Bayesian Model Selection with Latent Group-Based Effects and Variances with the R Package slgf

by Thomas A. Metzger and Christopher T. Franck

Abstract Linear modeling is ubiquitous, but performance can suffer when the model is misspecified. We have recently demonstrated that latent groupings in the levels of categorical predictors can complicate inference in a variety of fields including bioinformatics, agriculture, industry, engineering, and medicine. Here we present the R package slgf which enables the user to easily implement our recently-developed approach to detect group-based regression effects, latent interactions, and/or heteroscedastic error variance through Bayesian model selection. We focus on the scenario in which the levels of a categorical predictor exhibit two latent groups. We treat the detection of this grouping structure as an unsupervised learning problem by searching the space of possible groupings of factor levels. First we review the suspected latent grouping factor (SLGF) method. Next, using both observational and experimental data, we illustrate the usage of slgf in the context of several common linear model layouts: one-way analysis of variance (ANOVA), analysis of covariance (ANCOVA), a two-way replicated layout, and a two-way unreplicated layout. We have selected data that reveal the shortcomings of classical analyses to emphasize the advantage our method can provide when a latent grouping structure is present.

1 Introduction

Linear models with categorical predictors (i.e., factors) are pervasive in the social, natural, and engineering sciences, among other fields. Conventional approaches to fit these models may fail to account for subtle latent structures, including latent regression effects, interactions, and heteroscedasticity within the data. These latent structures are frequently governed by the levels of a factor. Several examples of such datasets can be found in Franck et al. (2013), Franck and Osborne (2016), Kharrati-Kopaei and Sadooghi-Alvandi (2007), and Metzger and Franck (2021). Our recent work (Metzger and Franck, 2021) developed latent grouping factor-based methodology to detect latent structures using Bayesian model selection. The current work provides an overview of the slgf package that enables users to easily implement the suspected latent grouping factor (SLGF) methodology, and expands on the previous work by allowing for more flexible model specification.

Consider Figure 1, which illustrates four relevant data sets analyzed in this paper. In each panel, the levels of a user-specified factor are found to exhibit a latent grouping structure that partitions the data into two groups with distinct regression effects (indicated by color-coding) and/or error variances (filled and open geometry). With the slgf package, the user specifies the factor suspected of governing this latent structure. The package protects the user against detecting spurious latent grouping structures since it can accommodate non-grouped candidate models. It can also accommodate additional linear model terms of interest. The slgf package then assesses the plausibility of each model and the corresponding structures via Bayesian model selection. An overview of slgf functionality for these data follows and full details of each analysis (including candidate models) appear in Section **Using the slgf package**. The slgf package focuses on assessing the plausibility of two-group structures in linear models with categorical predictors using fractional Bayes factors. A discussion comparing slgf and other R packages that address latent group models is in Section **Conclusion**.

The top left panel of Figure 1 represents a one-way analysis of variance (ANOVA) study where a continuous measurement of olfactory function (vertical axis) is modeled as a function of age, where age is a factor represented in five categories (horizontal axis) (O'Brien and Heft, 1995). We find the highest posterior model probability (61%) for the model where levels 1, 2, and 3 of the SLGF age have distinct mean effects and error variances from levels 4 and 5. We call this the `smell` data set.

The top right panel shows an analysis of covariance (ANCOVA), where the breaking strength of a starch film (vertical axis) is measured as a function of the SLGF (starch type) and a continuous measurement of film thickness (horizontal axis) (Furry, 1939). We find the highest posterior model probability (59%) for the model where potato starch (unshaded gray squares) have a larger error variance than the shaded points, and, the red points (canna and corn starch) have a distinct slope from the gray points. We call this the `textile` data set.

The bottom left panel shows the example described by Meek and Ozgur (1991), where the torque required to tighten a locknut (vertical axis) was measured as a function of a plating process and a threading technique. The plating processes analyzed included treatments with cadmium and wax
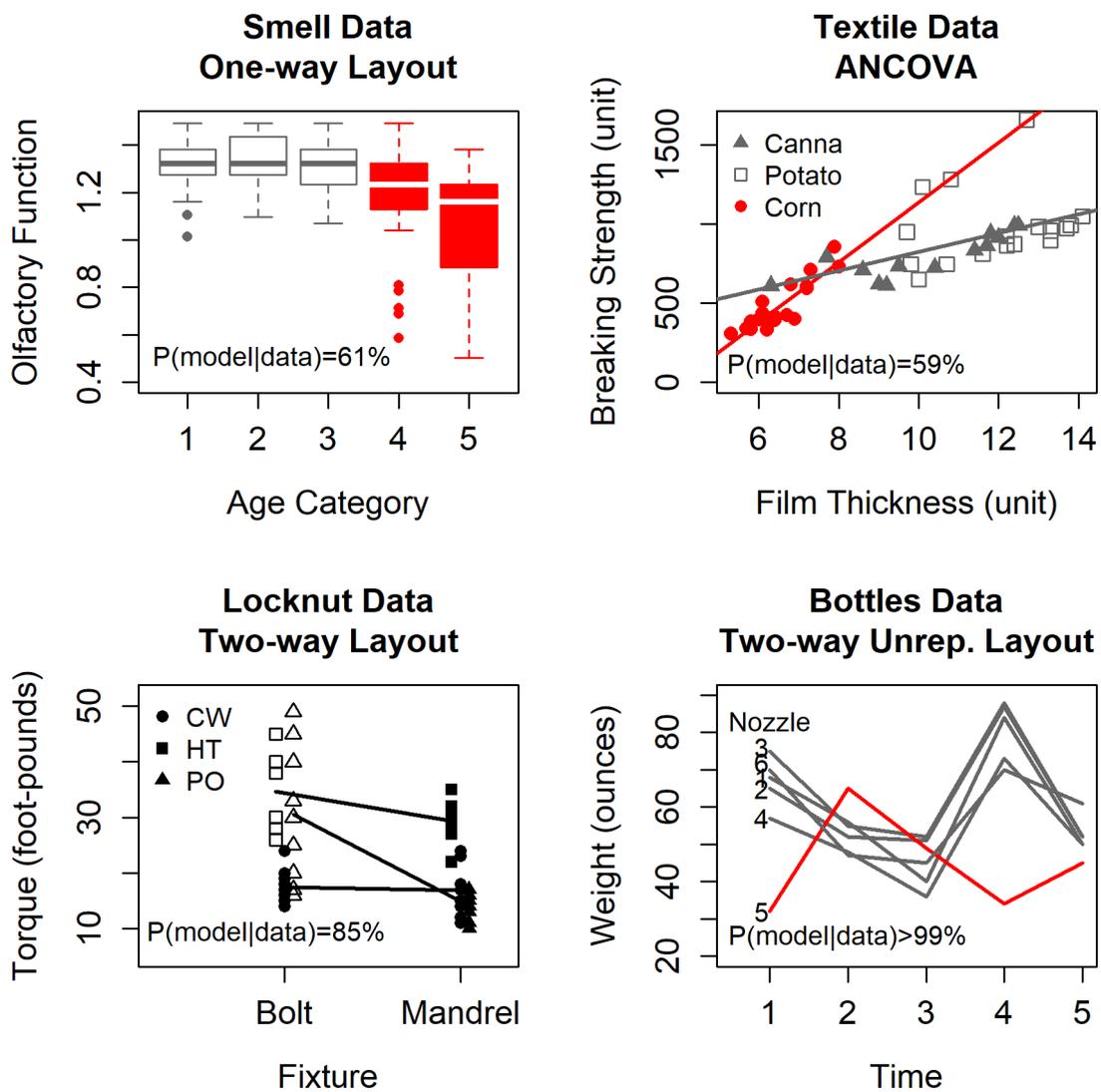
Figure 1: Smell data (O'Brien and Heft, 1995, top left), textile data (Furry, 1939, top right), locknut data (Meek and Ozgur, 1991, bottom left), and bottles data (Ott and Snee, 1973, bottom right). Color (red/gray) shows latent grouping structure (i.e., group-based regression effects) for smell, textile, and bottles data, and fill (solid/open geometry) shows group-based variances for smell, textile, and locknut data.

(CW), heat treating (HT), and phosphate and oil (PO). The threading techniques studied include bolt and mandrel, the types of fixture on which each locknut was affixed to conduct the test. We find the highest posterior model probability (85%) for the model where bolt by HT and bolt by PO measurements have a larger error variance than those from bolt by CW, mandrel by HT, mandrel by PO, and mandrel by CW. We call this the locknut data.

Finally, in the bottom right panel, the data set of Ott and Snee (1973) represents an unreplicated two-way layout where six machine nozzles were used to fill bottles on five occasions (horizontal axis). The weight of each bottle (vertical axis) was measured, and we find the highest posterior model probability for the structure where nozzle 5 is found to be out of alignment from the others ($> 99\%$). We call this the bottles data.

