011) (Cox-LASSO). For AFT-LASSO and Cox-LASSO we used functions `AEnet.aft` and `glmnet` in R packages `AdapEnetClass` and `glmnet`, and we set the penalization parameter via 10-fold cross-validation.

The simulation study extends that in Rossell and Rubio (2019) (section 6.2). Briefly, there are two covariates that truly have non-linear effects and 48 spurious covariates, generated from a zero-mean multivariate normal with unit variances and 0.5 correlation between all covariate pairs. The assumed model poses $E(y_i \mid x_i) = \sum_{j=1}^{50} x_{ij}\beta_j + \sum_{j=51}^{100} z_{ij}^T \beta_j$, where $x_{ij} \in \mathbb{R}$ and $z_{ij} \in \mathbb{R}^5$ captures deviations from linearity by projecting x_{ij} onto a spline basis and orthogonalizing the result to x_{ij}. We used a five-dimensional spline basis given that Rossell and Rubio (2019) found that larger dimensions gave very similar results. In our notation, there are $J = 100$ groups and $p = 50 + 250 = 300$ parameters.

We considered two data-generating truths. In Scenario 1 the truth is an accelerated failure time model and in Scenario 2 a proportional hazards model. Both scenarios are challenging in that there is a significant amount of censored data, which effectively reduces the information in the likelihood, and the dimension is moderately high.

- Scenario 1. Log-survival times are $x_{i1} + 0.5\log(|x_{i2}|) + \epsilon_i$, where $\varepsilon_i \sim N(0, \sigma = 0.5)$. All log-censoring times are 0.5, giving an average of 69% censored individuals.
- Scenario 2. Let h_0 be the log-Normal(0,0.5) baseline hazard. Log-survival times arise from a proportional hazards structure $h(t) = h_0(t)\exp\left\{3x_{i1}/4 - 5\log(|x_{i2}|)/4\right\}$. All log-censoring times are 0.55, giving an average of 68% censored individuals.

In Scenario 1 the assumed model is well-specified, except for approximating the non-linear truth by a finite spline basis, whereas in Scenario 2 the whole hazard function is misspecified. Figure 5 shows the proportion of correct model selections, across 250 independent simulations for $n \in \{100, 500\}$. Generally, ALA showed a competitive performance. The LA selected the data-generating model slightly more frequently under $n = 100$ and the well-specified Scenario 1 (upon inspection this was due to higher power to include x_{i1}), whereas it performed similar to ALA for larger n and in Scenario 2. For $n = 500$ the proportions of correct model selections for ALA were near 1. Both ALA and LA outperformed significantly AFT-LASSO and Cox-LASSO.

The ALA provided significant computational gains over LA. In most scenarios the computation time was reduced by a factor ranging from 20–70, see Table 1.

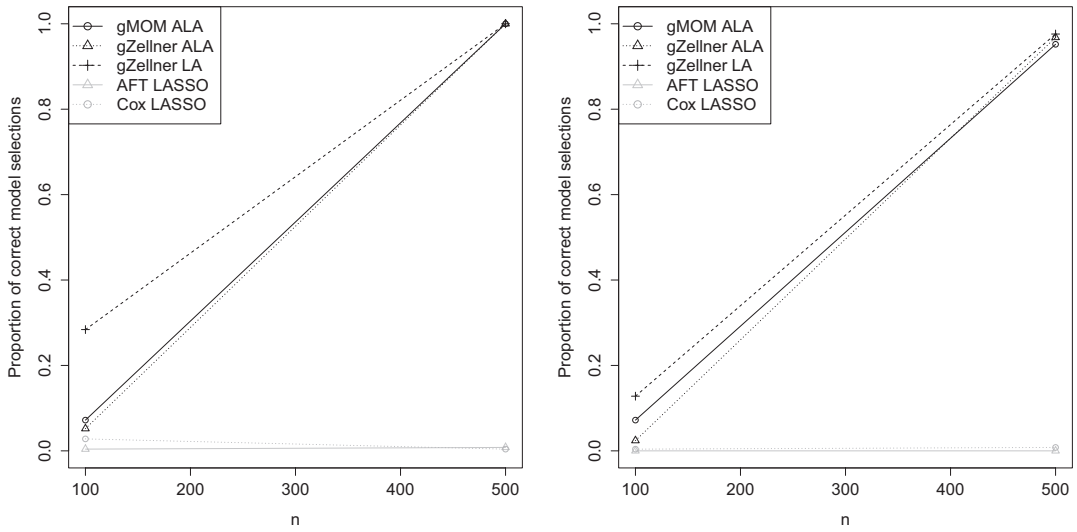

FIGURE 5 Non-linear survival regression ($J = 100$, $p = 300$). Proportion of correct model selections under a truly accelerated failure (left) and truly proportional hazards model (right)

