
Projection Predictive Inference for Generalized Linear and Additive Multilevel Models

Alejandro Catalina

Paul Bürkner

Aki Vehtari

Helsinki Institute for Information Technology, HIIT
Aalto University, Department of Computer Science

Abstract

Projection predictive inference is a decision theoretic Bayesian approach that decouples model estimation from decision making. Given a reference model previously built including all variables present in the data, projection predictive inference projects its posterior onto a constrained space of a subset of variables. Variable selection is then performed by sequentially adding relevant variables until predictive performance is satisfactory. Previously, projection predictive inference has been demonstrated only for generalized linear models (GLMs) and Gaussian processes (GPs) where it showed superior performance to competing variable selection procedures. In this work, we extend projection predictive inference to support variable and structure selection for generalized linear multilevel models (GLMMs) and generalized additive multilevel models (GAMMs). Our simulative and real-world experiments demonstrate that our method can drastically reduce the model complexity required to reach reference predictive performance and achieve good frequency properties.

1 INTRODUCTION

Variable selection is an important aspect of statistical and predictive modelling workflows, for example, when understanding a model’s predictions is important, or where there is a cost associated to collecting new data. From the perspective of predictive performance, one goal of variable selection is to find the smallest subset

of variables in a dataset yielding comparable predictive performance to the full model containing all the available variables. In this context, we assume that there might be variables with true non-zero coefficients that we cannot properly detect due to scarce data or the presence of a highly complex correlation structure.

In this paper, we substantially generalize projection predictive inference to perform variable selection and model structure selection in generalized linear multilevel models (GLMMs) (McCulloch, 2003, Gelman et al., 2013) and generalized additive multilevel models (GAMMs) (Hastie and Tibshirani, 1986, Verbyla et al., 1999). Both types of models are widely used across the quantitative sciences, for instance, in the social and political sciences (e.g. poll or elections data whose measurements are organized in regions or districts with multiple levels), or in the physical sciences (e.g., meteorological or medical data).

Projection predictive inference (Piironen et al., 2020b) is a general Bayesian decision theoretic framework that separates model estimation from decision. Given a reference model on the basis of all variables, it aims at replacing its posterior $p(\lambda_* | \mathcal{D})$ with a constrained projection $q_{\perp}(\lambda)$. This projection is solved so that its predictions are as close as possible to the reference model’s predictions. The uncertainties in the reference model related to the excluded model parts are also projected and thus partially retained in the projection.

In the context of variable selection, one typically constrains the projection to a smaller subset of variables, where the excluded variables have their coefficients fixed at zero. Then, the projection procedure sequentially projects the posterior onto an incremental subspace, until all the variables have entered the projection. At each step, the method selects the variable that most decreases the Kullback-Leibler (KL) divergence from the reference model’s predictive distribution to that of the projection model, a procedure known as forward search. This forms a *solution path* for the variables into the projection. This approach has been shown to provide better performance than state-of-the-art com-

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

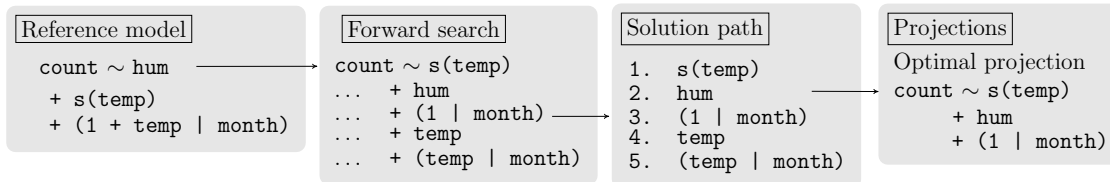


Figure 1: Illustration of projection predictive variable and structure selection workflow with BikeSharing data.

petitors (Piironen and Vehtari, 2017a, Piironen et al., 2020b, Pavone et al., 2020, Piironen and Vehtari, 2016). Piironen and Vehtari (2017a) demonstrate that, when using the projection approach, overfitting in the model space search is very small compared to other stepwise procedures and that, even in huge model spaces, the selected submodel has similar predictive performance to the reference model.

Previously, projection predictive inference has been used to perform variable selection only in generalized linear models (GLMs) (Piironen et al., 2020b) and Gaussian processes (GPs) (Piironen and Vehtari, 2016). However, the existing projection solutions do not translate to GLMMs or GAMMs because the projection is not identifiable for these models (Bickel and Doksum, 1977). That is, there is not a unique solution to the projection’s parameters. Nonetheless, it is possible to address these challenges and extend projection predictive inference to GLMMs and GAMMs as we show in this paper. Our contributions include:

- Discussing the identifiability issue for projecting to GLMMs and GAMMs.
- Extending projection predictive inference to support GLMMs and GAMMs. We address the challenging projection problem by learning a Laplace approximation to the marginal likelihood of the projection.
- Performing extensive simulations and real data experiments that validate the working of our method.
- Implementing our proposal in the open source `projpred` R package for projection predictive inference (Piironen et al., 2020a).

To give an illustration of our developed method’s application, consider the `BikeSharing` data we study in Section 5.2. These data include measurements for categorical variables for `month`, `season` or `weather`, and continuous variables for `temperature`, `humidity` and `windspeed`. A priori we assume all continuous variables are relevant for predictions and they may interact with all categorical variables. Our approach finds a projection model with substantially simplified model structure that still provides very good predictive performance. It only contains `month` group parameters

for `humidity` and `temperature` (i.e., parameters of humidity and temperature varying by month) while discarding all `weather` group parameters. The procedure is illustrated in Figure 1. Note we are using R formula syntax for conciseness, we point the reader to Bürkner (2018) for more details on the syntax.

2 RELATED METHODS

For GLMs, variable selection has been approached from different perspectives. Some methods (Breiman, 1995, Tibshirani, 1996, Fan and Li, 2001, Zou and Hastie, 2005, Candès and Tao, 2007) propose to deal with it by solving a penalized maximum likelihood formulation that enforces sparse solutions, while at the same time trying to select a subset of relevant variables (e.g. LASSO). These approaches suffer from confounding the estimation and selection of variables, often ending up selecting fewer variables than truly relevant in the data, as in the case of correlated covariates. For further information, see the comprehensive review by Hastie (2015). Similarly, Marra and Wood (2011) propose to add an additional penalty term to perform variable selection in GAMs, with similar shrinkage capacity as ridge regression. Another set of methods (George and McCulloch, 1993, Raftery et al., 1997, Ishwaran and Rao, 2005, Johnson and Rossell, 2012, Carvalho et al., 2010) suggests imposing a sparsifying prior on the coefficients that favours sparse solutions. Nonetheless, these priors do not actually produce sparse posteriors, because every variable has a non-zero probability of being relevant. One can obtain a truly sparse solution by selecting only those variables whose probability of being relevant is above a certain threshold (Barbieri and Berger, 2004, Ishwaran and Rao, 2005, Narisetty and He, 2014), but this approach ignores the uncertainty in the variables below the threshold.

Reference models have been used before for tasks other than variable selection, as in Afrabandpey et al. (2020), where the authors constrain the projection of a complex neural network to be interpretable (e.g., projecting onto decision trees). Piironen and Vehtari (2016) use projection predictive inference and impose further constraints on the projection of a GP reference model to perform

variable selection, given the identifiability issue of the direct projection.

While some alternative methods for variable selection in GLMMs and GAMMs exist, they either only allow variable selection for population parameters but not group parameters (Groll and Tutz, 2012, Tutz and Groll, 2012, Hubin and Storvik, 2018) or, when trying to use Bayes factors (Kass and Raftery, 1995), are computationally infeasible due to combinatorial explosion as soon as there are more than just a few variables. To the best of our knowledge, there are no practically applicable competing methods available in the literature, and so we only focus on the absolute performance of our method.

