Latent space Projection Predictive Inference

Anonymous authors
Paper under double-blind review

Abstract

Given a reference model that includes all the available variables, projection predictive inference replaces its posterior with a constrained projection including only a subset of all variables. We extend projection predictive inference to enable computationally efficient variable and structure selection in models outside the exponential family. By adopting a latent space projection predictive perspective we are able to: 1) propose a unified and general framework to do variable selection in complex models while fully honouring the original model structure, 2) properly identify relevant structure and retain posterior uncertainties from the original model, and 3) provide an improved approach also for non-Gaussian models in the exponential family. We demonstrate the superior performance of our approach by thoroughly testing and comparing it against popular variable selection approaches in a wide range of settings, including realistic data sets. Our results show that our approach successfully recovers relevant terms and model structure in complex models, selecting less variables than competing approaches for realistic datasets.

1 Introduction

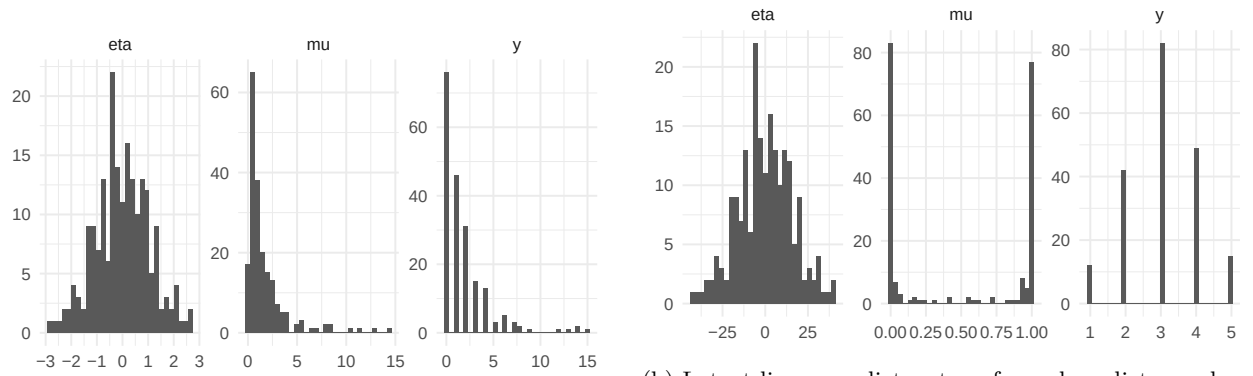
Variable and structure selection plays an important role in a robust Bayesian workflow (Gelman et al., 2020). While variable selection has been extensively studied and successfully applied for models in the exponential family (Koopman, 1936), it has not received the same attention regarding models outside the exponential family (e.g., advanced ordinal, count or time-to-event (also known as survival) data distributions), despite their important applications in different fields (Kelter, 2020; Nagler, 1994; Bürkner & Vuorre, 2019; Barron, 1992).

We propose an efficient, stable, and information theoretically justified method to make variable selection for non-normal observation models in or beyond the exponential family. The main benefits of the proposed *latent space projection predictive inference* are:

1. We enable the projection predictive variable and structure selection for models outside the exponential family while honouring the original model structure and its predictive uncertainty.
2. We obtain more stable projections for non-Gaussian exponential family models.
3. We demonstrate the superior performance of our method as compared to state-of-the-art competitors in both simulated and real-world scenarios.
4. We provide a ready-to-use open source implementation of the new methods.

2 Projection predictive inference

Given posterior draws $\{\lambda_*^{(s)}\}_{s=1}^S \sim p(\lambda \mid \mathcal{D})$ from a reference model with data $\mathcal{D} = \{X, y\}$, outcome observations y and predictor variables X, projection predictive inference (Piironen et al., 2020b; Catalina et al., 2022; McLatchie et al., 2023) learns a projection $q_\perp(\lambda)$ containing only a subset of variables that matches the reference predictive performance as close as possible. The solution is given by the minimiser of the Kullback-Leibler (KL) divergence from the reference model to the projection predictive distributions (Dupuis & Robert, 2003). Let $p(\tilde{y} \mid \mathcal{D})$ be the reference predictive distribution and $\{\lambda_\perp^{(s)}\}_{s=1}^S \sim q_\perp(\lambda)$ be draws from



(a) Latent linear predictor, transformed predictor and response space for a Poisson model with `log` link function.

(b) Latent linear predictor, transformed predictor and response space for an ordinal cumulative model with `probit` link function.

Figure 1: Latent (η), transformed (μ), and response (y) space representation for two models.

the projection, and $q_\perp(\tilde{y})$ be the projection predictive distribution, then:

$$\begin{aligned}&\mathrm{KL}\left(p\left(\tilde{y} \mid \mathcal{D}\right) \parallel q_\perp\left(\tilde{y}\right)\right)\\ &\quad= -\mathbb{E}_{\lambda_*}\left(\mathbb{E}_{\tilde{y}|\lambda_*}\left(\log \mathbb{E}_{\lambda_\perp}\left(p\left(\tilde{y} \mid \lambda_\perp\right)\right)\right)\right) + \mathrm{C}.\end{aligned} \tag{1}$$

As the integrals involved are in most cases computationally infeasible, Goutis & Robert (1998) suggested changing the order of integration and minimisation and solve the optimisation for each posterior draw separately

$$\arg\max_{\lambda} \mathbb{E}_{\tilde{y}|\lambda_*^{(s)}}\left(\log \mathbb{E}_{\lambda_\perp}\left(p\left(\tilde{y} \mid \lambda\right)\right)\right). \tag{2}$$

Piironen et al. (2020b) proposed a further speed-up by first clustering the posterior draws $\lambda_*^{(s)}$ and then solving the optimisation individually for each resulting cluster centre $\{\lambda_*^{(c)}\}_{c=1}^{C}$.

