Objective Bayesian Variable Selection in
Well-Formulated Models

Short Title: Variable Selection in Well-Formulated Models

Daniel Taylor-Rodriguez      Andrew Womack      Nikolay Bliznyuk

November 27, 2013

Abstract

This paper investigates objective Bayesian variable selection when there is a hier-
archical dependence structure on the inclusion of predictors in the model, that, when
ignored, makes the variable selection dependent on how the predictors are coded. In
particular, we study the type of dependence found in polynomial response surfaces
of orders two and higher. We develop three classes of priors on the model space,
investigate both their operating characteristics and their effect on the variable selec-
tion outcome, and provide a Metropolis-Hastings algorithm for searching the space of
models. The tools proposed allow fast and thorough exploration of model spaces that
account for hierarchical polynomial structure in the predictors. Proposed choices for
the model priors allow for strong control over the number of false positives included in
“good” models.

Keywords: Markov Chain Monte Carlo; mixtures of g-priors; model priors; multiple
testing; multiplicity penalization; well-formulated models.

1 Introduction

In modern regression problems choosing subsets of good predictors from a large pool is
standard practice, especially in the increasingly abundant big data problems. For regressions
with a large number of predictors the complete model space is too large to enumerate and
automatic selection algorithms are necessary to find appropriate parsimonious models. This
multiple testing problem is further complicated when it is critical to incorporate interactions
or powers of the predictors. Not only are there a large number of predictors, but also the 
that structure underlies the space of predictors must be taken into account.

More than two decades years ago, Peixoto (1990) exposed the relevance of respecting 
polynomial hierarchy among covariates in the variable selection context. To make model 
selection invariant to coding transformations in polynomial interaction models (e.g., to cen-
tering of the main effects), the selection must be constrained to the subset of models that 
respect the polynomial hierarchy (Griepentrog et al., 1982; Khuri, 2002; McCullagh and 
Nelder, 1989; Nelder, 2000; Peixoto, 1987, 1990). Models that respect the hierarchy are 
referred to as well-formulated models (WFMs). Succinctly, a model is well-formulated if for 
any predictor in the model every lower order predictor associated with it is also in the model. 
For example, models using $x_1^2$ or $x_1 x_2$ but not $x_1$ are not well formulated.

Although research on this topic started more than three decades ago (Nelder, 1977), only 
recently have modern variable selection techniques been adapted to account for the con-
straints imposed by the polynomial hierarchy. As described in Bien et al. (2013), the current 
literature on variable selection for polynomial response surface models can be classified into 
three broad groups: multistep procedures (Brusco et al., 2009; Peixoto, 1987), regularized 
regression methods (Bien et al., 2013; Yuan et al., 2009), and Bayesian approaches (Chip-
man, 1996, 1998; Chipman et al., 1997). The methods introduced in this paper build on the 
analysis of WFMs from a Bayesian perspective.

The Bayesian variable selection problem consists of comparing models $M$ in a model 
space $\mathcal{M}$ by their posterior probabilities, given by $p(M|y, \mathcal{M}) \propto m(y|M)\pi(M|\mathcal{M})$. Model 
posterior probabilities depend on the prior on the model space as well as on the priors for the 
model-specific parameters, implicitly through the marginals $m(y|M)$. Priors on the model-
specific parameters have been extensively discussed in the literature (Berger and Pericchi, 
1996; Berger et al., 2001; George, 2000; Jeffreys, 1961; Kass and Wasserman, 1996; Liang 
et al., 2008; Zellner and Siow, 1980). In contrast, the effect of the prior on the model space 
has until recently been neglected. Scott and Berger (2010) highlighted the relevance of the 
priors on the model space in the context of multiple testing.

Variable selection within the model space of well-formulated polynomial models poses
two challenges for automatic, objective model selection procedures. First, the notion of
model complexity takes on a new dimension. Complexity is not exclusively a function of the
number of predictors, but also depends upon the depth and connectedness of the predictor
associations defined by the polynomial hierarchy. Second, because the model space is shaped
by such relationships, stochastic search algorithms used to explore the models must also
conform to these restrictions.

In this paper, we develop priors for model spaces of WFM and use intrinsic priors
(Berger and Pericchi, 1996; Moreno et al., 1998) for model-specific parameters. We exploit
the graphical structure of WFM to construct samplers that efficiently traverse the model
space. To the best of our knowledge, this is the only Bayesian framework for exploring
model spaces that investigates the use of priors to penalize over all dimensions of the WFM
model space complexity. Although we focus on the polynomial response surface regression
models with independent homoscedastic normal errors, the proposed ideas can be extended
to well-formulated linear mixed models and generalized linear models.