3 PROJECTION PREDICTIVE INFERENCE

3.1 Formulation of the KL Projection

Because the latent (parameter) domains of the reference and projected models may be different, formulating the problem in terms of minimizing a discrepancy measure between $p(\lambda_* | \mathcal{D})$ and $q_{\perp}(\lambda)$ does not make sense. Instead, we minimize the KL divergence from the reference model’s predictive distribution to that of the constrained projection:

$$\begin{aligned} \text{KL}(p(\tilde{y} | \mathcal{D}) \parallel q(\tilde{y})) &= \mathbb{E}_{\tilde{y}}(\log p(\tilde{y} | \mathcal{D}) - \log q(\tilde{y})) \\ &= -\mathbb{E}_{\tilde{y}}(\log q(\tilde{y})) + C \\ &= -\mathbb{E}_{\tilde{y}}(\log \mathbb{E}_{\lambda}(p(\tilde{y} | \lambda))) + C \\ &= -\mathbb{E}_{\lambda_*}(\mathbb{E}_{\tilde{y}|\lambda_*}(\log \mathbb{E}_{\lambda}(p(\tilde{y} | \lambda)))) + C, \end{aligned} \quad (1)$$

where we have collapsed terms that don’t depend on λ into C . Here, the expectations over λ_* , $\tilde{y} | \lambda_*$, λ are over the posterior $p(\lambda_* | \mathcal{D})$, the posterior predictive distribution $p(\tilde{y} | \lambda_*)$, and the constrained projection $q_{\perp}(\lambda)$, respectively.

In practice, we approximate the KL minimization by changing the order of the integration and optimization in $\mathbb{E}_{\tilde{y}|\lambda_*}(\log \mathbb{E}_{\lambda}(p(\tilde{y} | \lambda)))$. To make this feasible, Goutis (1998) propose to do the projection draw-by-draw, where we find a direct mapping from a posterior draw of the reference model to the projection’s constrained space. Piironen et al. (2020b) propose a further speedup by demonstrating that it is possible to solve the projection employing only a small subset of posterior draws or even representative points that can be found, for instance, by clustering.

3.2 Variable and Structure Selection

A high level overview of the variable selection procedure of projective predictive inference includes the following steps (Piironen et al., 2020b):

1. Cluster the draws of the reference model’s posterior.
2. Perform forward search to determine the ordering of the terms for the projection. At each step, include the term that most decreases the KL divergence between the reference model’s predictions and the projection’s.
3. Then, we sequentially recompute the projections, starting from the null model, adding one term at a time. To ensure more accurate predictive performance, this time we project more draws from the reference model.

For a more robust variable selection, we perform a leave-one-out (LOO) cross validation procedure through the model space. In this approach, we repeat the full forward search N times by performing the selection with $N - 1$ data points and leaving the remaining point as test point each time, resulting in N different solution paths. Instead of running the procedure for every observation, Vehtari et al. (2019) show that we can achieve a similarly robust selection by running the procedure only on a carefully selected subset of points based on their estimated Pareto \hat{k} diagnostic.

In GLMMs and GAMMs, we no longer perform variable but model structure selection, that is, select additive model components to which we refer to as *terms*. In this context, a *term* may refer to a single variable with a single coefficient, a group level term, corresponding to all the coefficients of the group’s levels, or a smooth term, corresponding to all the coefficients associated with the smooth basis functions. The structure selection involves the same steps as the variable selection, only that variables are replaced by terms.

3.3 Solving the Projection for Exponential Family Models

For GLMs with observation models in the exponential family (McCullagh and Nelder, 1989), projecting a draw λ_* from the reference model’s posterior to the projection space λ_{\perp} in Equation (1) coincides exactly with maximizing its likelihood under the projection model. Given a new observation \tilde{y}_i , with expectation over the reference model $\mu_i^* = \mathbb{E}_{\tilde{y}|\lambda_*}(\tilde{y}_i)$, this reduces to (Piironen et al., 2020b):

$$\lambda_{\perp} = \arg \max_{\lambda} \sum_{i=1}^N \mu_i^* \xi_i(\lambda) - B(\xi_i(\lambda)), \quad (2)$$

which does not depend on the dispersion parameter ϕ , for some function B of the natural parameters $\xi_i(\lambda)$.

The above projection holds for observation models other than Gaussian or link functions other than the identity, as long as they belong to the exponential family. In these cases, there is no closed form solution for the projection parameters, and we run iteratively reweighted least squares (IRLS), where at every iteration one computes a pseudo Gaussian transformation of each log likelihood \mathcal{L}_i as a second order Taylor series expansion centered at the projection's prediction (McCullagh and Nelder, 1989, Gelman et al., 2013).

4 PROJECTION PREDICTIVE INFERENCE FOR GLMMS AND GAMMS

4.1 Generalized Linear Multilevel Models

GLMMs (McCulloch, 2003, Gelman et al., 2013) jointly estimate both *global population* and *group-specific* parameters. This approach allows the model to *partially pool* information across groups, which is particularly useful for the estimates of groups with few data points. In this case, we refer to multilevel structure as terms arising from the levels of a categorical variable and their interactions with other variables.

Given a response variable y as well as a population design matrix X and group design matrix Z (both being constructed from the predicting variables), we can write a GLMM as $y \sim \pi(g(\eta), \phi)$, where $g(\cdot)$ is the inverse link function of the generalized family π , and ϕ is its dispersion parameter. The only difference to a GLM comes in the linear predictor η that for GLMMs can be written as $\eta = X\beta + Zu$, where β are the population parameters and u are the group-specific parameters with $u \sim p(u | \theta)$, which may depend on some hyper parameters θ . Inference is concerned with accurately estimating the model parameters (β, u, ϕ, θ) .

4.2 Generalized Additive Multilevel Models

GAMMs (Hastie and Tibshirani, 1986, Verbyla et al., 1999) add further complexity to GLMMs by introducing smooth terms, which are presented as a linear combination of non-linear basis functions.

As for GLMMs, we can formulate the model as $y \sim \pi(g(\eta), \phi)$. In the case of generalized additive models (GAMs) without multilevel structure, the predictor η can be written as $\eta = \sum_{j=1}^J f_j(X)$, where each f_j is a function of the predictor matrix X (in practice, each f_j uses only a subset of columns of X). These functions are usually represented via additive spline basis expansion $f(x) = \sum_{k=1}^K \gamma_k b_k(x)$ with B-splines $b_k(x)$ (Eilers

and Marx, 1996). To avoid overfitting, we can either penalize some summary of the spline basis coefficients, or equivalently write it as a GLMM by splitting up the evaluated spline basis function into an unpenalized null-space (appended into X) and a penalized space (appended as group variables into Z) where the prior on u serves the same purpose (Wood, 2017). Standard multilevel terms of GLMMs can be combined with non-linear smooth terms of GAMs to form the even more powerful GAMM model class (Wood, 2017). However, as soon as smooth terms are added and translated to the GLMM framework, the resulting Z matrix becomes much denser than in a standard GLMM, thus further complicating inference.

4.3 Solving the Projection for GLMMs

Without further constraints, even if its observation model belongs to the exponential family, the projection (1) is not identifiable for GLMMs (Bickel and Doksum, 1977, Lee and Nelder, 1996, Gelman et al., 2013). This means that, given the mean prediction of the reference model μ_i^* , there is no unique solution for the parameters in the projection model fitted to μ_i^* .