For models in the exponential family distribution, Eq. (2) coincides with computing maximum likelihood estimates under the projection model as

$$\lambda_\perp = \arg\max_{\lambda} \sum_{i=1}^{N} \mu_i^* \xi_i(\lambda) - B(\xi_i(\lambda)), \tag{3}$$

where $\mu^* = \mathbb{E}_{\tilde{y}|\lambda_*}(\tilde{y})$ are mean predictions of the reference model, ξ_i are the natural parameters for the ith observation and $B(\cdot)$ is a function of the natural parameters. These maximum likelihood estimates can be efficiently computed by solving penalised iteratively reweighted least squares (PIRLS; Marx, 1996). If non-constant, the projected scale parameter of the exponential family model is then obtained as

$$\phi_\perp = \arg\max_{\phi} \sum_{i=1}^{N} \left(\frac{r_i(\lambda_\perp)}{A(\phi)} + \mathrm{E}_{\tilde{y}_i|\lambda_*}(H(\tilde{y}_i, \phi))\right), \tag{4}$$

where A, H are family-specific functions and $r_i(\lambda_\perp) = \mu_i^* \xi_i(\lambda_\perp) - B(\xi_i(\lambda_\perp))$ does not depend on ϕ.

3 Latent space projective inference

The equivalence between maximum likelihood estimates and KL minimiser does not hold for models outside the exponential family. For data $\mathcal{D} = \{X, y\}$, parameters λ and inverse link function g, we assume the general model formulation

$$y \sim p\left(\mu, \phi\right), \quad \mu = g(\eta), \quad \eta \sim p\left(\eta \mid \lambda, X\right), \tag{5}$$

where η is commonly called the latent predictor (McCullagh & Nelder, 1989). In Fig. 1 we show the difference between the linear and the transformed predictor spaces for two common models, a Poisson model with a `log` link function and an ordinal cumulative model (Bürkner & Vuorre, 2019) with `logit` link function.

We propose to reformulate the projection problem by solving the KL minimisation in the latent predictive space $p(\tilde{\eta} \mid \lambda_*)$ as:

$$\begin{aligned} &\mathrm{KL}\left(p\left(\tilde{\eta} \mid \mathcal{D}\right) \parallel q_{\perp}\left(\tilde{\eta}\right)\right) = \\ &\quad - \mathbb{E}_{\lambda_*}\left(\mathbb{E}_{\tilde{\eta}|\lambda_*}\left(\log \mathbb{E}_{\lambda_\perp}\left(p\left(\tilde{\eta} \mid \lambda_\perp\right)\right)\right)\right) + \mathrm{C}. \end{aligned}$$

3.1 Non-exponential family case

The distribution $p(\eta \mid \lambda, X)$ is model-dependent and in non-exponential family case does not, in general, have a nice closed form. We propose to approximate $p(\eta \mid \lambda, X)$ with a Gaussian distribution, a choice motivated by several reasons: 1) as potential boundaries on μ are enforced only via the inverse link g, η itself is unbounded so that its support matches the support of a Gaussian, 2) the Gaussian distribution belongs to the exponential family and is thus computationally easy to handle, and 3) the model on η is typically additive, such that commonly chosen Gaussian priors on the latent parameters λ make the implied distribution of η closer to Gaussian as well. Now that the transformed projection is again in the exponential family, we can apply the approach outlined in Section 2 to solve the projection.

Although the dispersion of the latent space is typically not a model parameter, we still need to approximate it to compute its projection in Eq. (4). The approximate dispersion is model dependent, and may sometimes even be known analytically, as in the case of an ordinal cumulative family, where the known latent dispersion is exactly 1.

3.2 Exponential family case

The latent projection formulation also works for models in the exponential family. For Gaussian observation models, it coincides exactly with the original framework, and therefore no improvement is gained with the latent approach.

For non-Gaussian exponential family models, the original framework computes an approximate solution to Eq. (1) via PIRLS (as in Eq. (3)), which often results in unstable solutions for models with complex structure or link functions (Catalina et al., 2022). In contrast, the latent approach computes an approximation on the latent space, removing the complexity of the link function and the response model. Experimental results show that the latent approach results in significant improvements also for non-Gaussian exponential family models.

4 Related work

For models in the exponential family, variable selection has a strong presence in the literature. Some methods perform variable selection by optimising a penalised likelihood formulation (Tibshirani, 1996; Zou & Hastie, 2005; Friedman et al., 2010a; Candes & Tao, 2007; Breiman, 1995; Fan & Li, 2001), while at the same time trying to select a subset of relevant variables. These methods impose a penalisation on large coefficients, effectively driving some of them towards zero, depending on the choice of regularization. For more details, we refer the reader to Hastie (2015). These methods suffer from several drawbacks. First, by dealing with the estimation of the model and the selection of variables at the same time, they often result in suboptimal solutions (Piironen & Vehtari, 2017a). Second, they are derived for specific likelihoods, and generalising them to other models is difficult, if at all possible. Most notably, these methods cannot perform variable selection on group-specific parameters in hierarchical models (Catalina et al., 2022).