The manuscript is organized as follows: In Section 2 we review WFM and their corre-
sponding model spaces in terms of directed acyclic graphs (DAGs), facilitating the definition
of the priors and random walks on WFM spaces. In Section 3, we propose, investigate, and
provide recommendations for three possible types of prior distributions on the model space.
We also consider two automatic choices of hyper-parameters and investigate heuristically
their operating properties. Section 4 provides a Metropolis-Hastings algorithm to sample
the model space that takes advantage of its structure. In Section 5, the performance of
the proposed methods is explored through simulation studies. In Section 6, we illustrate
the algorithm with real data and compare the results with other procedures from current
literature. Section 7 contains recommendations and discussion. Supplementary Materials
contain additional details about the methods used and some additional simulation results.
Suppose that the observations $y_i$ are modeled using the polynomial regression as

$$
y_i = \sum_{a_1,\ldots,a_p} \beta(a_1,\ldots,a_p) \prod_{j=1}^p x_{i,j}^{a_j} + \epsilon_i, \quad (1)$$

where $(\alpha_1,\ldots,\alpha_p) = \alpha \in \mathbb{N}_0^p$, $\mathbb{N}_0$ is the set of natural numbers including 0, $\epsilon_i \overset{iid}{\sim} N(0,\sigma^2)$, and only finitely many $\beta_\alpha$ are allowed to be non-zero. As an illustration, consider a model space that includes polynomial terms incorporating covariates $x_{i,1}$ and $x_{i,2}$ only. The terms $x_{i,2}^2$ and $x_{i,1}^2 x_{i,2}$ can be represented by $\alpha = (0,2)$ and $\alpha = (2,1)$, respectively.

The notation $y = Z(X)\beta + \epsilon$ is used to denote that observed response $y = (y_1,\ldots,y_n)$ is modeled via a polynomial function $Z$ of the original covariates $X = (x_1,\ldots,x_p)$, where $x_j = (x_{1,j},\ldots,x_{n,j})'$ and the coefficients of the polynomial terms are given by $\beta$. A specific polynomial model $M$ is defined by the set of coefficients $\beta_\alpha$ that are allowed to be non-zero. This definition is equivalent to characterizing $M$ through a collection of multi-indices $\alpha \in \mathbb{N}_0^p$. In particular, model $M$ is specified by $M = \{\alpha_{M,1},\ldots,\alpha_{M,|M|}\}$ for $\alpha_{M,k} \in \mathbb{N}_0^p$ where $\beta_\alpha = 0$ for $\alpha \not\in M$ and $|M|$ is the cardinality of $M$.

Any particular model $M$ uses a subset $X_M$ of the original covariates $X$ to form the polynomial terms in the design matrix $Z_M(X)$. Without ambiguity, a polynomial model $Z_M(X)$ on $X$ can be identified with a polynomial model $Z_M(X_M)$ on the covariates $X_M$. The coefficient vector and error variance of the model $M$ are denoted by $\beta_M$ and $\sigma^2_M$, respectively.

In vector-matrix notation, the model $M$ models the data as $y = Z_M(X_M)\beta_M + \epsilon_M$, where $\epsilon_M \sim N_n (0, \mathbf{I} \sigma^2_M)$. A model $M$ is said to be nested in the model $M'$ if $M \subset M'$. Model $M$ relates the response to the covariates in two distinct ways: choosing the set of meaningful covariates $X_M$ as well as choosing the polynomial structure of these covariates $Z_M(X_M)$.

The set $\mathbb{N}_0^p$ constitutes a poset that is partially ordered through the binary relation “$\preceq$”, defined between pairs $(\alpha, \alpha')$ by $\alpha' \preceq \alpha$ whenever $\alpha_j \geq \alpha'_j$ for all $j = 1,\ldots,p$ and $\alpha' < \alpha$ if, additionally, $\alpha_j > \alpha'_j$ for some $j$. The order of a term $\alpha \in \mathbb{N}_0^p$ is given by the sum of its elements, $\text{order}(\alpha) = \sum \alpha_j$. If $\text{order}(\alpha) = \text{order}(\alpha') + 1$ and $\alpha' < \alpha$ then $\alpha'$ is said
to immediately precede $\alpha$, which is denoted by $\alpha' \rightarrow \alpha$. The parent set of $\alpha$ is defined by $\mathcal{P}(\alpha) = \{ \alpha' \in \mathbb{N}_0^p : \alpha' \rightarrow \alpha \}$, which is the set of nodes that immediately precede $\alpha$. A polynomial model $M$ is said to be well-formulated if $\alpha \in M$ implies that $\mathcal{P}(\alpha) \subset M$. For example, any well-formulated model using $x_1^2 x_2$ must also include the parent terms $x_1 x_2$ and $x_1^2$, their corresponding parent terms $x_1$ and $x_2$, and the intercept term $1$.