To make the model identifiable and solve the projection, we propose to further restrict it by integrating out the group parameters u . The resulting likelihood can be written as:

$$\begin{aligned} \mathcal{L}(y | \beta, u, \phi) &= p(y | \beta, \phi, \theta) \\ &= \int p(y | \beta, \phi, u) p(u | \theta) du \\ &= \prod_i^N \int p(y_i | \beta, \phi_i, u) p(u_i | \theta) du_i, \end{aligned} \quad (3)$$

where u_i are the group parameters belonging to observation i and we have assumed conditional independence between datapoints given the group parameters. This integral cannot be evaluated in closed form. For simple models with a single group one can numerically integrate the above expression (employing Gaussian-Hermite quadrature) but this quickly becomes infeasible for higher dimensional problems. Maximizing this likelihood with respect to the model parameters, following Equation (2), gives the constrained projection.

There are many references in the literature that focus on practical approaches to obtain suitable approximate maximum likelihood estimates for GLMMs (McCulloch, 1997, Bates et al., 2015, Lee and Nelder, 2001, Lee et al., 2006, Ogden, 2013, Booth and Hobert, 1999). For our purposes, it is essential to use an approximate solution that still provides a good proxy for the KL divergence minimising solutions to Equation (1). Some of the methods cited above provide accurate and re-

liable solutions, but often at the expense of a higher computational cost (e.g., Ogden, 2013).

In the statistics literature, one finds simpler approximations that would scale better for our case, such as the well known Restricted Maximum Likelihood Estimates (REML) (Lee and Nelder, 2001, Lee et al., 2006). This approach obtains a solution by dealing with the group parameters as *fixed* data and appending them as an extension to the population parameters. One would then solve an augmented GLM.

4.4 Laplace Approximation

The REML approach does not provide a tractable approximation to the log marginal likelihood obtained from Equation (3), which takes the form

$$\log \mathcal{L}(\beta, u, \phi) = \sum_{i=1}^N \log \left\{ \int \exp(h_i) du_i \right\},$$

where $h_i = \log p(y_i | g(\eta_i), \phi_i) + \log p(u_i | \theta)$ is the (unnormalized) log joint density. The integrals in the above equation do not exist in closed form, except for models where $p(u | \theta)$ is conjugate to the likelihood π . Even in those cases, if the dimensionality of u is not very small, the computation is still intractable.

As a general purpose solution, we consider a first-order Laplace approximation to the integral (Ha, 2009, Barndorff-Nielsen and Cox, 1989, Bates et al., 2015). We split up the integration problem into sub-problems that are easier to solve. Given a value of θ , we can find the conditional mode $\tilde{u}(\theta)$ and conditional estimate $\tilde{\beta}(\theta)$ by solving the following optimization problem

$$\begin{bmatrix} \tilde{u}(\theta) \\ \tilde{\beta}(\theta) \end{bmatrix} = \arg \max_{u, \beta} h(u | y, \beta, \phi, \theta),$$

as the parameters that maximize the likelihood.

Usually, we express the conditional density on the *deviance scale*:

$$\begin{bmatrix} \tilde{u}(\theta) \\ \tilde{\beta}(\theta) \end{bmatrix} = \arg \min_{u, \beta} -2h(u | y, \beta, \phi, \theta).$$

This optimization problem can be solved efficiently using penalized iteratively re-weighted least squares (PIRLS), as implemented in the popular `lme4` (Bates et al., 2015) package. At each iteration, PIRLS performs a Gauss-Newton iteration in the space of u and β . See Bates et al. (2015) for more details.

The second order Taylor series expansion of $-2 \log h$ at $\tilde{u}(\theta)$ and $\tilde{\beta}(\theta)$ provides the Laplace approximation to the *profiled deviance*. On the deviance scale, the Laplace approximation is a function of the so-called

discrepancy measure and takes a sum of squares form:

$$d(u | y, \theta, \beta) = \left\| W^{1/2}(\mu) [y - \mu(u, \theta, \beta)] \right\|^2 + \|u\|^2,$$

where $\mu = g(\eta(u, \theta, \beta))$ is the inverse link transformation of the latent predictor η , and W is a diagonal matrix of weights. Optimizing this function with respect to θ provides the maximum likelihood estimates of β, ϕ by substituting θ_{ML} into $\tilde{\beta}(\theta)$ and solving for ϕ in $h(u | y, \beta, \theta_{\text{ML}}, \phi)$. Importantly, optimizing the Laplace approximation is a problem in the space of constrained θ , which is usually small and therefore easy to solve efficiently.

4.5 Solving the Projection for GAMMs

The identifiability issue that exists in GLMMs is further aggravated by having a dense Z matrix in GAMMs, which makes the likelihood in Equation (3) intractable to compute even in conjugate Gaussian models with a single smooth term. This also happens in GAMs without any multilevel structure.

In order to make these models identifiable, one has to 1) impose a quadratic penalty on the coefficients of the basis functions (Wood, 2017), which also helps in avoiding overfitting, and 2) integrate out the group parameters and group smooth terms. Solving the resulting maximum likelihood equations has similar issues as in the plain GLMM case. Given that GAMMs can be represented as GLMMs, the same Laplace approximation is commonly used to obtain maximum likelihood estimates in these models (Wood, 2010, 2017).

4.6 Consistency of the Selection

A variable selection is consistent if the probability of selecting the best model is 1 as the number of data points grows to infinity as long as the number of variables grows slower. Here, we define the best model according to a given utility function, which may favour specific properties, such as low number of covariates. The consistency of projection predictive inference is theoretically justified in Vehtari and Ojanen (2012).

4.7 Computational Cost

The computational budget of our approach is composed of the following components

- Running PIRLS, i.e. solving the projection, for a given subset of terms.
- Solving the same projection for a number of draws.
- Performing forward search to explore the model space.
- Running LOO cross validation for many data points.

The biggest limiting factors when applying projection predictive inference in complex models is twofold: 1) the complexity of the multilevel structure, which adds many parameters to estimate, and 2) the number of draws to project from the reference model, since the computational budget grows linearly with the number of draws.

Piironen et al. (2020b) demonstrate that a small number of posterior draws is sufficient to find a good solution path, which saves a lot of computation during the forward search. One can then increase the number of draws for the final computation of the projections for more accurate predictive performance. Nonetheless, running PIRLS for complex multilevel models is still expensive, especially when done repeatedly for different posterior draws.

Further, in our approach, we reduce the number of models to explore in forward search by considering only those that are sensible according to common modelling practices for GLMMs (Gelman and Hill, 2006). This means that we only consider a model with a certain group parameter if its population parameter has already entered the projection. Likewise, we only consider an interaction between two variables when both of them have already entered the projection separately. This saves the method from exploring many models that are not considered sensible in the first place.

4.8 Diagnosing the Approximation

Once we have computed the maximum likelihood estimates of a given submodel, we can diagnose the approximation by checking:

1. The Kullback-Leibler divergence from the reference model predictions to the submodel’s. As projection size increases, the KL divergence approaches 0 if the approximation is well-behaved.
2. The residual $y - \hat{y}$ histogram should resemble a Normal distribution centred around 0, and narrows down as projection size increases.

5 EXPERIMENTS

We now turn to validating our method with both simulated experiments and real world datasets.

5.1 Simulations

We first validate our method by running projection predictive variable and structure selection as implemented in `projpred` (Piironen et al., 2020a) on extensive simulations. We systematically test simple and more complex models with both an increasing number of grouping factors and variables. The complete

Table 1: Data generation process settings for the simulations.