The slgf package implements a combinatoric approach that evaluates all possible assignments of SLGF levels into two groups. We refer to each these assignments as *schemes*. For example, in the smell data, the scheme that is visualized assigns age levels 1, 2, and 3 into one group and levels 4 and 5 into the other, denoted {1,2,3}{4,5}. More details on how schemes are established can be found in Subsection Grouping schemes and model classes.

The user may specify an SLGF for regression effects, another SLGF for error variances, require them to be the same, or specify no SLGF for one or both of these. For example, the smell data has age as the SLGF for both. In Subsection **Case study 2: textile data**, we analyze a data set with distinct regression and error variance SLGFs.

In this paper, we provide an overview of the slgf package that enables analysis of data sets like those in Figure 1 via Bayesian model selection. In Section **SLGF methodology**, we briefly review the SLGF methodology. In Section **Using the slgf package**, we illustrate the package functionality for the four data sets illustrated in Figure 1. For each data set, we will demonstrate the relevant code and package functionality along with a comparison between the results of a classical approach and our approach. In Section **Conclusion**, we summarize the package and its functionality.

2 SLGF methodology

2.1 Model specification

For a thorough review of the SLGF model specification see Metzger and Franck (2021). First consider the linear model

$$Y = 1^T \alpha + W\nu + V\tau + U\rho + \varepsilon, \tag{1}$$

where 1^T is an $N \times 1$ vector of 1s, α is an intercept, ν represents the full SLGF effect with K degrees of freedom, τ represents the regression effects that do not arise from latent groupings (i.e., all other regression effects of interest), and the ρ terms indicate statistical interactions between SLGF and other regression effects; W, V, and U partition the overall model matrix into model matrices corresponding to the SLGF effects ρ, additional effects τ, and SLGF interactions, respectively; and finally ε represents an $N \times 1$ vector of errors where $\varepsilon \stackrel{\text{iid}}{\sim} N(0, \Sigma)$ for $\Sigma = \sigma^2 I$ where I is an $N \times N$ identity matrix.

Because a central goal of the SLGF methodology is to compare models with and without latent grouping structures, we next develop notation to indicate whether model terms in Equation (1) involve groupings of factor levels or not. If a model contains a one degree of freedom group effect instead of the full K degree of freedom SLGF effect, we denote the effect $\tilde{\nu}$ instead, with corresponding $\tilde{W}$ to ensure they remain conformable. Similarly, if the interaction ρ is with the group effect rather than the full SLGF effect, we denote it $\tilde{\rho}$. When there are group-based error variances, we let $\tilde{\varepsilon}$ denote the vector of heteroscedastic errors, where the elements of $\tilde{\varepsilon}$ are either $N(0, \sigma_1^2)$ or $N(0, \sigma_2^2)$ depending on their membership in group 1 or 2, respectively.

For example, for the smell data in the top left panel of Figure 1, the most probable model can be represented as $Y = 1^T \alpha + \tilde{W}\tilde{\nu} + \tilde{\varepsilon}$, with a 1 degree of freedom group effect $\tilde{\nu}$ (color-coding) and heteroscedastic error term $\tilde{\varepsilon}$ (shading). This model (posterior model probability 0.65) was found to be far more probable than the ordinary one way analysis of variance model $Y = 1^T \alpha + W\nu + \varepsilon$ (posterior model probability less than 0.0001), the model with a 4 degree of freedom mean effect ν and homoscedastic errors ε. Similarly, the bottles data (bottom right panel) most probable model is $Y = 1^T \alpha + W\nu + \tilde{U}\tilde{\rho} + \varepsilon$ with a 4 degree of freedom nozzle effect ν, an 8 degree of freedom group-by-nozzle interaction $\tilde{\rho}$, and homoscedastic errors ε.

2.2 Grouping schemes and model classes

Recall schemes are the possible assignments of factor levels to two latent groups. While the schemes shown in Figure 1 may seem visually obvious, the **slgf** package considers all possible such assignments of factor levels into two groups. This (i) obviates the need for the user to specify specific schemes, and (ii) apportions prior model probabilities commensurately with the actual number of models corresponding to a SLGF to prevent detection of spurious latent grouping structure. Problems will differ in the number of schemes under consideration. The package **slgf** automatically determines the schemes once the set of candidate models has been established by the user. The minimum number of levels that can comprise a grouping scheme can be adjusted by the user to lower the number of candidate models or to avoid creating model effects with too few degrees of freedom to be estimated. The user may specify the SLGF for regression effects and/or error variances, or neither. These SLGFs may or may not be different factors. If they are the same, the user may require that the grouping schemes must be equal or that they may be distinct. For example, in the textile data in the top right panel of Figure 1, the SLGF is starch for both regression effects and error variances, but the user should allow for distinct schemes since the variance scheme appears to be {potato}{canna,corn} and the regression effect scheme appears to be {corn}{canna,potato}.

A model *class* describes the structure of the model including specification of effects related to the hidden groups. Model classes may include, for example, the set of models with group-based regression effects but no group-based variances; or, a single model with no group-based regression effects or variances. For example, in the smell data represented in top left panel of Figure 1, we consider the following 62 models comprising six model classes:

1. A single model with a 1 degree of freedom global mean effect and homoscedastic error variance;

2. A single model with a 4 degree of freedom mean effect and homoscedastic error variance;

3. 15 models (corresponding to the 15 possible grouping schemes) with a 1 degree of freedom global mean effect and group-based heteroscedastic error variances;

4. 15 models with a 4 degree of freedom mean effect and group-based heteroscedastic error variances;

5. 15 models with a 1 degree of freedom group-based mean effect and homoscedastic error variance;

6. 15 models with a 1 degree of freedom group-based mean effect and group-based error variances.

For our analysis, we specified that the regression effect and variance grouping schemes must be equivalent, and that one level of the age factor could comprise a group. The user can relax these specifications as desired.

2.3 Parameter priors

With **slgf**, the user can choose to implement noninformative priors on the regression effects (default), or the Zellner-Siow mixture of g-priors on these effects. We first enumerate the noninformative priors. Let β represent the full set of regression effects. For simplicity, we parametrize on the precision scale where $\varphi = \frac{1}{\sigma^2}$ and the corresponding precision matrix $\varphi I_{n \times n}$ is denoted Φ. For a model m_s^c where c indexes class and s indexes grouping scheme, **slgf** imposes

$$P(\boldsymbol{\beta}, \varphi | m_s^c) \propto \varphi \tag{2}$$

for homoscedastic models, and

$$P(\boldsymbol{\beta}, \varphi_1, \varphi_2 | m_s^c) \propto \varphi_1 \cdot \varphi_2 \tag{3}$$

for heteroscedastic models.

Alternatively, in contexts with limited data, such as the two-way unreplicated bottles data in the bottom right panel of Figure 1, we recommend employing the Zellner-Siow mixture of g-prior (Zellner and Siow, 1980; Zellner, 1986; Liang et al., 2008), which reduces the minimal training sample size necessary for the computation of the fractional Bayes factor (see Subsection **Fractional Bayes factors and posterior model probabilities** for further detail). We have generally found that in cases where the number of data points is close to the number of parameters in some of the larger candidate models (e.g., case study 4, bottles data), the mixture of g-priors approach outperforms the noninformative priors due to the drastic reduction in the required proportion of the data needed to implement the fractional Bayes factor approach. For homoscedastic models, recall $\Phi = \phi I$ where I is an $N \times N$ identity matrix. Let

$$P(\alpha, \varphi | m_s^c) \propto \varphi \tag{4}$$

and

$$\boldsymbol{\beta}_{-\alpha}|\Phi, g, m_s^c \sim N(\mathbf{0}, g(X^T\Phi^{-1}X)^{-1}). \tag{5}$$

Next, for heteroscedastic models, first denote $\tilde{\Phi}$ as a diagonal precision matrix where the ith diagonal element is either φ_1 or φ_2, depending upon the grouping membership of the ith observation. Let

$$P(\alpha, \varphi_1, \varphi_2|m_s^c) \propto \varphi_1 \cdot \varphi_2 \tag{6}$$

and

$$\boldsymbol{\beta}_{-\alpha}|\tilde{\Phi}, g, m_s^c \sim N(\mathbf{0}, g(X^T\tilde{\Phi}^{-1}X)^{-1}); \tag{7}$$

In both homoscedastic and heteroscedastic cases,

$$g \sim \text{InvGamma}\Big(\frac{1}{2}, \frac{N}{2}\Big). \tag{8}$$

Thus for homoscedastic models, the full prior on all parameters is the product of Equations (4), (5), and (8). For heteroscedastic models, it is the product of Equations (6), (7), and (8).