These approaches have been generalized to specific likelihoods outside the exponential family, such as ordinal or Cox models (Wurm et al., 2017; Archer & Williams, 2012; Archer et al., 2014; Fan et al., 2005), but no general framework exists. Penalised likelihood approaches can be applied to models outside the exponential family by approximating the likelihood with an exponential family distribution. This enables variable selection,

but the resulting model is likely to underperform in terms of predictive performance. A model in the original response space can be obtained by fitting a Bayesian model including only the selected variables. We compare our method against baselines that follow this approach in Section 5.2.

Variable selection is also addressed from the Bayesian perspective (O'Hara & Sillanpää, 2009). This is typically done by imposing so-called sparsity priors, such as the horseshoe (Carvalho et al., 2010; Piironen & Vehtari, 2017b) or the spike-and-slab (Ishwaran & Rao, 2005). As the posterior itself is not fully sparse, a sparse solution is then obtained by thresholding based on posterior expectations. Hahn & Carvalho (2015) discus related decopuling shrinkage and selection, that loosely follows the original decision theoretical idea by Lindley (1968). Piironen & Vehtari (2017a) and Piironen et al. (2020b) discuss the relationship of Hahn & Carvalho (2015) approach to projection predictive inference and discuss why the latter is more principled by following the original approach by Lindley (1968) more closely.

If a Markov chain Monte Carlo (MCMC) algorithm (Robert & Casella, 2013) is used for inference, no strong assumptions on the likelihood or overall structure are needed. This makes Bayesian inference applicable to models outside the exponential family with multilevel (or other complex) structure. This helps with the generality of the inference, but still falls short on the selection, since the user still needs to manually decide which variables should be selected.

Bayesian reference models have been used for variable and structure selection tasks in the context of exponential family models, including (additive) multilevel models and discrete response families with finite support (Piironen et al., 2020b; Catalina et al., 2022; Pavone et al., 2022; Piironen & Vehtari, 2016; Weber et al., 2023). However, its application to models outside the exponential family has remained unexplored.

Our approach solves variable selection on non-exponential family models by performing the selection on the latent space of the original model, therefore keeping the original structure. We approximate the unknown latent distribution with a Gaussian distribution. The problem of learning an implicit distribution has been approached in the Bayesian inference literature from different angles. Some authors tackle this problem with variational inference (Blei et al., 2017), by learning an implicit mapping between samples of the implicit distribution and a powerful and expressive learner, typically normalising flows (Rezende & Mohamed, 2016; Huszár, 2017; Pequignot et al., 2020; Titsias & Ruiz, 2019). Optimising these flexible models is typically expensive, requiring many iterations and diagnostics to have any guarantee of convergence (Dhaka et al., 2020). On top of that, the main bottleneck in projection predictive inference is not solving this projection *once*, but possibly many times, as complex models require many posterior draws to be projected for the projections to fully capture the posterior uncertainty in the reference model.

5 Experiments

We evaluate the performance of the proposed method with simulated and real-world data experiments. In particular, we must account for complex high dimensional posterior geometries, the role of correlated terms, and the size of the selected subset of variables in different models outside the exponential family.

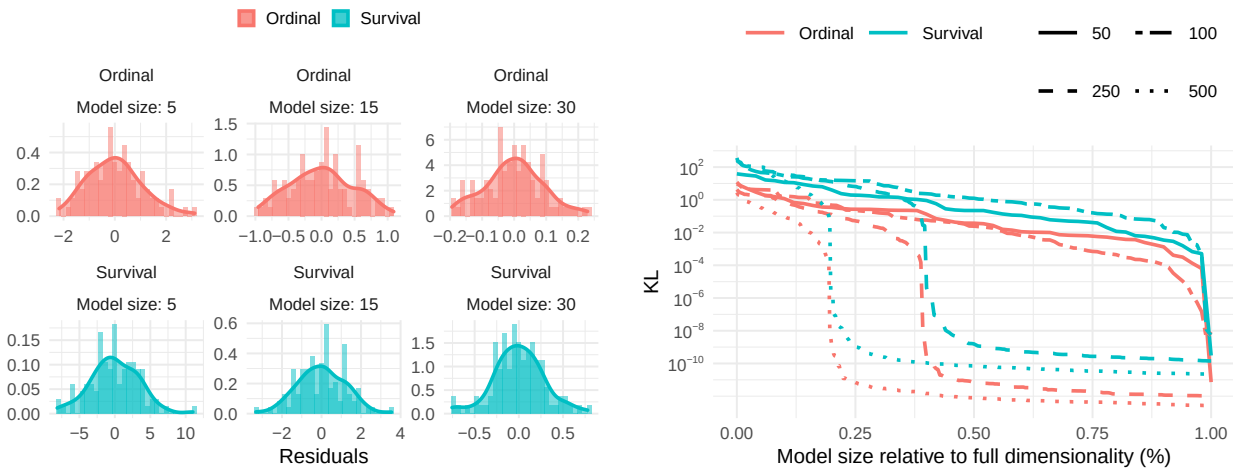
First, we demonstrate the use of different diagnostics for the latent approximation. Second, we demonstrate the variable selection itself in high dimensional non-exponential family examples. Then, we extend the experiments to real datasets. Finally, we show the benefits that our method brings to models in the exponential family, too.

For the experiments, we used modified `projpred` (Piironen et al., 2020a) available at `https://github.com/stan-dev/projpred/tree/e1913e4b59cff00f1e7a5386c068431aa3368dec`. The latent space projection is now also available in `projpred` package since release 2.4.0.