Abbr	Description	Values
D	# of variables	5, 7, 10
V	% of group parameters	0.33, 0.67, 1.0
K	# of grouping factors	1, 2, 3
ρ	Correlation	0.0, 0.33, 0.67, 0.9
L	Levels per grouping	5
N	Observations	300
π	Model	Gaussian, Bernoulli
s	Sparsity	0.4

settings of the simulations are shown in Table 1.

To reduce external variation, we fix the number of observations to 300, the number of levels in each grouping factor to 5 and the sparsity to 0.4 (40% of terms being zero) to make sure that some terms are irrelevant. For each simulation condition, we run 25 data realisations. The running time for these experiments ranges from a few seconds for the simplest simulations to a few minutes for the most complex ones.

The complete data generation process, common for all observation models π , is given as

$$\begin{aligned}
 x_{id} &\sim \text{Normal}(0, 1) & \beta_d &\sim \text{Normal}(\mu_{b,f}, \sigma_{b,f}^2) \\
 z_d &\sim \text{Bernoulli}(p = 0.6) \\
 g_{ik} &\sim \text{DiscreteUniform}(1, L) \\
 \mu_{g_k} &\sim \text{MultivariateNormal}(\mu_g, \sigma_g^2 I) \\
 v_d &\sim \text{Bernoulli}(p = V) \\
 u_{lk} &\sim \text{MultivariateNormal}(\mu_{g_k}, \Sigma_{\sigma_g^2, \rho}) \\
 \eta_i &= \sum_{d=0}^D z_d \beta_d x_{id} + \sum_{k=1}^K \sum_{l=1}^L \sum_{d=0}^D v_d z_d u_{lkd} x_{g_{ik}=l,d} \\
 y_i &\sim \pi(g(\eta_i), \phi_i),
 \end{aligned}$$

for data points $i = 1, \dots, N$, variables $d = 0, \dots, D$, grouping factors $k = 1, \dots, K$, inverse link function g , and covariance matrix $\Sigma_{\sigma_g^2, \rho}$ with diagonal entries σ_g^2 and off diagonal elements $\rho\sigma_g^2$. Note that we collapse the intercept ($d = 0$) into β , u , so that we have $D + 1$ dimensions by appending a column of ones to X . We choose the identity link function. We fix the mean and standard deviation hyperparameters for the intercept μ_f, σ_f^2 to 0, 20 respectively, for the main terms $\mu_{b,f}, \sigma_{b,f}^2$ to 5, 10 respectively and for the group terms μ_g, σ_g^2 to 0, 5 respectively. We choose large values to avoid simulating practically undetectable terms.

We sample group terms following a two-step procedure:

1. We first sample K means for all grouping factors from a $D + 1$ dimensional multivariate normal

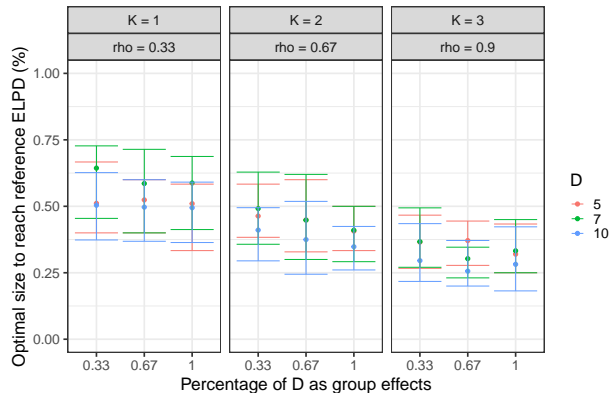


Figure 2: Optimal size for the projections to achieve the reference LOO performance (ELPD). Each column shows a different correlation and number of grouping factors. We include 95% uncertainty interval for the size of the projection.

distribution as

$$\mu_{g_k} \sim \text{MultivariateNormal}(\mu_g, \sigma_g^2 I).$$

- Then, for each level l and grouping factor k , we sample from a $D + 1$ dimensional multivariate normal (with mean μ_{g_k}) as

$$u_{lk} \sim \text{MultivariateNormal}(\mu_{g_k}, \Sigma_{\sigma_g^2, \rho}).$$

We divide the analysis of these experiments in two aspects. First, we focus on the predictive performance of our method in terms of projection’s model size and expected log predictive density (ELPD). Second, we analyse the solution terms by checking the true and false positives rate for different selection thresholds.

Figure 2 shows the size of the optimal projections that achieve the reference ELPD for Gaussian simulated data. For easier comparison, we show the projection’s optimal size as a percentage of the reference model’s size. Most importantly, the optimal projections always achieve the reference predictive performance with only a subset of terms. The exact size will depend on both the correlation factor and the number of actually irrelevant terms. Higher correlations correspond to smaller optimal subsets, as there are more redundant terms that can be discarded. In this regard, `projpred` is able to identify a subset of *predictive* relevant terms, which is the main goal of our approach.

We now analyse the selected optimal terms from the selection point of view. As the projection is based on the reference model’s posterior predictive draws, it will select terms that are relevant from the predictive perspective. Thus, it is possible that some selected terms are not *truly* relevant when compared to the true data generation process. Naturally, that is not usually

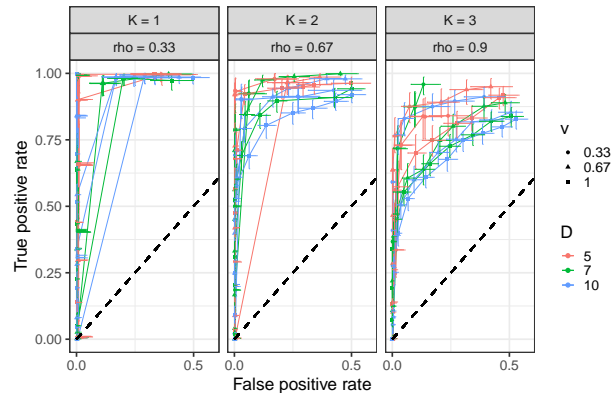


Figure 3: False against true positive rate in a Gaussian model for varying selection thresholds. Each column shows a different correlation and number of grouping factors. We include 95% uncertainty interval for both true and false positive rates. Chance selection in dashed black lines.

feasible in real world scenarios, as the existence of a *true model* is not clear in the first place. However, this validation is sensible for simulated data, and we show true against false positives rates in Figure 3.

To compute true and false positive rates, we select relevant terms by looking at the ELPD improvement of each projection with respect to the previous one. For a threshold $t \in [0, 1]$, we select all terms whose projection’s ELPD improvement is above the t th quantile of all. Then, we compare the selected terms against the ground truth. For models with only a few terms, the discrete set of thresholds results in straight jumps in the figure.

As the reference model grows in complexity, either by adding more variables D , grouping factors K or percentage of group terms V , we expect the method to return more false positives. This is reflected on the figure, as more complex models are usually closer to the identity line, with the exception of very simple models. The number of grouping factors K has the largest impact on the performance, as it multiplies the number of actual coefficients to estimate in the model. While higher correlation factors induced smaller optimal models, they also induce more false positives. Highly correlated variables contain very similar information, and the method has a hard time selecting the *true* one.

In the supplementary material we include more experiments that include Bernoulli simulations, more levels per grouping factor and more sparsity thresholds.

5.2 Real Data Experiments

We now validate the performance of our method in real world datasets, including a Bernoulli classification model and a Poisson count data model. The overall running time of this experiments is about 30 minutes, except for the more complex `BikeSharing` GAMM, which takes about two hours.

5.2.1 Bernoulli Classification Model

We fit a Bernoulli model to the `VerbAgg` (Bates et al., 2015) dataset. This dataset includes item responses to a questionnaire on verbal aggression, used throughout De Boeck and Wilson (2004) to illustrate various forms of item response models. It consists of 7584 responses of a total of 316 participants on 24 items. We randomly draw 50 individuals and their responses to increase the difficulty of the selection.