2.4 Fractional Bayes factors and posterior model probabilities

Note that if we form a standard Bayes factor for models using improper priors on parameters, the unspecified proportionality constants associated with the improper priors (Equations 2, 3, 4, and 6) would not cancel one another and the Bayes factor would be defined only up to an unspecified constant. Thus we invoke a fractional Bayes factor approach (O'Hagan, 1995) to compute well-defined posterior model probabilities for each model. More details follow.

The **slgf** package obtains posterior model probabilities through the use of fractional Bayes factors. Briefly, a Bayes factor is defined as the ratio of two models' integrated likelihoods. The integrated likelihood is obtained by integrating parameters out of the joint distribution of data and parameters. In some cases, this integration is analytic, but in others, it is undertaken with a Laplace approximation; the corresponding simplified expressions and methods used to optimize them are described in detail later in this section. In the SLGF context, let $\mathcal{M}$ represent the full set of models under consideration, representing all classes and grouping schemes of interest. Denote $\boldsymbol{\theta}$ as the full set of unknown parameters associated with a model $m_s^c \in \mathcal{M}$ and $\pi(\boldsymbol{\theta}|m_s^c)$ as the prior on these parameters given model m_s^c. The parameter vector $\boldsymbol{\theta}$ depends on class and scheme of model m_s^c. The integrated likelihood is

$$P(\boldsymbol{Y}|m_s^c) = \int_{\boldsymbol{\Theta}} P(\boldsymbol{Y}|\boldsymbol{\theta}, m_s^c)\pi(\boldsymbol{\theta}|m_s^c)d\boldsymbol{\theta},$$

with Bayes factor comparing models m_s^c and $m_{s'}^{c'}$

$$BF = \frac{P(\boldsymbol{Y}|m_s^c)}{P(\boldsymbol{Y}|m_{s'}^{c'})}.$$

Since the priors used by the **slgf** package are improper, $\pi(\boldsymbol{\theta}|m_s^c)$ is defined only up to an unspecified constant. Thus, BF is defined only up to a ratio of unspecified constants. To overcome this issue and enable improper priors on parameters to be used in the course of Bayesian model selection, the fractional Bayes factor (O'Hagan, 1995) was developed. A fractional Bayes factor is a ratio of two fractional marginal model likelihoods, where a fractional marginal likelihood is defined as

$$q^b(\boldsymbol{Y}|m_s^c) = \frac{\int P(\boldsymbol{Y}|\boldsymbol{\theta}, m_s^c)\pi(\boldsymbol{\theta}|m_s^c)d\boldsymbol{\theta}}{\int P(\boldsymbol{Y}|\boldsymbol{\theta}, m_s^c)^b\pi(\boldsymbol{\theta}|m_s^c)d\boldsymbol{\theta}}. \tag{9}$$

The $q^b(\boldsymbol{Y}|m_s^c)$ quantity in Equation (9) is the integrated likelihood based on the $1-b$ fraction of the data where the improper prior has been updated to become proper with b fraction of the data. Thus all normalizing constants are specified. The fractional Bayes factor is thus

$$FBF = \frac{q^b(\boldsymbol{Y}|m_s^c)}{q^b(\boldsymbol{Y}|m_{s'}^{c'})}.$$

for some fractional exponent $0 < b < 1$. Thus we must compute the integrals $\int P(\boldsymbol{Y}|\boldsymbol{\theta}, m_s^c)\pi(\boldsymbol{\theta}|m_s^c)d\boldsymbol{\theta}$ and $\int P(\boldsymbol{Y}|\boldsymbol{\theta}, m_s^c)^b\pi(\boldsymbol{\theta}|m_s^c)d\boldsymbol{\theta}$, the numerator and denominator of Equation (9), respectively, for all $m_s^c \in \mathcal{M}$. Although O'Hagan (1995) provides several recommendations for choice of b, **slgf** exclusively

implements $b = \frac{m_0}{N}$ where m_0 is the minimal training sample size required for the denominator of Equation (9) to be proper for all models. If m_0 is too small, then the denominator of Equation (9) diverges. The user must specify m_0; if their choice is too low, then **slgf** increases it until all relevant integrals converge. For further details, see O'Hagan (1995), p. 101; for recommendations on choosing m_0 in practice, see Subsection **Choice of m$_0$**.

Next we discuss the technical details on how these integrals are computed via Laplace approximation. Specifically, we will describe how $\log\left(q^b(\boldsymbol{Y}|m_s^c)\right)$ is computed in each case. In the case of noninformative regression priors for homoscedastic models, $\boldsymbol{\beta}$ and σ^2 are integrated analytically. Let $\hat{\boldsymbol{Y}}$ represent the fitted values of m_s^c and SSResid the residual sum of squares of this model. We obtain

$$
\log\left(q^b(\boldsymbol{Y}|m_s^c)\right) = \left(-\frac{N(1-b)}{2}\right)(\log \pi + \log(\text{SSResid})) + \left(\frac{Nb-1}{2}\right)\log b + \log\left(\frac{\Gamma\left(\frac{N-P}{2}\right)}{\Gamma\left(\frac{Nb-P}{2}\right)}\right)
$$
(10)

In the case of noninformative regression priors for heteroscedastic models, both the numerator and denominator integrals of Equation (9) must be approximated with a Laplace approximation because although $\boldsymbol{\beta}$ can be integrated analytically, σ_1^2 and σ_2^2 cannot be. The integrals are computed on the log-scale for numeric stability. Equation (9) on the log-scale simplifies to:

$$
\log\left(q^b(\boldsymbol{Y}|m_s^c)\right) = \frac{N(b-1)}{2}\log(2\pi) + \frac{P+1}{2}\log b + \frac{1}{2}\log\left(\frac{|H_b^\star|}{|H^\star|}\right) + \log\left(\frac{P(\boldsymbol{Y}|\boldsymbol{\theta}^\star)\pi(\boldsymbol{\theta}^\star|m_s^c)}{P(\boldsymbol{Y}|\boldsymbol{\theta}_b^\star)^b\pi(\boldsymbol{\theta}_b^\star|m_s^c)}\right)
$$
(11)

where $\boldsymbol{\theta}^\star$ and $H^\star$ denote the mode and Hessian of $P(\boldsymbol{Y}|\boldsymbol{\theta}, m_s^c)\pi(\boldsymbol{\theta}|m_s^c)$, and $\boldsymbol{\theta}_b^\star$ and $H_b^\star$ denote the mode and Hessian of $P(\boldsymbol{Y}|\boldsymbol{\theta}, m_s^c)^b\pi(\boldsymbol{\theta}|m_s^c)$. These modes and Hessians are computed with **optim** using the Nelder-Mead algorithm.

In the Zellner-Siow mixture of g-prior case, α and $\boldsymbol{\beta}_{-\alpha}$ are integrated analytically. For homoscedastic models, σ^2 is as well, and only g is integrated with a Laplace approximation. Again marginal model likelihoods are computed on the log-scale. The log of the mode of $P(\boldsymbol{Y}|g, m_s^c)^b\pi(\boldsymbol{\theta}|m_s^c)$, denoted $g_b^\star$, is found by solving the closed-form equation $\frac{(Nb-1-P)}{2}\log(1+bg) + \frac{Nb-1}{2}\log(1+bg(1-R^2)) - \frac{3}{2}\log g - \frac{N}{2g} := 0$ with the base **R** function **uniroot** where R^2 is the coefficient of determination for m_s^c. The Hessian is then evaluated at this solution $g_b^\star$; the closed-form Hessian of $P(\boldsymbol{Y}|g, m_s^c)^b\pi(\boldsymbol{\theta}|m_s^c)$ evaluated at $g^\star$ is given by $H_b^\star = \frac{1}{2}\left(\frac{((Nb-1)b^2(1-R^2)^2}{(1+bg^\star)(1-R^2)^2} - \frac{(Nb-P-1)b^2}{(1+bg^\star)^2} + \frac{3}{g^{\star 2}} - \frac{2N}{g^{\star 3}}\right)$. For $b=1$, this expression describes the numerator of Equation (9); see Liang et al. (2008) for further mathematical details. The Laplace approximation for Equation (9) on the log-scale then is given by:

$$
\log\left(q^b(\boldsymbol{Y}|m_s^c)\right) = \log\left(\frac{\Gamma\left(\frac{N-1}{2}\right)}{\Gamma\left(\frac{Nb-1}{2}\right)}\right) + \frac{Nb-1}{2}(\log(\text{SSTotal}) + \log \pi) + \frac{1}{2}\log\left(\frac{|H_b^\star|}{|H^\star|}\right) +
$$
$$
\log\left(\frac{P(\boldsymbol{Y}|\boldsymbol{\theta}^\star)\pi(\boldsymbol{\theta}^\star|m_s^c)}{P(\boldsymbol{Y}|\boldsymbol{\theta}_b^\star)^b\pi(\boldsymbol{\theta}_b^\star|m_s^c)}\right).
$$
(12)