6 | DISCUSSION

Our main contribution was proposing an approximate inference tool that can be particularly helpful in non-Gaussian regression, and in Gaussian outcomes where one wishes to use non-local priors. The proposal focuses on scoring models quickly for structural learning problems, and can be combined with parallel computing strategies to accelerate model search. The posterior probability rates require the same type of technical conditions than exact inference, and attain essentially the same functional rates in n. Importantly, ALA and exact calculations asymptotically recover, in general, different models. However we characterized situations where the ALA recovers the correct model, even under misspecification, and showed numerous examples where ALA results agree with exact inference. We also illustrated a significant applied potential in reducing computation times, enabling the use of Bayesian model selection to settings where it was previously impractical.

We focused our theory and examples on a simple strategy where Taylor expansions are taken around an $\eta_{\gamma 0}$ with zero regression coefficients. We also illustrated that $\eta_{\gamma 0}$ given by 1–2 Newton-Raphson steps typically improves precision, though computations scale more poorly with (n, p). In future work it may be interesting to study alternative choices of $\eta_{\gamma 0}$ that balance computational cost and accuracy and/or model selection properties. For instance, under concave log-likelihoods and minimal regularity conditions, it is possible to show that if $\eta_{\gamma 0} = \eta_{\gamma}^{*} + O_p(1/\sqrt{n})$, then the ALA provides a consistent estimator of the exact integrated likelihood.

The ALA can in principle be applied to any model, but one expects it to work best when the log-likelihood is concave or at least locally concave around $\eta_{\gamma 0}$, e.g. models satisfying local asymptotic normality. Yet another avenue is to apply ALA to conditionally concave settings in the spirit of INLA for latent Gaussian models (Rue et al., 2009).

From a foundational Bayesian point-of-view, a limitation of ALA is it they only provides approximate inference. It would be interesting to explore strategies to combine the ALA with exact computation, e.g. to build proposal distributions for MCMC algorithms, or sequential Monte

Carlo strategies as in Schäfer and Chopin (2013). Such extensions require care as both our theory and examples show that a naive combination of ALA and importance sampling can lead to degenerate weights. In summary, the current work provides a basis which we hope may lead the ground for multiple interesting extensions.

ORCID

David Rossell http://orcid.org/0000-0002-7104-5789

REFERENCES

Bakin, S. (1999) Adaptive regression and model selection in data mining problems. PhD thesis, The Australian National University, Canberra, Australia, 5.

Barbieri, M.M. & Berger, J.O. (2004) Optimal predictive model selection. *The Annals of Statistics*, 32(3), 870–897.

Bien, J., Taylor, J. & Tibshirani, R. (2013) A lasso for hierarchical interactions. *Annals of Statistics*, 41(3), 1111–1141.

Breheny, P. & Huang, J. (2015) Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, 25, 173–187.

Carbonetto, P. & Stephens, M. (2012) Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1), 73–108.

Castillo, I., Schmidt-Hieber, J. & van der Vaart, A.W. (2015) Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5), 1986–2018.

Dawid, A.P. (1999) The trouble with Bayes factors. Technical report, University College London.

Fan, J. & Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.

Flood, S., King, M., Rodgers, R., Ruggles, S. & Warren, J.R. (2020) Integrated public use microdata series, current population survey: Version 7.0 [dataset]. *Minneapolis, MN: IPUMS*, doi: http://dx.doi.org/10.18128/D030.V7.0

Griffin, J.E., Latuszynski, K.G. & Steel, M.F.J. (2020) In search of lost (mixing) time: adaptive Markov chain monte carlo schemes for Bayesian variable selection with very large p. *Biometrika*, 108(1), 53–69.