Following R’s `formula` syntax, we fit the reference model $r2 \sim \text{btype} + \text{mode} + \text{situ} + (\text{btype} + \text{mode} + \text{situ} \mid \text{id})$, which includes `btype`, `situ` and `mode` as group terms across participants. The full reference model contains 7 terms, counted as the simplest individual components of the model excluding the global intercept, namely `btype`, `mode`, `situ`, $(\text{btype} \mid \text{id})$, $(\text{mode} \mid \text{id})$, $(\text{situ} \mid \text{id})$ and $(1 \mid \text{id})$. We fit the reference model with a regularised horseshoe prior (with global scale 0.01) (Piironen and Vehtari, 2017b, Carvalho et al., 2010) with `rstanarm` (Goodrich et al., 2018) and run a LOO-cross validated-selection procedure (Piironen et al., 2020a, Vehtari et al., 2016).

We show ELPD summaries and their standard error in Figure 4a for incremental projections. The red dash line shows ELPD performance for the full reference model. The optimal projection includes the first 6 terms, resulting in the model $r2 \sim \text{btype} + \text{mode} + \text{situ} + (\text{btype} + \text{situ} \mid \text{id})$, implying that almost all terms are relevant. However, the standard error increases for the last two terms, indicating a slight identifiability issue for those parameters, possibly caused by higher correlations between their estimates. For more accurate projections, one could project more posterior draws or improve the reference model (e.g., with some PCA-like approach).

5.2.2 Poisson GLMM Model

We fit a Poisson model to the `BikeSharing` dataset (Fanaee-T and Gama, 2013). These data contain the hourly and daily count of rental bikes between 2011 and 2012 in London’s capital bikeshare system with the corresponding weather and season information. We only use the daily averaged dataset, with 731 observations for 2 years. We build a knowingly overly complicated

model, including highly correlated group parameters for different grouping factors. The reference model includes population and group terms for `windspeed`, `temp` and `humidity` across the grouping factors `month`, `weekday`, `weather`, `season` and `holiday`.

We fit the reference model with a regularised horseshoe prior (with global scale 0.01) using `rstanarm` (Goodrich et al., 2018) and run `projpred`’s LOO-cross-validated selection procedure.

We show ELPD summaries for each incremental projection in Figure 4b. The optimal projection is $\text{count} \sim \text{temp} + \text{humidity} + \text{windspeed} + (\text{humidity} + \text{temp} \mid \text{month}) + (1 \mid \text{weather}) + (1 \mid \text{season})$, including only 8 terms out of 23.

5.2.3 Poisson GAMM Model

We also used the same data and model as in the example above to build a GAMM reference model that now adds smooth terms for the population terms `windspeed`, `temp` and `humidity`. We also include group intercepts for all grouping factors. Note that, for computational reasons, this model is simpler than the full GLMM we have considered in the previous experiment. We make use of `rstanarm` again to build this model with a normal prior and run `projpred`’s LOO-cross-validated selection.

We show ELPD summaries for each incremental projection in Figure 4c. The optimal projection is $\text{count} \sim \text{s}(\text{temp}) + \text{s}(\text{humidity}) + \text{s}(\text{windspeed}) + (1 \mid \text{season})$, including only 4 out of 8 terms. The final projection is smaller than the optimal GLMM projection in the previous experiment, and still achieves a very similar performance.

6 CONCLUSION

In this work, we have extended projection predictive inference to variable and structure selection in complex model classes, namely GLMMs and GAMMs. For these models, the original GLM projection predictive approach cannot be directly translated as it would result in unidentifiable models. To resolve this issue, we proposed to combine projection predictions with a Laplace approximation and demonstrated that this not only enables accurate variable selection but also scales well to larger number of variables and grouping factors. We have validated our proposal by performing extensive simulations that test the boundaries of our method in extreme settings. We also showed that our method works well in real world scenarios with highly correlated grouping factors. We leave the extension of our current framework to other models which do not belong to the exponential family for future work.

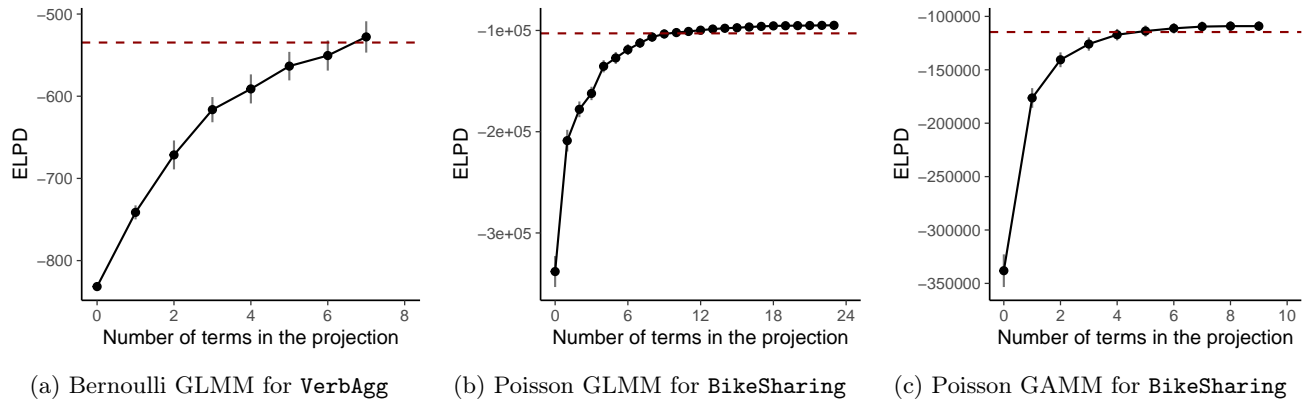


Figure 4: Incremental ELPD summaries of real data experiments.

In such cases, the KL minimization in Equation (1) does not coincide with maximum likelihood estimates anymore.

Acknowledgements

We would like to thank Akash Dhaka, Kunal Gosh, Charles Margossian and Topi Paananen for helpful comments and discussions. We also thank Frank Weber for helping in the development of the code package. We also acknowledge the computational resources provided by the Aalto Science-IT project.

References

- H. Afrabandpey, T. Peltola, J. Piironen, A. Vehtari, and S. Kaski. A decision-theoretic approach for model interpretability in Bayesian framework. *Machine Learning*, 109(9-10):1855–1876, Sep 2020. ISSN 1573-0565. doi: 10.1007/s10994-020-05901-8. URL <http://dx.doi.org/10.1007/s10994-020-05901-8>.
- M. M. Barbieri and J. O. Berger. Optimal predictive model selection. *The Annals of Statistics*, 32(3): 870–897, Jun 2004. ISSN 0090-5364. doi: 10.1214/009053604000000238. URL <http://dx.doi.org/10.1214/009053604000000238>.
- O. E. Barndorff-Nielsen and D. R. Cox. *Asymptotic techniques for use in statistics*. London ; New York : Chapman and Hall, 1989. ISBN 0412314002.
- D. Bates, M. Mächler, B. Bolker, and S. 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>.
- P. J. Bickel and K. A. Doksum. *Mathematical statistics: Ideas and concepts*. San Francisco: Holden-Day, 1977.
- J. G. Booth and J. P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):265–285, Feb 1999. ISSN 1467-9868. doi: 10.1111/1467-9868.00176. URL <http://dx.doi.org/10.1111/1467-9868.00176>.
- L. Breiman. Better subset regression using the non-negative garrote. *Technometrics*, 37(4):373–384, 1995. ISSN 00401706. URL <http://www.jstor.org/stable/1269730>.
- P.-C. Bürkner. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1): 395–411, 2018. doi: 10.32614/RJ-2018-017.
- E. Candes and T. 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>.
- C. M. Carvalho, N. G. Polson, and J. 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>.
- P. De Boeck and M. Wilson. *Explanatory item response models: A generalized linear and nonlinear approach*. Springer Science & Business Media, 2004.
- P. H. C. Eilers and B. D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121, May 1996. ISSN 0883-4237. doi: 10.1214/ss/1038425655. URL <http://dx.doi.org/10.1214/ss/1038425655>.
- J. Fan and R. 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>.