For heteroscedastic models, a three-dimensional Laplace approximation is used to integrate σ_1^2, σ_2^2, and g. To obtain $\boldsymbol{\theta}_b^\star$ and $\boldsymbol{\theta}^\star$, we first transform $\gamma_1 = \log\left(\frac{1}{\sigma_1^2}\right)$ and $\gamma_2 = \log\left(\frac{1}{\sigma_2^2}\right)$ to stabilize the optimization. We optimize $\log P(\boldsymbol{Y}|g, \sigma_1^2, \sigma_2^2)^b\pi(\sigma_1^2, \sigma_2^2, g) = \frac{n_1 b}{2}\log \gamma_1 + \frac{n_2 b}{2}\gamma_2 - \frac{P}{2}\log g + \frac{1}{2}|X^T\tilde{\Sigma}X| - \frac{1}{2}\log|\frac{bg+1}{bg}X^T(\tilde{\Sigma} - Z_{\tilde{\Sigma}})X| - \frac{b}{2}\boldsymbol{Y}^T\left(\tilde{\Sigma} - Z_{\tilde{\Sigma}} - (\tilde{\Sigma} - Z_{\tilde{\Sigma}})X\left(\frac{bg+1}{bg}X^T\tilde{\Sigma}X - X^TZ_{\tilde{\Sigma}}X\right)^{-1}X^T(\tilde{\Sigma} - Z_{\tilde{\Sigma}})\right)\boldsymbol{Y} - \frac{3}{2}\log(g) - \frac{N}{2g} + \log(J)$ using the Nelder-Mead method from **optim** where $Z_{\tilde{\Sigma}} = \tilde{\Sigma}Z(Z^T\tilde{\Sigma}Z)^{-1}Z^T\tilde{\Sigma}$, $Z = \boldsymbol{1}^T$, and $\log(J) = -(\gamma_1 + \gamma_2)$ represents the determinant of the log-precision transformation. For $b=1$ these equations yield integrand of the numerator of (9).

With the modes computed, the Hessians of $\log P(\boldsymbol{Y}|g, \sigma_1^2, \sigma_2^2)^b\pi(\sigma_1^2, \sigma_2^2, g)$ are calculated with the function **Hessian** from the package **numDeriv**. Finally with the modes and Hessians computed, the Laplace approximation for Equation (9) is given by:

$$\log\left(q^b(\mathbf{Y}|m_s^c)\right) = \frac{Nb-1}{2}\log(2\pi) + \frac{P+1}{2}\log(b) + \frac{1}{2}\log\left(\frac{|H_b^\star|}{|H^\star|}\right) + \log\left(\frac{P(\mathbf{Y}|\boldsymbol{\theta}^\star)\pi(\boldsymbol{\theta}^\star|m_s^c)}{P(\mathbf{Y}|\boldsymbol{\theta}_b^\star)^b\pi(\boldsymbol{\theta}_b^\star|m_s^c)}\right).$$
(13)

For the sake of consistency, all models, even with fully tractable marginal model likelihoods, are computed with a FBF. Once log-fractional marginal likelihoods have been computed for all models, we subtract the maximum from this set so that the set of log-fractional marginal likelihoods has been rescaled to have a maximum of 0. Each value is exponentiated to obtain a set of fractional marginal likelihoods with maximum 1. This adjustment helps to avoid numerical underflow when computing posterior model probabilities.

2.5 Choice of m_0

The user must specify the argument m_0, the minimal training sample size such that all marginal model likelihoods are well-defined. If `prior="flat"`, then we recommend that the user begins by letting `m0` equal the dimension of the improper priors: that is, the number of coefficients in most complex model under consideration plus the number of variances under consideration. If `prior="zs"`, then `m0` can generally be much smaller (in practice, we have found that `m0=2` performs well) as the prior on the regression effects is proper. If the user's choice is too low, then `ms_slgf` will incrementally increase it by 1 until all marginal model probabilities are numerically stable. If `m0` reaches n, corresponding to 100% of data used for training, `ms_slgf` will terminate and the user should specify a different set of models.

2.6 Model priors

With this adjusted set of fractional marginal likelihoods, we next consider the priors for the model space. The function `ms_slgf` imposes a uniform prior by model class, and for classes containing multiple models, the prior on each class is uniformly divided among the models it contains. We finally compute posterior model probabilities for each model:

$$P(m'|\mathbf{Y}) = \frac{P(\mathbf{Y}|m')P(m')}{\sum_{\mathcal{M}}P(\mathbf{Y}|m)P(m)}.$$
(14)

The prior probability placed on each model can be found in the `models$ModPrior` vector in output from `ms_slgf`.

2.7 Parameter estimation

Our approach provides maximum *a posteriori* (MAP) estimates for all relevant quantities: $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2 = \{\hat{\sigma}^2\}$ or $\hat{\sigma}^2 = \{\hat{\sigma}_1^2, \hat{\sigma}_2^2\}$ in the homoscedasitc and heteroscedastic cases respectively, and g in the Zellner-Siow mixture of g-prior case.

Because the prior on $\boldsymbol{\beta}$ is either flat or centered at 0, the MAP estimator is simply the usual maximum likelihood estimator:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} P(\mathbf{Y}|X, \boldsymbol{\beta}, \Sigma)$$
(15)

so that $\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{Y}$. The variance(s) and g were computed via the base **R** function **optim** during the Laplace approximation stage. For computational efficiency, $\boldsymbol{\beta}$ is integrated out of $P(\mathbf{Y}|X, \boldsymbol{\theta})P(\boldsymbol{\theta})$ and the variances are estimated on the log-scale, so we let $\hat{\boldsymbol{\lambda}} := \{\hat{\lambda}\}$ in homoscedastic models or $\{\hat{\lambda}_1, \hat{\lambda}_2\}$ in heteroscedastic models. Then

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} \int P(\mathbf{Y}|X, \boldsymbol{\beta}, \Sigma)P(\boldsymbol{\beta})P(\Sigma)d\boldsymbol{\beta}$$
(16)

or,

$$\{\hat{\boldsymbol{\lambda}}, \hat{g}\} = \arg\max_{\boldsymbol{\lambda}, g} \int P(\mathbf{Y}|X, \boldsymbol{\beta}, \Sigma, g)P(\alpha)P(\boldsymbol{\beta}_{-\alpha}|\Sigma)P(g)d\boldsymbol{\beta}.$$
(17)

Then, $\hat{\sigma}^2 = \exp\{\hat{\lambda}\}$ for $\hat{\sigma}^2 = \{\hat{\sigma}^2\}$ or $\hat{\sigma}^2 = \{\hat{\sigma}_1^2, \hat{\sigma}_2^2\}$. The output values `coefficients`, `variances`, and `gs` (only if `prior="zs"`) are lists where each element contains the estimates for each model's $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2$, and $\hat{g}$, respectively.

3 Using the slgf package

The function `ms_slgf()` is the main function of **slgf** that implements the methodology we have described. Each argument of `ms_slgf()` and its output will be illustrated in the case studies found in Subsections **Case study 1: smell data**, **Case study 2: textile data**, **Case study 3: locknut data**, and **Case study 4: bottles data**. The `ms_slgf()` function requires several inputs to compute and output posterior model probabilities for all models, schemes, and model classes of interest. The user begins with a `data.frame` containing a continuous response, at least one categorical predictor, and any other covariates of interest. The `data.frame` cannot contain column names with the character string `group`, because `ms_slgf()` will search for this string when fitting group-based models. The user must first identify an SLGF for the fixed effects and/or the variance. The user indicates, via the arguments `response`, `slgf_beta`, and `slgf_Sigma`, character strings corresponding to the response, the suspected latent fixed effect grouping factor, and the suspected latent variance grouping factors, respectively. If no latent regression effect structure or variance structure is to be considered, the user may specify `slgf_beta=NA`, `slgf_Sigma=NA`, or both. We note that if the user does not specify any SLGFs, the model selection is still undertaken through fractional Bayes factors as described previously. If the user chooses the same categorical variable for both latent grouping factors, the argument `same_scheme`, which defaults to `FALSE`, can indicate whether the grouping schemes for the regression effect and variance structures must be equivalent.