5.1 Diagnosing the quality of the latent approximation

We use two diagnostics to assess the quality of the approximate projections:

- We perform projection predictive checks on the residuals $\tilde{\eta}_* - \tilde{\eta}_\perp$, where $\tilde{\eta}_*$ corresponds to the latent predictions of the reference model and $\tilde{\eta}_\perp$ to the predictions of the projection.



(a) Residual histogram plots for three selected projections for a reference model of size $D = 50$. In all cases, the residual distribution is close to normal with a decreasing standard deviation as more terms are introduced in the projection. Notice the shrinking x-axis as the projections include more terms.

(b) KL divergence between reference and projection predictive distributions. Different line types indicate size of the reference model. As projections include a higher proportion of the reference model, the KL-divergence approaches 0 faster.

Figure 2: Residual histograms and KL-divergence plots for projections on a simulated time-to-event survival analysis model and an ordinal cumulative model, both with $N = 100$ observations, and correlation factor $\rho = 0.3$.

- We check the Kullback-Leibler divergence between the reference's and projection's predictions after convergence, which should approach 0 as more terms are included in the projection.

We use simulated data from 1) an ordinal cumulative model with `probit` link function and 2) a time-to-event survival model with a `log` link function. The outcome in both cases is generated as a function of $D \in [50, 100, 250, 500]$ sampled predictors with a uniform correlation factor of 0.3, where only 60% of the predictors have a non-zero effect on the response. The full generation process for these data is detailed and further analysed in Section 5.2.

Projection predictive checks assess normality assumptions. Histograms of projection residuals (Fig. 2a) for various model sizes are a practical diagnostic for the normality assumption. As more terms enter the projection, the residuals get smaller and more concentrated, as indicated by the shrinking x-axis in the figure.

Kullback-Leibler divergence shows that the latent projections eventually match the reference model predictions. Even for the most challenging scenarios, the KL-divergence of the latent predictions shows that the projection predictive distribution gets closer (the KL-divergence approaches 0) to the reference predictive distribution as more terms enter the projection (Fig. 2b).

5.2 Non-exponential family models with simulated data

We compare the predictive performance of the optimal submodels in terms of held-out expected log predictive density (ELPD) (Vehtari & Ojanen, 2012) for two types of models: an ordinal cumulative model and a time-to-event survival analysis model with a Weibull hazard process. Additionally, we examine the performance regarding the selection of truly relevant variables.

For the simulated high-dimensional data, we compare the performance of our approach to other popular sparsity promoting solutions in the literature:

- Elastic net regularization as implemented by `glmnet` (Friedman et al., 2010b), abbreviated as `glmnet`.
- Spike-and-slab priors as implemented by `spikeSlabGAM` (Scheipl, 2011), abbreviated as `ss`.
- Spike-and-slab LASSO priors as implemented by `SSLASSO` (Rockova & George, 2018), abbreviated as `sslasso`.
- Projection predictive inference on the approximate response space as implemented by `projpred` (Piironen et al., 2020a), abbreviated as `projpred`.

Since these approaches do not exist for the specific likelihoods we use, or for models outside the exponential family in general, we run variable selection on an approximate model, where we assume a normal likelihood of the response rather than the appropriate likelihood (cumulative or Weibull in our examples). Note that this normal approximation is different from our latent approach, since the latter approximates the latent predictor, not the response.

We first fit ordinal and time-to-event survival regression models on extensive simulation conditions. The generative process is

$$
\begin{aligned}
&x_n \sim \text{Normal}(0, \Sigma_\rho), \quad z_d \sim \text{Bernoulli}(0.6), \\
&\beta_d \sim z_d \cdot \text{Normal}(0, 1.5), \\
&\eta_n = x_n^T \beta, \quad y_n \sim p(g(\eta_n), \phi),
\end{aligned}
$$

where $y \in \mathbb{R}$, $x_n \in \mathrm{R}^D$ are the outcome and predictors respectively, β denotes the unknown coefficients, number of variables D is varied in $[50, 100, 250, 500]$ and the number of observations N is also varied in $[100, 200, 300]$. ρ indicates the uniform correlation between predictors, g is the model-specific inverse link function and p is either an ordinal cumulative likelihood (with 5 ordered categories) or a time-to-event likelihood with Weibull base hazard process (maximum time-to-event is capped to 5), and ϕ are model specific parameters. Here, we show results for $N = 100$ observations and $\rho = 0$.

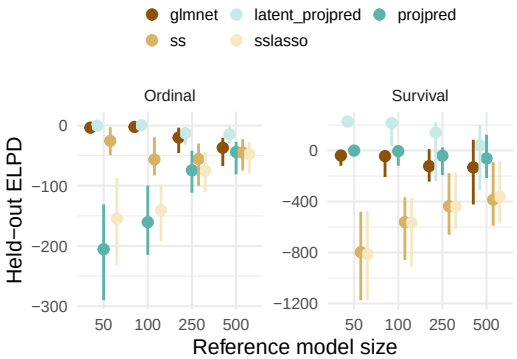
We fit the reference model making use of `brms` (Bürkner, 2018) with a regularized horseshoe prior (Piironen & Vehtari, 2017b) accounting for the sparsity in the predictors. The regularized horseshoe prior helps in avoiding overfitting, particularly for the models with a larger number of covariates.

Both the KL and residuals diagnostics indicate a good approximation for these models as shown in Section 5.1. Here we analyse the predictive and selection performance.