- H. Fanaee-T and J. Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, pages 1–15, 2013. ISSN 2192-6352. doi: 10.1007/s13748-013-0040-3. URL [WebLink].
- A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006. ISBN 9780511790942. doi: 10.1017/cbo9780511790942. URL <http://dx.doi.org/10.1017/cbo9780511790942>.
- A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013. ISBN 9781439840955.
- E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, Sep 1993. ISSN 1537-274X. doi: 10.1080/01621459.1993.10476353. URL <http://dx.doi.org/10.1080/01621459.1993.10476353>.
- B. Goodrich, J. Gabry, I. Ali, and S. Brilleman. rstanarm: Bayesian applied regression modeling via Stan., 2018. URL <http://mc-stan.org/>. R package version 2.17.4.
- C. P. Goutis, Constantinos 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>.
- A. Groll and G. Tutz. Variable selection for generalized linear mixed models by l1-penalized estimation. *Statistics and Computing*, 24(2):137–154, Oct 2012. ISSN 1573-1375. doi: 10.1007/s11222-012-9359-z. URL <http://dx.doi.org/10.1007/s11222-012-9359-z>.
- I.-D. Ha. *Maximum Likelihood Estimation Using Laplace Approximation in Poisson GLMMs*, volume 16. The Korean Statistical Society, Korean International Statistical Society, 2009.
- T. 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>.
- T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–310, Aug 1986. ISSN 0883-4237. doi: 10.1214/ss/1177013604. URL <http://dx.doi.org/10.1214/ss/1177013604>.
- A. Hubin and G. Storvik. Mode jumping mcmc for bayesian variable selection in glmm. *Computational Statistics Data Analysis*, 127:281–297, 2018. ISSN 0167-9473. doi: <https://doi.org/10.1016/j.csda.2018.05.020>. URL <https://www.sciencedirect.com/science/article/pii/S016794731830135X>.
- H. Ishwaran and J. S. 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>.
- V. E. Johnson and D. Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660, Jun 2012. ISSN 1537-274X. doi: 10.1080/01621459.2012.682536. URL <http://dx.doi.org/10.1080/01621459.2012.682536>.
- R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, Jun 1995. ISSN 1537-274X. doi: 10.1080/01621459.1995.10476572. URL <http://dx.doi.org/10.1080/01621459.1995.10476572>.
- Y. Lee and J. A. Nelder. Hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(4):619–656, 1996. doi: 10.1111/j.2517-6161.1996.tb02105.x. URL <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02105.x>.
- Y. Lee and J. A. Nelder. Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88(4):987–1006, Dec 2001. ISSN 1464-3510. doi: 10.1093/biomet/88.4.987. URL <http://dx.doi.org/10.1093/biomet/88.4.987>.
- Y. Lee, J. A. Nelder, and Y. Pawitan. *Generalized Linear Models with Random Effects*. Chapman and Hall/CRC, Jul 2006. ISBN 9780429144714. doi: 10.1201/9781420011340. URL <http://dx.doi.org/10.1201/9781420011340>.
- G. Marra and S. N. Wood. Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, 55(7):2372–2387, Jul 2011. ISSN 0167-9473. doi: 10.1016/j.csda.2011.02.004. URL <http://dx.doi.org/10.1016/j.csda.2011.02.004>.
- P. McCullagh and J. 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>.
- C. E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92(437):162–170, Mar 1997. ISSN 1537-274X. doi: 10.1080/01621459.1997.10473613. URL <http://dx.doi.org/10.1080/01621459.1997.10473613>.

- C. E. McCulloch. Generalized linear mixed models. *NSF-CBMS Regional Conference Series in Probability and Statistics*, 7:i–84, 2003. ISSN 19355920, 23290978. URL <http://www.jstor.org/stable/4153190>.
- N. N. Narisetty and X. He. Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817, Apr 2014. ISSN 0090-5364. doi: 10.1214/14-aos1207. URL <http://dx.doi.org/10.1214/14-aos1207>.
- H. Ogden. A sequential reduction method for inference in generalized linear mixed models. *arXiv preprint 1312.1903v2*, 2013.
- F. Pavone, J. Piironen, P.-C. Bürkner, and A. Vehtari. Using reference models in variable selection. *arXiv preprint arXiv:2004.13118*, 2020.
- J. Piironen and A. Vehtari. Projection predictive model selection for Gaussian processes. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2016. doi: 10.1109/MLSP.2016.7738829.
- J. Piironen and A. 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>.
- J. Piironen and A. 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>.
- J. Piironen, M. Paasiniemi, A. Catalina, and A. Vehtari. *projpred: Projection Predictive Feature Selection*, 2020a. <https://mc-stan.org/projpred>, <https://discourse.mc-stan.org/>.
- J. Piironen, M. Paasiniemi, and A. 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>.
- A. E. Raftery, D. Madigan, and J. A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, Mar 1997. ISSN 1537-274X. doi: 10.1080/01621459.1997.10473615. URL <http://dx.doi.org/10.1080/01621459.1997.10473615>.
- R. 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>.
- G. Tutz and A. Groll. Regularization for generalized additive mixed models by likelihood-based boosting. *Methods of Information in Medicine*, 51(02):168–177, 2012. ISSN 2511-705X. doi: 10.3414/me11-02-0021. URL <http://dx.doi.org/10.3414/me11-02-0021>.
- A. Vehtari and J. 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>.
- A. Vehtari, A. Gelman, and J. Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, Aug 2016. ISSN 1573-1375. doi: 10.1007/s11222-016-9696-4. URL <http://dx.doi.org/10.1007/s11222-016-9696-4>.
- A. Vehtari, D. Simpson, A. Gelman, Y. Yao, and J. Gabry. Pareto smoothed importance sampling. *arxiv preprint:1507.02646*, 2019.
- A. P. Verbyla, B. R. Cullis, M. G. Kenward, and S. J. Welham. The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):269–311, Aug 1999. ISSN 1467-9876. doi: 10.1111/1467-9876.00154. URL <http://dx.doi.org/10.1111/1467-9876.00154>.
- S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36, Sep 2010. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2010.00749.x. URL <http://dx.doi.org/10.1111/j.1467-9868.2010.00749.x>.
- S. N. Wood. *Generalized Additive Models*. Chapman and Hall/CRC, May 2017. ISBN 9781315370279. doi: 10.1201/9781315370279. URL <http://dx.doi.org/10.1201/9781315370279>.
- H. Zou and T. 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>.

Supplementary Material: Projection Predictive Inference for Generalized Linear and Additive Multilevel Models

A ADDITIONAL RESULTS FOR MORE SPARSITY THRESHOLDS

In this section we provide more results for Gaussian simulated data with group sparsity thresholds $s = \{0, 0.11, \dots, 0.89\}$. To isolate the effect of the sparsity we fix the number of variables to $D = 5$ and $K = 1$ grouping factors. For a more complex setting, we allow all $D = 5$ variables to vary within the grouping factor.