Next the user determines the model classes they wish to evaluate. The argument `usermodels` is a `list` where each element contains a string of R class `formula` or `character`. The user also specifies which classes should also be considered in a heteroscedastic context via the argument `het`, which is a vector of the same length as `usermodels`, containing an indicator 1 or 0 corresponding to each model class specified in `usermodels` where 1 indicates the model will be considered with group-based variances and 0 indicates it will not. Together the arguments `usermodels` and `het` indicate which fixed effect structures are of interest, and which should be further considered for heteroscedasticity, thus implicitly creating the full set of model classes considered.

Next the user chooses a prior to place on the regression effects. As described in Subsection **Parameter priors**, `prior="flat"` (the default) implements the noninformative prior and `prior="zs"` imposes the Zellner-Siow mixture of g-prior.

Finally the user must specify the minimum number of levels of the SLGF that can comprise a group, via the arguments `min_levels_beta` and `min_levels_Sigma`, which default to 1. The number of possible grouping schemes increases with the number of levels of the SLGF. To speed up the computation, the user can increase these arguments and thus reduce the number of candidate models. Because we partition into two groups, note these arguments may not exceed half the number of levels of the SLGF. Additionally, when considering data with limited degrees of freedom, increasing `min_levels_beta` and/or `min_levels_Sigma` may be necessary to ensure effects can be computed.

3.1 Case Study 1: smell data

First we revisit the smell data set analyzed by O'Brien and Heft (1995). They measured olfactory acuity (denoted `olf`) on a continuous scale as a function of age (`agecat`), where age groups were divided into five categorical levels. See Figure 2. We note that levels 4 and 5 of `agecat` appear to have larger variance than levels 1, 2, and 3, but standard analysis of variance models assume homoscedasticity. We first demonstrate how a classical analysis might misrepresent the data. A usual one-way ANOVA analysis compares the null model, with a single mean, against the alternative model, with 4 degrees of freedom for the mean effects, with homoscedastic error variance.

```
 % remove smell null model
> smell$agecat <- as.factor(smell$agecat) # coerce agecat to a factor variable
> smell_null <- lm(olf~1, data=smell) # fit a null model with a single mean
> smell_full <- lm(olf~agecat, data=smell) # fit a full model with a 4 agecat effects
> print(smell_null)
Call:
lm(formula = olf ~ 1, data = smell)

Coefficients:
(Intercept)
      1.234
> print(smell_full)
Call:
lm(formula = olf ~ agecat, data = smell)
```
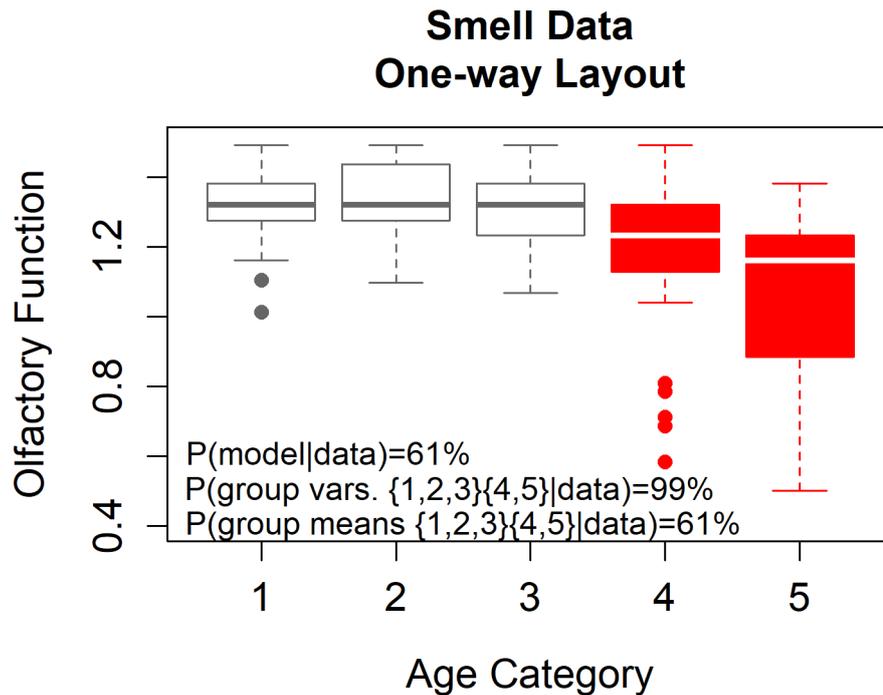
Figure 2: The smell data (O'Brien and Heft, 1995) is analyzed for group-based means and variances. We find posterior model probability of 61% for the model with group-based means and variances with scheme {1,2,3}{4,5}. We also find overall posterior probability of grouping scheme

.

```
Coefficients:
(Intercept)      agecat2      agecat3      agecat4      agecat5
   1.31689      0.02824     -0.01075     -0.11580     -0.25728
> anova(smell_null, smell_full) # compare the null and full models
Analysis of Variance Table

Model 1: olf ~ 1
Model 2: olf ~ agecat
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    179 7.7585
2    175 5.6197  4    2.1388 16.651 1.395e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(smell_null)$sigma^2
0.04334349
> summary(smell_full)$sigma^2
0.03211259
```

This approach, which assumes all levels of agecat have equal error variance, favors the model with a 4 degree of freedom agecat effect. Note we obtain maximum likelihood estimates for the error variance of $\hat{\sigma}^2_{\text{full}} = 0.03211$. Based on Figure 2, we suspect this value may overestimate the error variance for levels 1, 2, and 3, while underestimating that of levels 4 and 5. We also suspect that the full model may be overly complex, as the means for levels 1, 2, and 3 appear to be plausibly equivalent. That is, the apparent latent grouping scheme for both regression effects and error variances is {1,2,3}{4,5}, or equivalently, {4,5}{1,2,3}.

Next, consider the **slgf** approach. We will consider the classes of models with group-based means, group-based variances, and both group-based means and variances. We specify dataf=smell and response="olf", along with slgf_beta="agecat" and slgf_Sigma="agecat" as the suspected latent grouping factor for both regression effects and variances. We set the minimum number of levels for a group to 1 with min_levels_beta=1 and min_levels_Sigma=1. Note that fewer grouping

schemes would be considered if we let these arguments equal 2. For simplicity, since the mean and variance grouping schemes both visually appear to be {1,2,3}{4,5}, we will restrict the schemes to be equivalent with `same_scheme=TRUE`. Via the `usermodels` argument, we will consider the null model class olf∼1, the full model class olf∼agecat, and the group-means model class olf∼group, which will automatically consider all possible grouping schemes. Similarly, we will consider each of these formulations with the class of both homoscedastic and group-based variances via the argument `het=c(1,1,1)`. With a relatively large amount of data, we will use the uninformative `prior="flat"`. Finally we specify a minimal training sample size of `m0=9`, although if we specify this value to be too small, `ms_slgf()` will automatically increase it to the smallest value for which the relevant integrals converge and/or the necessary optimizations can be performed. We run `ms_slgf` to obtain the posterior model probabilties for all 62 models under consideration. We inspect the two most probable models, with indices 62 and 32, which comprise over 99% of the posterior probability over the model space considered:

```
> smell_out <- ms_slgf(dataf=smell, response="olf", lgf_beta="agecat",
                       min_levels_beta=1, lgf_Sigma="agecat",
                       min_levels_Sigma=1, same_scheme=TRUE,
                       usermodels=list("olf~1", "olf~agecat", "olf~group"),
                       het=c(1,1,1), prior="flat", m0=9)
> smell_out$models[c(1,2),c(1,2,3,5,7)]
        Model  Scheme.beta Scheme.Sigma  FModProb Cumulative
62  olf~group {4,5}{1,2,3} {4,5}{1,2,3} 0.6054935  0.6054935
32 olf~agecat         None {4,5}{1,2,3} 0.3878754  0.9933688
```

The most probable model, as suspected, is olf∼group, indicating group-based means where `Scheme.beta` is {4,5}{1,2,3}. Note also `Scheme.Sigma` indicates group-based heteroscedasticity with the same scheme. This model received posterior probability of approximately 61%. The next most probable model also has group-based heteroscedasticity with scheme {4,5}{1,2,3}, but note the model is olf∼agecat, containing the full model not with group-based mean effects, but rather 4 degrees of freedom for the agecat effect. By inspecting `smell_out$scheme_probabilities_Sigma`, we see that models with variance grouping scheme {4,5}{1,2,3} comprise over 99% of the posterior probability. By contrast, the models with fixed effect grouping scheme {4,5}{1,2,3} (that is, the homoscedastic and heteroscedastic versions) comprise 61% of the posterior probability. We find these posterior probabilities intuitive, easy to interpret quantifications of uncertainty.