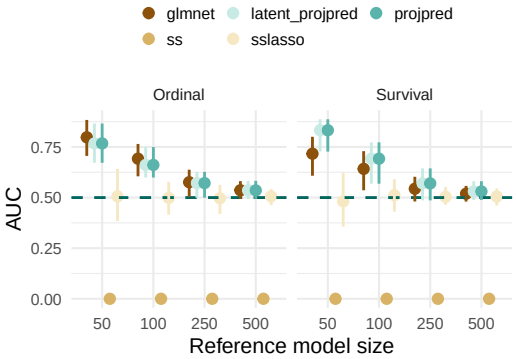
Latent projections achieve the best held-out ELPD performance. Fig. 3a shows that the latent approach always achieves the best performance, even surpassing the full reference model with models that include substantially fewer terms. This can be explained by a slight overfitting in the reference model. In the most complex scenarios, we see that the performance of all methods suffers from slightly larger variance, including the full reference model.

Approximate likelihood methods result in underperforming selections. While the performance of `glmnet` is the closest to our proposed method on average, its performance distribution is wider and less reliable, while other methods suffer greatly from the approximate likelihood representation (Fig. 3a). Spike-and-slab priors (`ss` in the figures) impose a very strong penalisation, which, together with the arbitrary choice of threshold (0.5 in our experiments, as is common, e.g. in Scheipl (2011)), can result in suboptimal selections. The ill-informed approximate likelihood results in very few terms crossing the selection threshold, particularly in `ss` and naive `projpred`. Even though `sslasso` does not need an arbitrary threshold for the selection, it suffers from similar problems.

Other approaches require refitting a Bayesian model. Other approaches operate on an approximate likelihood model that does not result in a sensible predictive model. To have a functional reduced model in the same domain as the reference model, all approaches except ours require refitting a full Bayesian model with the selected predictors and correct (non-normal) likelihood, increasing their cost. Our latent approach fully honours the reference model structure, and predictions in the original model space only require evaluating the inverse link function for the latent predictions.

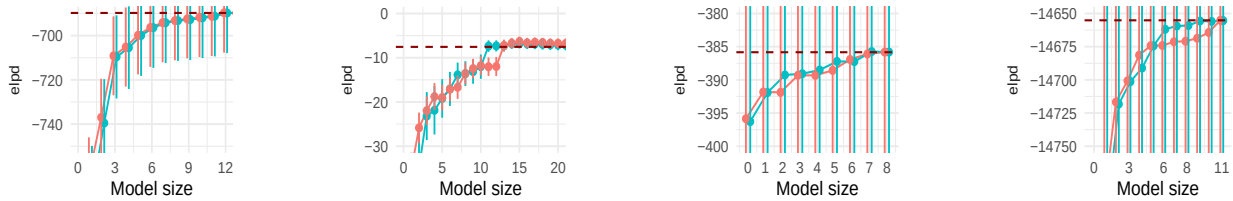


(a) Median and 90% confidence interval ELPD difference to the reference model for the selected subset of predictors for an ordinal cumulative model (left) and a time-to-event survival model (right) with $N = 100, \rho = 0$. Higher values are better. Competing methods are described in Section 5.2.



(b) Median and 90% confidence interval of the area under the curve (AUC) for the selection of truly relevant terms in an ordinal cumulative model (left) and a time-to-event survival model (right), with $N = 100$, and $\rho = 0$ for all competing methods. Higher values are better. Performance of pure random selection is shown as dashed dark green line. Competing methods are described in Section 5.2.

Figure 3: Mean and median values for ELPD difference and AUC for the selected submodels.



(a) Eyedisease (b) HCCFrame (c) Cancer (d) Rotterdam

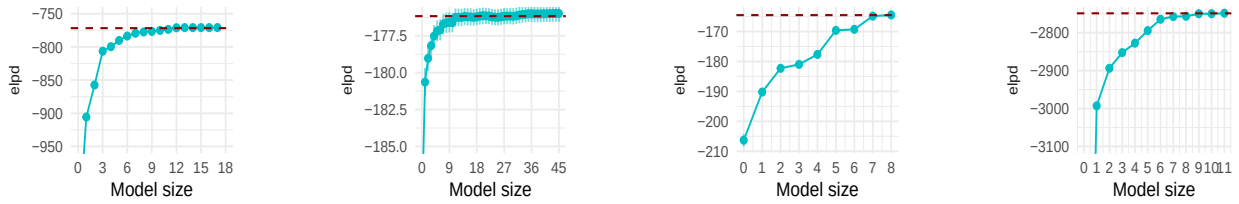
Figure 4: Full data zoomed-in ELPD performance comparison in the response model space. For `latent_projpred`, the projections are computed in the latent space, and the predictions are then transformed. Red shows `glmnet`, and cyan shows `latent_projpred`. Error bars indicate ELPD 95% quantiles. Larger error bars correspond to datasets with fewer observations.

Latent approach retains reference model uncertainty. The uncertainty present in the reference model is projected as well, contributing to better informed solutions than the other approaches, which cannot represent the original model likelihood. Fig. 3b shows that both `projpred` approaches achieve the highest AUC, even though naive `projpred` fails to assess the predictive performance of the projections. For the simpler `ordinal` experiments, `glmnet` achieves similar AUC. Interestingly, `ss` has trouble identifying any relevant features, and therefore its selection is not accurate, whereas `sslasso` manages to identify at least some.

5.3 Non-exponential family models with real data

In this section, we assess the selection and predictive capabilities of our latent approach on real datasets and compare it against `glmnet` as the best competitor in our simulated examples. We focus on the quality of the selection path and the properties of the latent projections.