In Figure 5a we show the ROC curve for this experiment's results. To trace each sparsity's ROC curve we vary the selection threshold for relevant group effects. As we did in the main text, we decide which group effects are relevant by looking at the ELPD improvement of a projection with respect to the previous one after including a given group effect. For a selection threshold t , we select as relevant all group effects whose projection's ELPD improvement is over the t th quantile of all ELPD improvements. This gives a curve starting at 0% true positive rate when $t = 0$, to 100% for $t = 1$ for every sparsity threshold.

Nonetheless, the sparsity threshold inherently sets the portion of actually relevant group effects. As the sparsity increases, the number of relevant group effects gets lower, and therefore the false positives rate can only increase as we return more and more group effects as relevant. On the other hand, for low sparsity thresholds we have the opposite behaviour, as there are many actually relevant group effects, and therefore the false positive rate gets lower as we select more group effects. The most difficult setting corresponds to medium sparsity thresholds, and that is reflected in the figure with higher false positive rates.

B ADDITIONAL RESULTS FOR MORE LEVELS

In this case we focus on studying the effect of larger number of levels in each grouping factor. We fix the number of variables $D = 5$ and $K = 1$ grouping factor. We allow only 0.33 portion of $D = 5$ total variables to be group effects. We now vary the number of levels in $\{5, 10, 20, 50\}$. To avoid other factors to intervene, we simulate Gaussian observations.

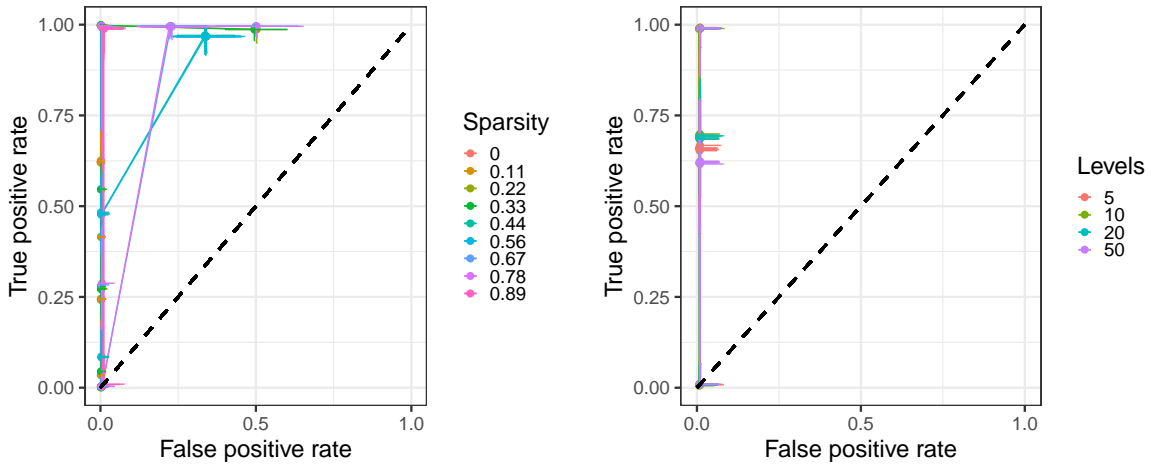
We show ROC curves for all choices of levels in Figure 5b. We use the same mechanism to decide relevant group effects and contrast them to the ground truth. In this case it's clear that the number of levels is not inducing more false positives to be returned, rather affecting only the true positives rate and how quickly the method is able to get all of them right.

C BERNOULLI SIMULATIONS

Finally, we show optimal projection size and ROC curve results for Bernoulli simulated data following the same structure as in the main text. In this case we choose a `logit` link function. For non-Gaussian observation models there is an added challenge as the projection is computing a linear Gaussian model as an approximation to the non-Gaussian observation model. This increases the computational budget of the method and sometimes results in unstable projections, even though still resulting in good performance overall. The source of the instabilities is the complexity of the link function. This kind of models require more posterior draws to be projected in order to maintain accurate predictive performance, which in turn increases the computational budget. The increase in posterior draws can multiply the runtime by a factor of four, from about half an hour to couple hours.

On top of that, the hyper parameters we have used on the sampling procedure for Bernoulli data imply smaller coefficients (see Table 2). This is due the further complexity added by the link function, that usually causes extreme values for large effects.

We follow the same analysis structure as in the main text. We first focus on the optimality of the projections and



(a) True positive rate against false positive rate in a Gaussian model for varying sparsity thresholds. We include 95% uncertainty interval for both true and false positive rates. Chance selection in dashed black lines.
 (b) True positive rate against false positive rate in a Gaussian model for varying number of levels. We include 95% uncertainty interval for both true and false positive rates. Chance selection in dashed black lines.

Figure 5: Additional results for Gaussian simulations.

Table 2: Hyper parameters for the Bernoulli simulations. They are propagated to a constant vector of the corresponding value.

Parameter	Value
μ_f, σ_f^2	0, 4
$\mu_{b,f}, \sigma_{b,f}^2$	0, 2
μ_g, σ_g^2	0, 3

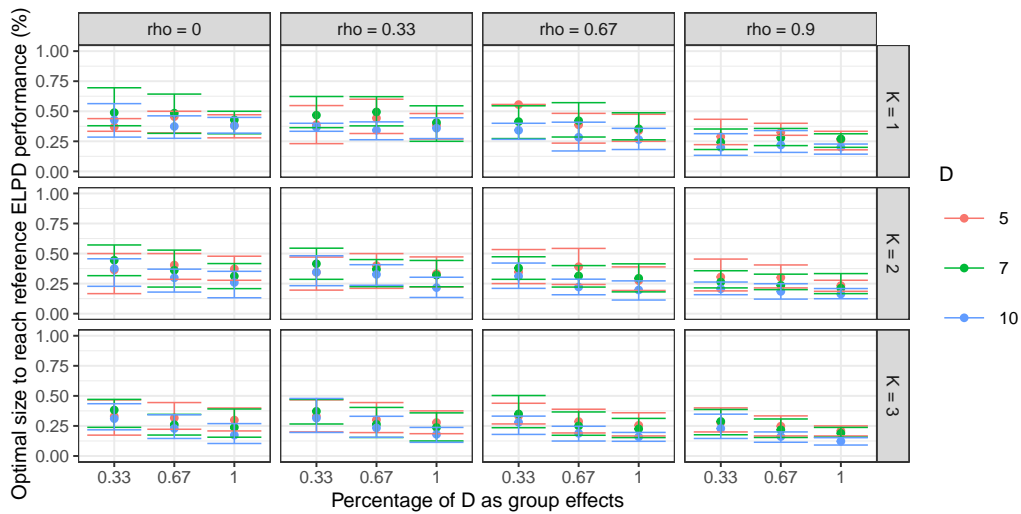


Figure 6: Optimal size for the projections to achieve the reference LOO performance (ELPD). Each column shows a different correlation factor. Each row shows a different number of grouping factors. We include 95% uncertainty interval for the size of the projection.