The poset $\mathbb{N}_0^p$ can be represented by a directed acyclic graph (DAG), denoted by $\Gamma(\mathbb{N}_0^p)$, with directed edges to a node from its parents. Without ambiguity, we can identify nodes in the graph $\alpha \in \mathbb{N}_0^p$ with monoids on the set of covariates. Any well-formulated model $M$ is represented by a subgraph $\Gamma(M)$ of $\Gamma(\mathbb{N}_0^p)$ with the property that if node $\alpha \in \Gamma(M)$, then the nodes corresponding to $\mathcal{P}(\alpha)$ are also in $\Gamma(M)$. Figure 2.1 illustrates a well-formulated and a non-well-formulated model.

![Figure 2.1: Examples of graphs for WFM and non-WFM spaces.](image)

The motivation for considering only well-formulated polynomial models is compelling. The subspace of $y$ modeled by $Z_M$, given by the hat matrix $H_M = Z_M(Z_M'Z_M)^{-1}Z_M'$, is invariant to affine transformations of the matrix $X_M$ if and only if $M$ corresponds to a well-formulated polynomial model (Peixoto, 1990). For example, if $p = 2$ and $y = \beta_{(0,0)} + \beta_{(1,0)} x_1 + \beta_{(0,1)} x_2 + \beta_{(1,1)} x_1 x_2 + \epsilon$, then the hat matrix is invariant to any covariate transformation of the form $(x_1 \ x_2)A + b$ for any real nonsingular $2 \times 2$ matrix $A$ and any real column vector $b$ of length two. In contrast, if $y = \beta_{00} + \beta_{20} x_1^2 + \epsilon$, the hat matrix formed after applying the transformation $x_1 \mapsto x_1 + c$ for real $c \neq 0$ is not the same as the hat matrix formed by the original $x_1$. This will make the selection depend on how the variables are coded, which is
undesirable when conducting a model selection procedure aimed at understanding the true
data generating mechanism.

WFMs and their associated DAGs are often discussed in terms of heredity. WFMs exhibit
strong heredity (Chipman, 1996), where all lower order terms dividing an included higher
order term must also be included. An alternative is weak heredity (Chipman, 1996), which
only forces some of the lower order terms in the corresponding polynomial hierarchy to
be in the model. However, Nelder (1998) demonstrated that the conditions, under which
weak heredity allows the design matrix to be invariant to coding, are seldom met exactly
in practice. Thus, this paper focuses on polynomial response surface models that exhibit
strong heredity.

The spaces of WFMs, $\mathcal{M}$, considered in this paper are characterized by two WFMs: $M_B$,
the base model, and $M_F$, the full model, where $1 \in M_B \subset M_F$. For example, $M_B$ can
incorporate covariates describing the structure of an experimental design. The model space
$\mathcal{M}$ is populated by all well formulated models $M$ that contain $M_B$ and are nested in $M_F$.

For convenience, define $\Upsilon(M) = M \setminus M_B$. Model $M \in \mathcal{M}$ can be uniquely identified by
two different sets of nodes in $M_F$: the set of extreme nodes, and the set of children nodes,
defined by $\mathcal{E}(M) = \{\alpha \in \Upsilon(M) : \alpha \notin P(\alpha') \forall \alpha' \in M\}$ and $\mathcal{C}(M) = \{\alpha \in \Upsilon(M_F) \setminus \Upsilon(M) : P(\alpha) \subset M\}$, respectively. The extreme nodes are the leaves in the DAG; if the set $\mathcal{E}(M)$
is removed from $M$, $M \setminus \mathcal{E}(M)$ is also a WFM. The children nodes are those that, when
added to $M$ give rise to a WFM in $\mathcal{M}$. Because $M_B \subseteq M$ for all $M \in \mathcal{M}$, the set of nodes
$\mathcal{E}(M) \cup M_B$ determines $M$ by beginning with this set and iteratively adding parent nodes.

Similarly, the nodes in $\mathcal{C}(M)$ determine the set $\{\alpha' \in P(\alpha) : \alpha \in \mathcal{C}(M)\} \cup \{\alpha' \in \mathcal{E}(M_F) : \alpha \not\preceq \alpha' \text{ for all } \alpha \in \mathcal{C}(M)\}$, which also uniquely identifies $M$.