Latent projections identify superior solution path. Fig. 4 shows the comparison of the latent approach against `glmnet`. From a predictive performance point of view, the latent approach identifies a better solution path for all datasets we tested, except for `eyedisease`, where both approaches are equal. Note that, along



(a) Eyedisease (b) HCCFrame (c) Cancer (d) Rotterdam

Figure 5: Full data ELPD performance in the latent predictor space for `latent_projpred`. The reference model performance is shown with dark red dashed horizontal line.

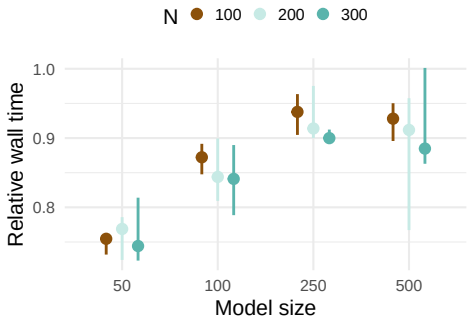


Figure 6: Median and 90% interval of relative wall time of `latent_projpred` over `projpred` for 50 Bernoulli data realizations and varying number of observations (different colors). Lower than 1 means `latent_projpred` is faster.

the path, there are multiple model sizes for which the latent approach's projections' performance is superior to their `glmnet` counterpart. Datasets with few observations (`cancer` and `rotterdam` particularly) suffer from higher variance in performance.

The response space is noisier. Our latent approach allows us to perform the variable selection on either the response scale or the latent predictor scale. When compared to the selection in response scale (Fig. 4), the latent predictor space is significantly less noisy (Fig. 5), mostly due to the complex observation model and sometimes relatively small datasets. The reference model filters noise from the data and thus the latent variables have less noise, resulting in reduced variance in the selection criterion (Fig. 5), even in data with few observations (`Cancer` and `Rotterdam` in particular).

5.4 Non-Gaussian exponential family models with simulated data

We simulate 50 realizations of a Bernoulli data with $N = 300$ observations and $D = 50, 100, 250, 500$ uncorrelated variables.

The latent approach is faster to compute. The latent normal approximation is solved efficiently in closed form, which is faster to compute than the iterative solution of the original projection approach, resulting in significantly faster solutions (Fig. 6).

Latent approach has less variability and selects fewer variables. We simulate 50 bootstrap samples from the same Bernoulli data generation process with $N = 100$ observations and $D = 50$ variables and compare the selection of the original framework against the latent approach. Fig. 7 shows that the latent approach results in smaller subsets of relevant variables with a smaller variability across bootstrap samples, while the original framework fails at identifying a suitable subset and overselects variables in every case. The difference in the behavior is mostly due to the model size selection rule. As seen in Figures 4 and 5,

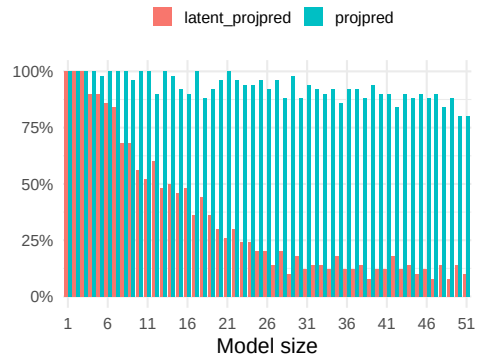


Figure 7: Bootstrap inclusion frequencies, for a simulated Bernoulli dataset with $N = 100$ observations and $D = 50$ variables, computed from 50 bootstrap samples.

latent projection has smaller uncertainty on the submodel performance and thus there is less variation in the selected model sizes.

5.5 Non-Gaussian exponential family models with real data

For complex non-Gaussian models, the original framework provides a solution that might suffer from unstable projections and often fail to converge to sensible results. We show that the latent approach obtains a superior solution for a real model in the context of endangered species conservation taken from an ongoing collaboration with domain experts (Digby et al., 2023).

These data consist of a set of 211 kākāpō individuals (it is not possible to obtain more observations due to the very nature of the data set) and the aim of the study is to find a model for the fertility of upcoming eggs. We constructed a hierarchical Bernoulli model (the response is 0 for fertile and 1 for infertile eggs) with a varying intercept per individual. In short, the model can be summarized as

$$
\begin{aligned}
y_i &\sim \text{Bernoulli}(\mu_i) \\
\mu_i &= \text{inv_logit}(\eta_i) \\
\eta_i &= \alpha + X\beta + z_i
\end{aligned}
$$

where X is the design matrix containing all covariates, β are the population parameters, α denotes the population intercept and z_i stands for the varying intercept for the ith individual. The group effect parameters can dominate the outcome variance in the model, which paired with the nonlinear link function, can often result in PIRLS solver convergence issues.

Latent approach provides superior performing projections. Fig. 8 shows that the latent approach results in more stable and efficient projections for kākāpō fertility model. The latent approach accurately captures the reference model predictive performance, and also a better ordering in the solution path.