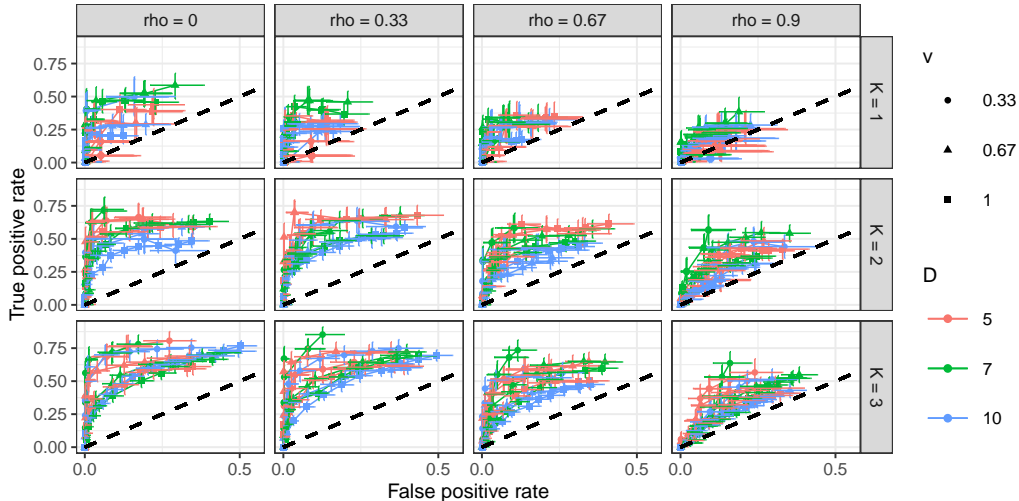


Figure 7: False against true positive rate in a Bernoulli model for varying selection thresholds. Each column shows a different correlation factor. Each row shows a different number of grouping factors. We include 95% uncertainty interval for both true and false positive rates. Chance selection in dashed black lines.

analyze the optimal projection size that the method suggests to achieve the reference performance. We show these results in Figure 6.

The projections achieve optimal predictive performance. The results show that the optimal projection is able to discard on average 60% of the terms while achieving the reference predictive performance. This proportion of relevant terms is dependent on the sparsity threshold selected as analyzed in Section A. These results hold even for larger number of grouping factors.

The projections are smaller with higher correlation. In presence of highly correlated variables the optimal projection is able to identify a smaller subset of relevant variables that achieve the reference predictive projection. As a result, the optimal size decreases as we increase the overall correlation.

The projections are robust. Even as we increase the complexity of the reference models, the method is able to find optimal projections every time. The robustness of the results is indicated by the small confidence interval around the mean estimates.

Now, we analyze the performance of the method regarding the proper identification of the truly relevant simulated terms. We show these results in Figure 7. We employ the same method to compute true and false positive rates for these experiments as we did in the main text for Gaussian data. That is, for a varying threshold $t \in [0, \dots, 1]$, we select as relevant those effects that imply an ELPD improvement over the t th quantile of all ELPD improvements after being included in the projection.

The number of variables D has a moderate effect. Increasing the number of variables in the data has a direct effect on the complexity of the reference model. This is also reflected in our results, as the ratio of true vs false positive is better for smaller number of variables. Nonetheless, the overall performance stays quite close.

The number of grouping factors K is the most significant factor. Increasing the number of grouping factors in the data multiplies the number of total parameters to estimate. As a result, the individual contribution of each parameter is diluted, which makes its identification even more opaque. In the figure we can clearly see how the ratio of true vs false positives rates gets closer to random choice as we increase the number of grouping variables.

The percentage of group terms V is important. As we increase the number of terms that vary within grouping factors we expect the method to confound the relevance of each of them, as the variability of the data

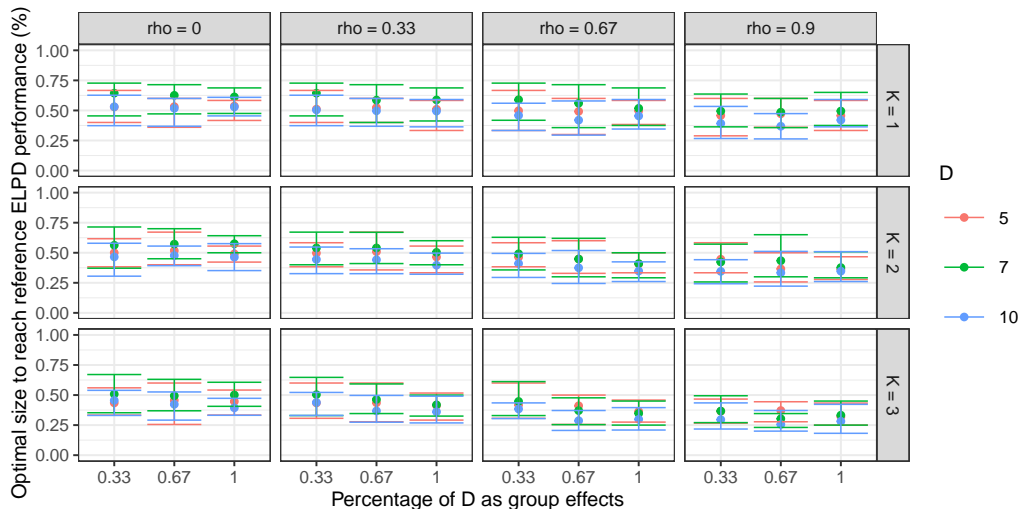


Figure 8: Optimal size for the projections to achieve the reference LOO performance (ELPD). Each column shows a different correlation factor. Each row shows a different number of grouping factors. We include 95% uncertainty interval for the size of the projection.

increases. We can see, in turn, that the method returns more false positives for higher proportions of group terms.

High correlation ρ induces more false positives. Another important factor that induces more false positives is the correlation between terms. This adds further complexity to the identification of truly relevant terms and is clearly reflected in our results, as the ratio of true vs false positive rates worsens for increasing correlation factors. Note that, even in the extreme unrealistic case of 0.9 correlation, the method still identifies around half of the truly relevant effects.

In general, the method does a good job at obtaining optimal projections with the smallest possible subset of relevant terms that still achieves the reference performance.

C.1 Full Gaussian Figures

In this section we show full sized figures of the Gaussian experiments in the main text.

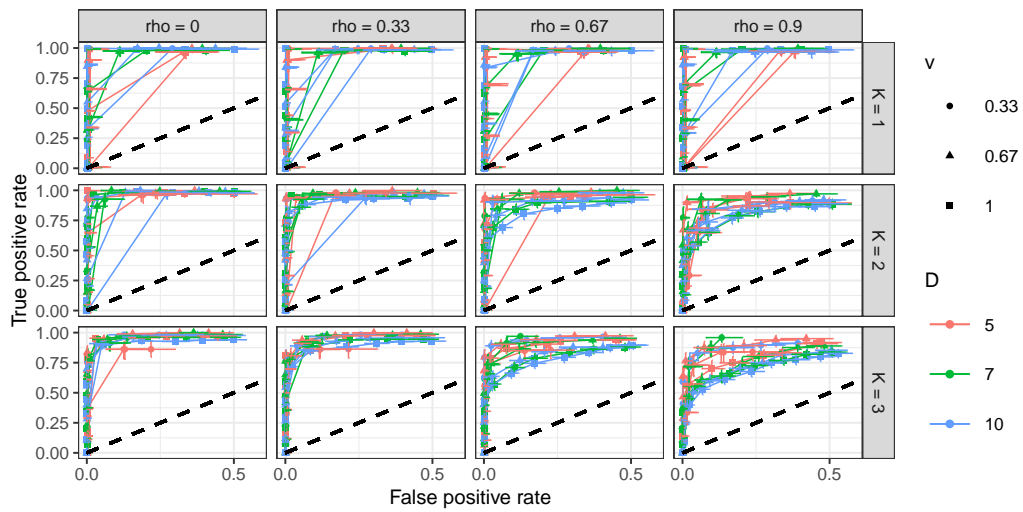


Figure 9: False against true positive rate in a Gaussian model for varying selection thresholds. Each column shows a different correlation factor. Each row shows a different number of grouping factors. We include 95% uncertainty interval for both true and false positive rates. Chance selection in dashed black lines.