The output fields `coefficients` and `variances` contain lists with the coefficients and variance(s) associated with each model. The output field `model_fits` contains the output from a linear model fit to the model specification in question, containing the , and Note the most probable model has index 62, so we inspect the 62nd elements of the coefficient and variance lists `smell_out$coefficients` and `smell_out$variances`, which contain the MAP estimates for each model's regression effects and variance(s), respectively. The group-based variance estimates are $\hat{\sigma}^2_{\{4,5\}} = 0.0587$ and $\hat{\sigma}^2_{\{1,2,3\}} = 0.0121$. We contrast these variances against the estimate $\hat{\sigma}^2_{\text{full}} = 0.032$, which appears to have overestimated the variance of levels 1, 2, and 3, while simultaneously underestimating that of levels 4 and 5.

```
> smell_out$coefficients[[62]]
(Intercept)  group{4,5}
  1.3252211  -0.1940328
> smell_out$variances[[62]]
     {4,5}     {1,2,3}
0.05868885 0.01211084
```

3.2 Case study 2: textile data

We reanalyze the breaking strength data set of Furry (1939), also investigated by Metzger and Franck (2021), to illustrate the additional flexibility of **slgf** beyond the original work. The breaking strength of a starch film strength (measured in grams) is analyzed according to the thickness of the film, denoted film (measured in 10^{-4} inches), and the type of starch starch used to create the film (canna, corn, or potato). As usual, we begin by plotting the data to ascertain whether there is a latent grouping factor present. By inspection we note that the potato films, represented by squares in Figure 3, appear to have a higher variability than the corn (filled red circles) and canna (filled gray triangles) films.

We first illustrate a typical ANCOVA approach, in which three parallel lines for each level of starch are fit with a common error variance. This model leads to the fit shown in the center panel of Figure 3. Note only the film thickness effect is statistically significant according to a traditional hypothesis testing approach with $\alpha = 0.05$. The residual standard error of this model is $\hat{\sigma}^2_{\text{ANCOVA}} = 27126.09$.
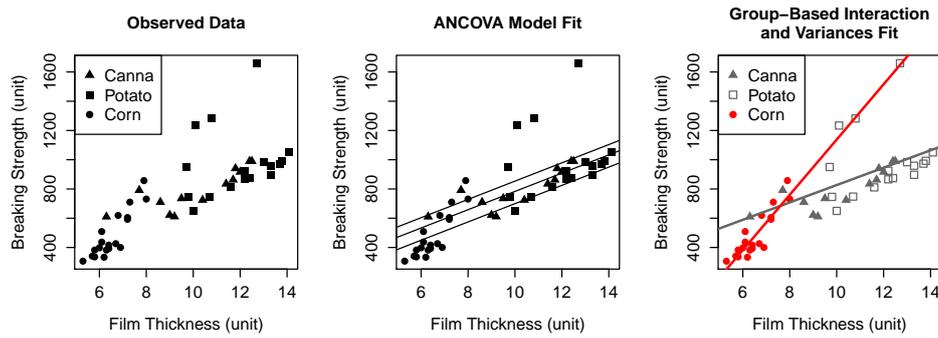
Figure 3: The breaking strength data set from Furry (1939) represents the breaking strength of starch films depending on the thickness of a film coating and the type of starch used to make the film. The left panel shows the data. The center panel shows an ANCOVA model. The right panel shows the most probable model ($P(m|Y) \approx 66\%$) containing a latent group-based interaction between groups {canna, potato}{corn} (gray points vs. red points) and film thickness, as well as distinct variances between groups {canna, corn}{potato} (filled points vs. open points).

```
> textile_ancova <- lm(strength~film+starch, data=textile)
> summary(textile_ancova)

Call:
lm(formula = strength ~ film + starch, data = textile)

Residuals:
    Min     1Q  Median     3Q    Max
-203.63  -99.45  -57.84  56.72  637.61

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   158.26     179.78   0.880 0.383360
film           62.50      17.06   3.664 0.000653 ***
starchcorn    -83.67      86.10  -0.972 0.336351
starchpotato   70.36      67.78   1.038 0.304795
```

We contrast these findings against our methodology with **slgf**. The following arguments are input: `dataf=textile` specifies the data frame; `response="strength"` specifies the column of `textile` that contains the response variable; `slgf_beta="starch"` and `slgf_Sigma="starch"` indicate that the categorical variable `starch` should be used as the latent grouping factor for both regression effects and variances; `same_scheme=FALSE` indicates that the latent regression effect and variance grouping structures do not need to be partitioned by the same levels of `starch`; `min_levels_beta=1` and `min_levels_Sigma=1` indicate that a latent group can contain only one level of `starch`; the usermodels argument indicates that we will consider main effects models `strength~film+starch` and `strength~film+starch+film*starch`, and models with group-based regression effects including `strength~film+group` and `strength~film+group+film*group`; the argument `het=c(1,1,1,1)` indicates that each of these four model specifications will also be considered with group-based variances; `prior="flat"` places a flat prior on the regression effects; and `m0=8` specifies the minimal training sample size.

```
> data(textile)
> out_textile <- ms_slgf(dataf = textile, response = "strength",
               lgf_beta = "starch", lgf_Sigma = "starch",
               same_scheme=FALSE, min_levels_beta=1, min_levels_Sigma=1,
               usermodels = list("strength~film+starch", "strength~film*starch",
                              "strength~film+group", "strength~film*group"),
               het=c(1,1,1,1), prior="flat", m0=8)
> out_textile$models[1:5,c(1,2,3,5)]
                Model         Scheme.beta       Scheme.Sigma      FModProb
31  strength~film*group {corn}{canna,potato} {potato}{canna,corn}  0.6596667376
8   strength~film*starch              None {potato}{canna,corn}  0.3337588991
30  strength~film*group {canna}{corn,potato} {potato}{canna,corn}  0.0018692078
```

```
28  strength~film*group {corn}{canna,potato} {corn}{canna,potato}  0.0010854755
7   strength~film*starch                 None {corn}{canna,potato}  0.0006831597
```

Refer to code and output above, where we provide the five most probable models. Note the three most probable models all have the latent variance grouping scheme {potato}{canna, corn}; again over 99% of the posterior model probability is accounted for by this variance scheme. This visually agrees with the plot, which shows that the potato starch films seem to have higher variability than the canna and corn starch films. The regression effect structure is less clear: the most probable model selects the film*group model, which contains main effects for film and group as well as their interaction, with scheme {canna}{corn, potato}. We plot this model in the right panel of Figure 3 to illustrate its plausibility. It does appear that the slope for corn is steeper than that of potato and canna, which can be contracted into a single level to simplify the model. However, the error variance for potato appears larger than that of canna and potato, as evidenced by the large spread of square potato points around the gray line. Thus we assert that the most probable model under our methodology is reasonable and appropriate. The group standard errors are $\sigma^2_{\{potato\}} = 57734.046$ and $\sigma^2_{\{canna,corn\}} = 5791.713$, indicating the standard ANCOVA model underestimates the error variance of the potato observations, and overestimates those of the canna and corn observations.

Finally we illustrate the output scheme_probabilities_beta and scheme_probabilities_Sigma, which sum up the probabilities for all model specifications associated with each possible grouping scheme. We see moderately high cumulative probability for models with regression grouping scheme {corn}{canna,potato}, followed closely be models with no grouping scheme for regression effects:

```
> out_textile$scheme_probabilities_beta
          Scheme.beta  Cumulative
2 {corn}{canna,potato} 0.592860983
4                 None 0.403632744
1 {canna}{corn,potato} 0.002502435
3 {potato}{canna,corn} 0.001003838
```

Intuitively, based on the wider spread of the square potato points in Figure 3, we see high cumulative probability for the variance grouping scheme {potato}{canna,corn}:

```
> out_textile$scheme_probabilities_Sigma
          Scheme.Sigma    Cumulative
3 {potato}{canna,corn} 9.975853e-01
2 {corn}{canna,potato} 2.184257e-03
1 {canna}{corn,potato} 2.304323e-04
4                 None 1.632320e-08
```

3.3 Case study 3: locknut data

We consider the two-way replicated layout of Meek and Ozgur (1991), where the torque (torque) required to tighten a locknut was measured as a function of a plating process (plating) and a threading method (fixture).