A note on the convergence issues for original `projpred`. As reported in lme4 issue tracker [1,2], it seems that the underlying solver for the projections in `projpred` may have issues in the case of non-integer responses for Binomial or Poisson likelihoods. Our investigation and diagnosing of the issue seems to indicate that the underlying cause would be in the solving of the integral for estimating group-effect parameters (Catalina et al., 2022; Bates et al., 2015). Models suffering from this issue show the estimated effects shrinking towards

[1] `https://github.com/lme4/lme4/issues/682`
[2] `https://github.com/lme4/lme4/issues/180`

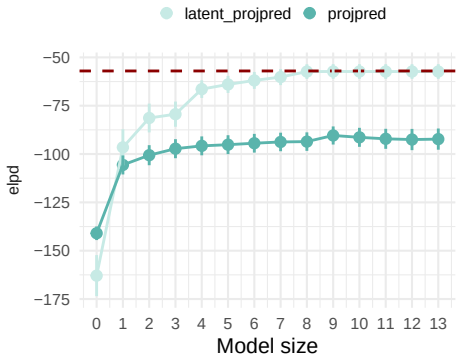


Figure 8: Comparison of full data expected log predictive density (ELPD) performance for a hierarchical Bernoulli model of fertility among kākāpō population. Reference performance in dark red dashed horizontal line. The original `projpred` solution does not reach the reference model predictive performance due to convergence issues in the underlying projections when including the varying intercept per individual.

zero. This is a natural effect in binomial models, since responses closer to 0.5 indicate inability of discerning both groups. However, the issue present here is that the group-effects estimates are also shrunk towards zero already with response values are as high as 0.9 (quite close to 1), which in the end are translated into worse projections. These worse projections can partially explain some of the differences between `latent_projpred` and original `projpred`, and we must be cautious when drawing conclusions from these results only. It is also noteworthy that the latent approach did not show such issues.

6 Discussion

We proposed a novel latent space projection predictive approach to perform variable and structure selection in models with non-exponential family observation model. As shown in the experiments, the approach is also beneficial for non-Gaussian exponential family models.

We have shown that our approach offers superior performance in extensive experiments on high dimensional generalized linear models and several real datasets. The proposed method is not limited to latent linear models, as it can readily be used for variable and structure selection in more complicated models, including linear and non-linear hierarchical models (as demonstrated in Fig. 8).

Our approach not only improves variable and structure selection in a whole new set of models while respecting their original structure, but also offers superior performance in currently supported models in `projpred`.

The method may have worse performance for models whose latent space is far from normal. We provide diagnostics that alert the user in these cases, but further research is needed to improve the performance in such cases.

References

Kellie J. Archer and Andre A.A. Williams. L1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Statistics in Medicine*, 31:1464–1474, 2012. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3718008/`.

Kellie J Archer, Jiayi Hou, Qing Zhou, Kyle Ferber, John G Layne, and Amanda E Gentry. ordinalgmifs: An R package for ordinal regression in high-dimensional data settings. *Cancer informatics*, 13:187—195, 2014. ISSN 1176-9351. doi: 10.4137/cin.s20806. URL `https://europepmc.org/articles/PMC4266195`.

David N Barron. The analysis of count data: Overdispersion and autocorrelation. *Sociological methodology*, pp. 179–220, 1992.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, Articles*, 67(1):1–48, 2015. ISSN 1548-7660. doi: 10.18637/jss.v067.i01. URL https://www.jstatsoft.org/v067/i01.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Leo Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995. ISSN 00401706. URL http://www.jstor.org/stable/1269730.

Paul-Christian Bürkner. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1):395, 2018. ISSN 2073-4859. doi: 10.32614/rj-2018-017. URL http://dx.doi.org/10.32614/rj-2018-017.

Paul-Christian Bürkner and Matti Vuorre. Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2(1):77–101, 2019.

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, Dec 2007. ISSN 0090-5364. doi: 10.1214/009053606000001523. URL http://dx.doi.org/10.1214/009053606000001523.

Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, Apr 2010. ISSN 1464-3510. doi: 10.1093/biomet/asq017. URL http://dx.doi.org/10.1093/biomet/asq017.

Alejandro Catalina, Paul-Christian Bürkner, and Aki Vehtari. Projection predictive inference for generalized linear and additive multilevel models. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151, pp. 4446–4461. PMLR, 2022.

Akash Kumar Dhaka, Alejandro Catalina, Michael Riis Andersen, Måns Magnusson, Jonathan H. Huggins, and Aki Vehtari. Robust, accurate stochastic optimization for variational inference. In *NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/7cac11e2f46ed46c339ec3d569853759-Abstract.html.

Andrew Digby, Daryl Eason, Alejandro Catalina, Michael Lierz, Stephanie Galla, Lara Urban, Marissa F. Le Lec, Joseph Guhlin, Tammy E. Steeves, Peter K. Dearden, Tineke Joustra, Caroline Lees, Tane Davis, Deidre Vercoe, and Kākāpō Recovery Team. Hidden impacts of conservation management on fertility of the critically endangered kākāpō. *PeerJ*, 11:e14675, February 2023. ISSN 2167-8359. doi: 10.7717/peerj.14675. URL https://doi.org/10.7717/peerj.14675.

Jerome A. Dupuis and Christian P. Robert. Variable selection in qualitative models via an entropic explanatory power. *Journal of Statistical Planning and Inference*, 111, 2003. ISSN 03783758. doi: 10.1016/S0378-3758(02)00286-0.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. doi: 10.1198/016214501753382273. URL https://doi.org/10.1198/016214501753382273.

Jianqing Fan, Gang Li, and Runze Li. An overview on variable selection for survival analysis. In *Contemporary Multivariate Analysis And Design Of Experiments: In Celebration of Professor Kai-Tai Fang's 65th Birthday*, pp. 315–336. World Scientific, 2005.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 2010a. ISSN 1548-7660. doi: 10.18637/jss.v033.i01. URL http://dx.doi.org/10.18637/jss.v033.i01.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010b. URL http://www.jstatsoft.org/v33/i01/.

Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020.

Constantinos Goutis and Christian P. Robert. Model choice in generalised linear models: A Bayesian approach via Kullback-Leibler projections. *Biometrika*, 85(1):29–37, Mar 1998. ISSN 1464-3510. doi: 10.1093/biomet/85.1.29. URL http://dx.doi.org/10.1093/biomet/85.1.29.

P Richard Hahn and Carlos M Carvalho. Decoupling shrinkage and selection in bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448, 2015.

Trevor Hastie. *Statistical Learning with Sparsity*. Chapman and Hall/CRC, May 2015. ISBN 9781498712170. doi: 10.1201/b18401. URL http://dx.doi.org/10.1201/b18401.

Ferenc Huszár. Variational inference using implicit distributions. *CoRR*, abs/1702.08235, 2017. URL http://arxiv.org/abs/1702.08235.

Hemant Ishwaran and J. Sunil Rao. Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773, Apr 2005. ISSN 0090-5364. doi: 10.1214/009053604000001147. URL http://dx.doi.org/10.1214/009053604000001147.

Riko Kelter. Bayesian survival analysis in stan for improved measuring of uncertainty in parameter estimates. *Measurement: Interdisciplinary Research and Perspectives*, 18(2):101–109, 2020. doi: 10.1080/15366367.2019.1689761. URL https://doi.org/10.1080/15366367.2019.1689761.

B. O. Koopman. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, 39(3):399–399, Mar 1936. ISSN 0002-9947. doi: 10.1090/s0002-9947-1936-1501854-3. URL http://dx.doi.org/10.1090/s0002-9947-1936-1501854-3.

D. V. Lindley. The choice of variables in multiple regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(1):31–66, 1968.

Brian D Marx. Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics*, 38(4):374–381, 1996.

Peter McCullagh and John A. Nelder. *Generalized Linear Models*. Springer US, 1989. ISBN 9781489932426. doi: 10.1007/978-1-4899-3242-6. URL http://dx.doi.org/10.1007/978-1-4899-3242-6.

Yann McLatchie, Sölvi Rögnvaldsson, Frank Weber, and Aki Vehtari. Robust and efficient projection predictive inference. *arXiv preprint arXiv:2306.15581*, 2023.

Jonathan Nagler. Scobit: An alternative estimator to logit and probit. *American Journal of Political Science*, 38(1):230–255, 1994. ISSN 00925853, 15405907. URL http://www.jstor.org/stable/2111343.

R. B. O'Hara and M. J. Sillanpää. A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, 4(1):85 – 117, 2009. doi: 10.1214/09-BA403. URL https://doi.org/10.1214/09-BA403.

Federico Pavone, Juho Piironen, Paul Christian Bürkner, and Aki Vehtari. Using reference models in variable selection. *Computational Statistics*, May 2022. ISSN 0943-4062. doi: 10.1007/s00180-022-01231-6.

Yann Pequignot, Mathieu Alain, Patrick Dallaire, Alireza Yeganehparast, Pascal Germain, Josée Desharnais, and François Laviolette. Implicit variational inference: the parameter and the predictor space. 2020. URL http://arxiv.org/abs/2010.12995.

Juho Piironen and Aki Vehtari. Projection predictive model selection for Gaussian processes. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2016. doi: 10.1109/MLSP.2016.7738829.

Juho Piironen and Aki Vehtari. Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735, 2017a. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-016-9649-y. URL `http://link.springer.com/10.1007/s11222-016-9649-y`.

Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051, 2017b. ISSN 1935-7524. doi: 10.1214/17-EJS1337SI. URL `https://projecteuclid.org/euclid.ejs/1513306866`.

Juho Piironen, Markus Paasiniemi, Alejandro Catalina, Frank Weber, and Aki Vehtari. *projpred: Projection Predictive Feature Selection*, 2020a. https://mc-stan.org/projpred, https://discourse.mc-stan.org/.

Juho Piironen, Markus Paasiniemi, and Aki Vehtari. Projective inference in high-dimensional problems: Prediction and feature selection. *Electronic Journal of Statistics*, 14(1):2155–2197, 2020b. doi: 10.1214/20-EJS1711. URL `https://doi.org/10.1214/20-EJS1711`.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. 2016. URL `http.://arxiv.org/abs/1505.05770`.

Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

Veronica Rockova and Edward George. The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444, 2018.

Fabian Scheipl. spikeSlabGAM: Bayesian variable selection, model choice and regularization for generalized additive mixed models in R. *Journal of Statistical Software*, 43(14):1–24, 2011. URL `http://www.jstatsoft.org/v43/i14/`.

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL `http://www.jstor.org/stable/2346178`.

Michalis K Titsias and Francisco Ruiz. Unbiased implicit variational inference. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 167–176. PMLR, 2019.

Aki Vehtari and Janne Ojanen. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6(none):142 – 228, 2012. doi: 10.1214/12-SS102. URL `https://doi.org/10.1214/12-SS102`.

Frank Weber, Änne Glass, and Aki Vehtari. Projection predictive variable selection for discrete response families with finite support. *arXiv preprint arXiv:2301.01660*, 2023.

Michael J. Wurm, Paul J. Rathouz, and Bret M. Hanlon. Regularized Ordinal Regression and the ordinalNet R Package, 2017.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, Apr 2005. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2005.00503.x. URL `http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x`.