Hjort, N.L. & Pollard, D. (2011) Asymptotics for minimisers of convex processes. *arXiv, 1107.3806*, 1–24.

Huang, X., Wang, J. & Liang, F. (2016) A variational algorithm for Bayesian variable selection. *arXiv*, 1602.07640, 1–33.

Johnson, V.E. (2013). Uniformly most powerful Bayesian tests. *Annals of Statistics*, 41(4), 1716–1741.

Johnson, V.E. & Rossell, D. (2010) On the use of non-local prior densities for default Bayesian hypothesis tests. *Journal of the Royal Statistical Society B*, 72, 143–170.

Johnson, V.E. & Rossell, D. (2012) Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 24(498), 649–660.

Jordan, M.I., Ghahramani, Z., Jaakkola, T.S. & Saul, L.K. (1999) An introduction to variational methods for graphical models. *Machine Learning*, 37(2), 183–233.

Kan, R. (2008) From moments of sum to moments of product. *Journal of Multivariate Analalysis*, 99, 542–554.

Kass, R.E., Tierney, L. & Kadane, J.B. (1990) The validity of posterior expansions based on Laplace's method. *Bayesian and Likelihood Methods in Statistics and Econometrics*, 7, 473–488.

Minka, T.P. (2001) A family of algorithms for approximate Bayesian inference. PhD thesis, Massachusetts Institute of Technology, 4.

Muirhead, R.J. (2009) *Aspects of multivariate statistical theory*. Hoboken: John Wiley & Sons.

Narisetty, N.N. & He, X. (2014) Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2), 789–817.

Rahaman-Khan, M.H. & Shaw, J.E.H. (2019) Variable selection for accelerated lifetime models with synthesized estimation techniques. *Statistical Methods in Medical Research*, 28(3), 937–952.

Rossell, D. (2018) A framework for posterior consistency in model selection. *arXiv*, 1806.04071, 1–58.

Rossell, D. & Rubio, F.J. (2018) Tractable bayesian variable selection: Beyond normality. *Journal of the American Statistical Association*, 113(524), 1742–1758.

Rossell, D. & Rubio, F.J. (2019) Additive Bayesian variable selection under censoring and misspecification. *arXiv*, 1907.13563, 1–57.

Rossell, D. & Telesca, D. (2017) Non-local priors for high-dimensional estimation. *Journal of the American Statistical Association*, 112, 254–265.

Rue, H., Martino, S. & Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society B*, 71(2), 319–392.

Ruli, E., Sartori, N. & Ventura, L. (2016) Improved Laplace approximation for marginal likelihoods. *Electronic Journal of Statistics*, 10(2), 3986–4009.

Schäfer, C. & Chopin, N. (2013) Sequential Monte Carlo on large binary sampling spaces. *Statistics and Computing*, 23(2), 163–184.

Scheipl, F., Fahrmeir, L. & Kneib, T. (2012) Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, 107(500), 1518–1532.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.

Scott, J.G. & Berger, J.O. (2006) An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7), 2144–2162. doi: http://dx.doi.org/10.1016/j.jspi.2005.08.031.

Shin, M., Bhattacharya, A. & Johnson, V.E. (2018) Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statistica Sinica*, 28(2), 1053–1078.

Silvapulle, M.J. & Burridge, J. (1986) Existence of maximum likelihood estimates in regression models for grouped and ungrouped data. *Journal of the Royal Statistical Society B*, 48(1), 100–106.

Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. (2011) Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5), 1.

van der Vaart, A.W. (1998) *Asymptotic statistics*. New York: Cambridge University Press.

Yang, Y., Wainwright, M.J. & Jordan, M.I. (2016) On the computational complexity of highdimensional Bayesian variable selection. *The Annals of Statistics*, 44(6), 2497–2532.

Zanella, G. & Roberts, G. (2019) Scalable importance tempering and Bayesian variable selection. *Journal of the Royal Statistical Society B*, 81(3), 489–517.

Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2), 894–942.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the Supporting Information section.

How to cite this article: Rossell D, Abril O, Bhattacharya A. Approximate Laplace approximations for scalable model selection. *J R Stat Soc Series B*. 2021;00:1–27. https://doi.org/10.1111/rssb.12466