A two-way analysis with an interaction yields the following ANOVA table. The fixture and plating main effects, along with fixture by plating interaction, are all statistically significant at level $\alpha = 0.005$. Additionally, we find $\hat{\sigma}^2_{\text{Full}} = 36.58$:

```
> anova(lm(Torque~Fixture+Plating+Fixture*Plating, data=locknut))
Analysis of Variance Table

Response: Torque
                Df Sum Sq Mean Sq F value    Pr(>F)
Fixture          1  821.4  821.40 22.4563 1.604e-05 ***
Plating          2 2290.6 1145.32 31.3118 9.363e-10 ***
Fixture:Plating  2  665.1  332.55  9.0916 0.0003952 ***
Residuals       54 1975.2   36.58
```

Upon inspection of Figure 4, we suspect that two latent characteristics are at play. First, based on the non-parallel lines representing the plating effects, there may be a group-by-plating interaction, so we will consider slgf_beta="Plating". Note since fixture has only two levels, it is not feasible to consider group-based effects based on fixture since the one degree of freedom fixture effect would be equivalent to a group effect.
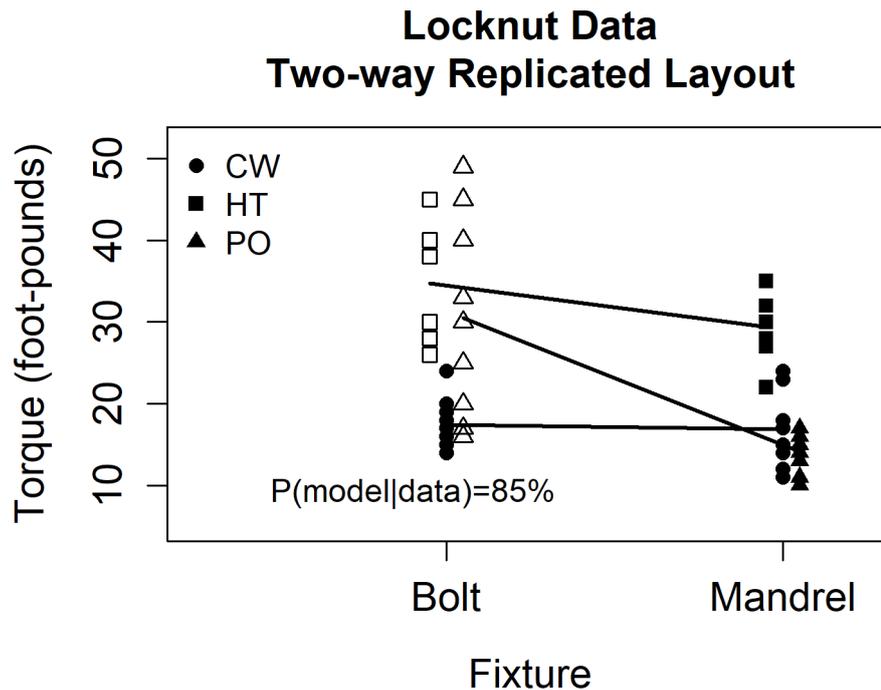
Figure 4: The most probable model ($P(m|Y) \approx 85\%$) contains a full fixture by plating interaction effect with no grouping structure, and group-based variances based on the levels of this interaction with scheme {bolt*CW, mandrel*CW, mandrel*HT, mandrel*PO}{bolt*HT, bolt*PO} (filled points vs. open points).

Regarding the variance structure, the variance of the torque amount at levels PO and HT appears higher, but only for the bolt fixture. This suggests that the levels of the interaction govern the variance groups; that is, `slgf_Sigma="Fixture*Plating"`. Since this specific variable header does not appear in the `locknut` data set, we manually create a new variable with each interaction level by pasting together the main effect variables:

```
locknut$Interaction <- paste0(locknut$Fixture, "*", locknut$Plating)
```

Thus we consider the following model specifications. Liang et al. (2008) (p. 420) note that the Zellner-Siow mixture of *g*-prior provides a fully Bayesian, consistent model selection procedure for small *n* along with relatively straightforward expressions for the marginal model probabilities. This approach is implemented by the user with the argument `prior="zs"`:

```
> data(locknut)
> locknut$Interaction <- paste0(locknut$Fixture, "*", locknut$Plating)
> out_locknut <- ms_slgf(dataf=locknut, response="Torque", same_scheme=FALSE,
              lgf_beta="Plating", min_levels_beta=1,
              lgf_Sigma="Interaction", min_levels_Sigma=1,
              usermodels=list("Torque~Fixture+Plating+Fixture*Plating",
                          "Torque~Fixture+group+Fixture*group"),
              het=c(1,1), prior="zs", m0=2)
```

This formulation favors the same main and interaction effects favors by the standard model. However, **slgf** favors group-based variances with scheme {bolt*HT, bolt*PO}{bolt*CW, mandrel*CW, mandrel*HT, mandrel*PO} with posterior probability of approximately 85%. This variance structure was expected based on the relatively larger spread of the open points in Figure 4. As we have noted previously, the group variance estimates show that the heteroscedastic model overestimates the variance for some levels of fixture and plating, and underestimates it for others. Since model '13' was the model probable model, we print these variances, obtaining $\hat{\sigma}^2_{\text{bolt*HT,bolt*PO}} \approx 85.0$ and $\hat{\sigma}^2_{\text{bolt*CW,mandrel*CW,mandrel*HT,mandrel*PO}} \approx 11.6$:

```
> out_locknut$variances[[13]]
{bolt*HT,bolt*PO} {bolt*CW,mandrel*CW,mandrel*HT,mandrel*PO}
       85.00448                                   11.58652
```

3.4 Case study 4: bottles data

Finally, we consider the data of Ott and Snee (1973), where a machine with six heads (head) is designed to fill bottles (weight). The weight of each bottle is measured once over five time points (time) as a two-way unreplicated layout. A visual inspection of the data (Figure 5, left panel) indicates that one of the filling heads is behaving distinctly than the other five. There appears to be an interaction between head and time, but we lack the degrees of freedom to fit such a model. If we were to fit the standard main effects model, we obtain the clearly inappropriate model fit in the center panel of Figure 5.

Since it appears that head {5} is out of calibration in some way as compared to heads {1,2,3,4,6}, we instead consider the group-based interaction model weight~time+group:time where 'head' is the regression effect SLGF. For this illustration, we consider only homoscedastic models. In this data-poor context, we recommend the use of the Zellner-Siow mixture of g-prior by specifying prior="zs" in the ms_slgf function. The minimal training sample size can be much lower, as this prior is proper. We inspect the posterior model probabilities of the most probable model and the additive main effects model:

```
> bottles_me <- lm(weight~time+heads, data=bottles)
> bottles2 <- data.frame(weight=bottles$weight, time=as.factor(bottles$time),
                         heads=as.factor(bottles$heads))
> bottles_out <- ms_slgf(dataf=bottles2, response="weight", lgf_beta="heads",
                min_levels_beta=1, lgf_Sigma=NA, min_levels_Sigma=NA, same_scheme=FALSE,
                usermodels=list("weight~time+group:time", "weight~time+heads"),
                het=c(0,0), prior="zs", m0=2)
> bottles_out$models[1:2,c(1,2,4,5)]
                 Model      Scheme.beta Log-Marginal   FModProb
5  weight~time+group:time {5}{1,2,3,4,6}    -103.168    0.9991932
32      weight~heads+time          None    -114.726 0.0002158313
```

The group-based approach overwhelmingly favors the model with a main effect for 'time' along with the group-based interaction 'group:time' with scheme {5}{1,2,3,4,6}. We also note that the error variance for the main effects model is $\hat{\sigma}^2_{\text{ME}} = 130.1233$, while the estimate for the group-based interaction model is $\hat{\sigma}^2_{\{5\}\{1,2,3,4,6\}} = 39.76$, suggesting the main effects model seriously overestimates the error variance and thus may lead to misleading inference on regression effects.
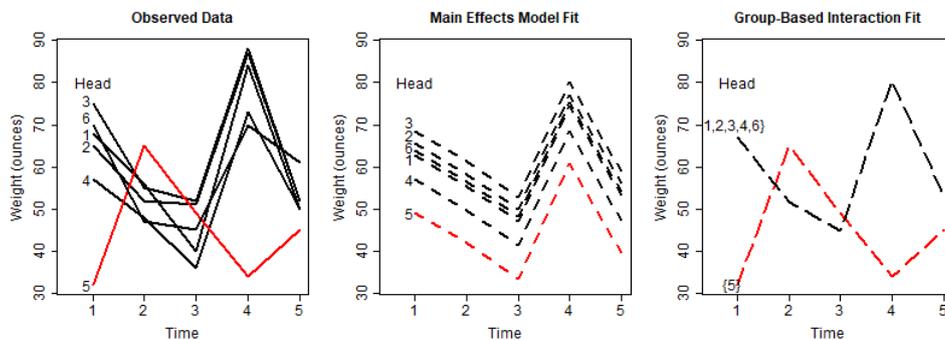

Figure 5: The bottles data set from Ott and Snee (1973) represents fill weights by six machine heads over five time points. The left panel shows the data, with head 5 appearing to be out of calibration. The center panel shows a main effects model, with a realistic fit for heads 1, 2, 3, 4, and 6, but not 5. The right panel shows the most probable group-based interaction ($P(m|\boldsymbol{Y}) > 99.9\%$) with main effects for time and a group-by-time interaction with scheme {5}{1,2,3,4,6}.

We note that there will be a linear dependency between the group-by-time interaction and the time main effect for time 5. The NA values can be seen by inspecting the coefficients of the corresponding model. These effects are not counted in the dimensionality of the model when computing $q^b(\boldsymbol{Y}|m)$.

```
> bottles_out3$coefficients[[5]]
          (Intercept)             heads2                   heads3                   heads4
                53.24               1.80                     4.80                    -6.80
               heads5             heads6     group{1,2,3,4,6}:time1        group{5}:time1
                -8.24              -1.00                    14.00                   -13.00
 group{1,2,3,4,6}:time2    group{5}:time2    group{1,2,3,4,6}:time3        group{5}:time3
```

```
                -1.40               20.00                     -8.20                  4.00
group{1,2,3,4,6}:time4      group{5}:time4  group{1,2,3,4,6}:time5     group{5}:time5
                27.40              -11.00                        NA                    NA
```

4   Conclusion

This manuscript has provided an overview of the **slgf** package in R, which is available from the Comprehensive R Archive Network. Source code can be found on Github at `https://github.com/metzger181osu/slgf`. The **slgf** package allows the user to determine whether latent groupings of categorical predictor's levels provide a better characterization of the response variable compared with ordinary linear models that do not account for the suspected latent groupings. This is accomplished through the *suspected latent grouping factor* methodology of Metzger and Franck (2021). The methodology allows for formal comparisons between ordinary linear models and latent grouping models, which protects the user from automatically selecting a spurious clustering structure that is not well supported by the data. We illustrate the ability to detect the lack of a grouping structure in the simulation studies of Metzger and Franck (2021).

The **slgf** package allows the user to (i) explore different grouping schemes for fixed effects and error variances, and (ii) specify entirely separate latent grouping factors for fixed effects and variances. We illustrate (i) in Case Study 2: Textile data, where the top model shows a different regression line for corn compared to canna and potato, but the error variance for potato is different from canna and corn (see Figure 3). To show (ii), we considered the locknut example of Subsection **Case study 3: locknut data**, where we considered whether fixture (bolt, mandrel) exhibited a fixed effect latent grouping structure, and whether interaction (bolt*CW, bolt*HT, bolt*PO, mandrel*CW, mandrel*HT, mandrel*PO) exhibited a variance latent grouping structure. As described in Subsection **Case study 3: locknut data**, we found no latent grouping structure for fixed effects, but torque error variance for bolt*HT and bolt*PO differ from the other interaction levels. The analysis supported no latent grouping structure for plating.

The **slgf** package provides functionality to detect plausible underlying cluster structures among levels of categorical predictors in linear models. This exercise in cluster detection is in some ways similar to considering a finite mixture model. R packages already exist to fit finite mixture models using the EM algorithm, such as **mixtools** (Benaglia et al., 2009). The **flexmix** package (Gruen and Leisch, 2023) in particular is notable for its ability to fit mixture models to regression data (including Gaussian, binomial, and Poisson models). Additionally, the package **MultiLCIRT** also considers latent variables for the item response theory setting; see Bartolucci et al. (2014), who use BIC for model selection rather than fractional Bayes factors.

In contrast to fitting finite mixture models for the purpose of parameter estimation and inference, **slgf** assesses the plausibility of cluster structures for small to medium-sized data sets via model selection. Additionally, **slgf** can avoid problems with convergence in the EM algorithm that may arise in small-sample scenarios, particularly when the number of data points is relatively low and the model being fit (e.g., a two component mixture model) is larger than the actual model generating the data (e.g., a one component mixture model with no cluster structure).

By contrast, **slgf** circumvents convergence issues by considering all possible groupings of points within the user-specified model classes, obtaining integrated likelihoods and posterior model probabilities for each model, and quantifying overall probability of a cluster structure as the sum of all posterior probabilities for models with two groups by the law of total probability. The **slgf** package thus excels in smaller-data settings where assessing the plausibility of a cluster structure is the core goal, and packages like **flexmix** will excel in cases where the main goal is to fit specified mixture models and conduct inference on parameters.

In addition to the basic **slgf** demonstration shown in Case Study 1: Smell data, we illustrate **slgf** functionality for mixtures of g-priors (Liang et al., 2008) and a two way unreplicated layout in Case Study 4: Bottles data. Mixtures of g-priors have been shown to work well with fractional Bayes factor methods to reduce the training fraction when sample size is small relative to the number of model parameters (Metzger and Franck, 2021).

Finally, although the methodology described here and in Metzger and Franck (2021) exclusively handles two latent groups, we call on any readers with a compelling data set that may exhibit more than two latent groups to contact the authors so that we might explore a generalization of our method to more than two groups.

We have provided an overview of functionality that we hope will enable scientists from diverse fields to access the SLGF methodology of Metzger and Franck (2021) via the **slgf** package to detect hidden groupings in the levels of categorical predictors that might impact outcomes of interest across a wide range of human endeavors.

References

F. Bartolucci, S. Bacci, and M. Gnaldi. Multilcirt: An r package for multidimensional latent class item response models. *Computational Statistics & Data Analysis*, 71:971–985, 2014. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2013.05.018. URL https://www.sciencedirect.com/science/article/pii/S0167947313002053. [p189]

T. Benaglia, D. Chauveau, D. R. Hunter, and D. Young. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29, 2009. URL https://www.jstatsoft.org/v32/i06/. [p189]

C. T. Franck and J. A. Osborne. Exploring Interaction Effects in Two-Factor Studies using the hiddenf Package in R. *The R Journal*, 8(1):159–172, 2016. URL https://journal.r-project.org/archive/2016/RJ-2016-011/index.html. [p175]

C. T. Franck, D. M. Nielsen, and J. A. Osborne. A method for detecting hidden additivity in two-factor unreplicated experiments. *Computational Statistics & Data Analysis*, 67(Supplement C):95 – 104, 2013. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2013.05.002. URL http://www.sciencedirect.com/science/article/pii/S0167947313001618. [p175]

M. S. Furry. Breaking strength, elongation and folding endurance of films of starches and gelatin used in textile sizing. *Technical Bulletin (United States Department of Agriculture)*, 674:1–36, 1939. [p175, 176, 184, 185]

B. Gruen and F. Leisch. *flexmix: Flexible Mixture Modeling*, 2023. URL https://CRAN.R-project.org/package=flexmix. R package version 2.3-19. [p189]

M. Kharrati-Kopaei and S. M. Sadooghi-Alvandi. A new method for testing interaction in unreplicated two-way analysis of variance. *Communications in Statistics - Theory and Methods*, 36(15):2787–2803, 2007. doi: 10.1080/03610920701386851. URL http://dx.doi.org/10.1080/03610920701386851. [p175]

F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger. Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008. doi: 10.1198/016214507000001337. URL https://doi.org/10.1198/016214507000001337. [p178, 180, 187, 189]

G. Meek and C. Ozgur. Torque variation analysis. *Journal of the Industrial Mathematics Society*, 41:1–16, 1991. [p175, 176, 186]

T. A. Metzger and C. T. Franck. Detection of latent heteroscedasticity and group-based regression effects in linear models via bayesian model selection. *Technometrics*, 63(1):116–126, 2021. doi: 10.1080/00401706.2020.1739561. URL https://doi.org/10.1080/00401706.2020.1739561. [p175, 177, 184, 189]

R. G. O'Brien and M. W. Heft. New discrimination indexes and models for studying sensory functioning in aging. *Journal of Applied Statistics*, 22:9–27, 1995. [p175, 176, 182, 183]

A. O'Hagan. Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):99–138, 1995. ISSN 00359246. URL http://www.jstor.org/stable/2346088. [p179, 180]

E. R. Ott and R. D. Snee. Identifying useful differences in a multiple-head machine. *Journal of Quality Technology*, 5(2):47–57, 1973. [p176, 177, 188]

A. Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. In P. Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243. Elsevier Science Publishers, Inc., 1986. [p178]

A. Zellner and A. Siow. Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadistica y de Investigacion Operativa*, 31(1):585–603, Feb 1980. ISSN 0041-0241. doi: 10.1007/BF02888369. URL https://doi.org/10.1007/BF02888369. [p178]

Thomas A. Metzger
Department of Statistics, The Ohio State University
1958 Neil Avenue, Columbus, Ohio 43210
United States of America

ORCiD: 0000-0003-3620-1405
metzger.181@osu.edu

Christopher T. Franck
Department of Statistics, Virginia Tech
403E Hutcheson Hall, Blacksburg, VA 24060
United States
(ORCiD: 0000-0003-1251-4378)
chfranck@vt.edu