In Figure 2.2, the extreme and children sets for model $M = \{1, x_1, x_2^2\}$ are shown for the
model space characterized by $M_F = \{1, x_1, x_2, x_1^2, x_1x_2, x_2^2\}$. In Figure 2(a), the solid nodes
represent nodes $\alpha \in M \setminus \mathcal{E}(M)$, the dashed node corresponds to $\alpha \in \mathcal{E}(M)$, and the dotted
nodes are not in $M$. The solid nodes in Figure 2(b) correspond to those in $M$, the dashed
node is the single node in $\mathcal{C}(M)$, and the dotted nodes are not in $M \cup \mathcal{C}(M)$.
Figure 2.2: $E(M)$ (left) and $C(M)$ (right) in the model space $M$ defined by a quadratic surface in two main effects for model $M = \{1, x_1, x_1^2\}$.

3 Priors on the model space

In this section, we develop three different prior structures on the model space for WFM$s$, discuss their advantages and disadvantages, and describe reasonable choices for their hyperparameters. The hyper-parameter choices are motivated by the multiplicity prior from Scott and Berger (2010) and the penalization prior of Wilson et al. (2010). Finally, we investigate how the choices of prior structure and hyper-parameter affect the posterior probabilities for predictor inclusion and provide recommendations for several scenarios.

3.1 Model prior definition

As was discussed in Section 2, the graphical structure for the model spaces suggests a method for prior construction on $M$, guided by the notion of inheritance. A node $\alpha$ is said to inherit from a node $\alpha'$ if $\alpha' \preceq \alpha$. The inheritance is said to be immediate if $\alpha \rightarrow \alpha'$.

For $\alpha \in \Upsilon(M_F)$, let $\gamma_\alpha(M)$ be the indicator function describing whether $\alpha$ is included in $M$. Let $\gamma^\nu(M) = \{\gamma_\alpha(M) : \text{order}(\alpha) = \nu\}$ and $\gamma^{<\nu}(M) = \bigcup_{j=0}^{\nu-1} \gamma^j(M)$. With these definitions, the prior probability of any model $M \in M$ can be factored as

$$\pi(M | M) = \prod_{j=J_M^{\min}}^{J_M^{\max}} \pi(\gamma^j(M) | \gamma^{<j}(M), M),$$

where $J_M^{\min}$ and $J_M^{\max}$ are, respectively, the minimum and maximum orders of nodes in $\Upsilon(M_F)$,
and \( \pi(\gamma_{\mathcal{M}}^{\text{min}}(M) | \gamma^{< \text{min}}_{\mathcal{M}}(M), \mathcal{M}) = \pi(\gamma_{\mathcal{M}}^{\text{min}}(M) | \mathcal{M}). \)

Prior distributions on \( \mathcal{M} \) can be simplified by making two assumptions (Chipman, 1996). First, if order(\( \alpha \)) = order(\( \alpha' \)) = \( j \) then \( \gamma_{\alpha} \) and \( \gamma_{\alpha'} \) are assumed to be conditionally independent when conditioned on \( \gamma^{< j} \), denoted by \( \gamma_{\alpha} \perp \perp \gamma_{\alpha'} | \gamma^{< j} \). Second, by invoking immediate inheritance, if order(\( \alpha \)) = \( j \) then \( \pi(\gamma_{\alpha}(M) | \gamma^{< j}(M), \mathcal{M}) = \pi(\gamma_{\alpha}(M) | \gamma_{\mathcal{P}(\alpha)}(M), \mathcal{M}), \) where \( \gamma_{\mathcal{P}(\alpha)}(M) \) is the inclusion indicator for the set of parent nodes of \( \alpha \). Figure 3.1 helps to visualize these assumptions when \( M_F \) is an order-two surface in two main effects.

Let \( \pi_{\alpha} = \pi(\gamma_{\alpha}(M) = 1 | \gamma_{\mathcal{P}(\alpha)}(M), \mathcal{M}) \) be the conditional inclusion probability of node \( \alpha \) in model \( M \). Under the assumptions of conditional independence and immediate inheritance the prior probability of \( M \) is \( \pi(M | \pi_{\mathcal{M}}, \mathcal{M}) = \prod \pi_{\alpha}^{\gamma_{\alpha}(M)}(1 - \pi_{\alpha})^{1 - \gamma_{\alpha}(M)}, \) with \( \pi_{\mathcal{M}} = \{ \pi_{\alpha} : \alpha \in \Upsilon(M_F) \} \). Because \( M \) must be well-formulated, \( \pi_{\alpha} = \gamma_{\alpha} = 0 \) if \( \gamma_{\mathcal{P}(\alpha)}(M) = 0 \). Thus, the product can be restricted to the set of nodes \( \alpha \in \Upsilon(M) \cup \mathcal{C}(M) \). Structure can be built into the prior on \( \mathcal{M} \) by making assumptions about the inclusion probabilities \( \pi_{\alpha} \), such as equality assumptions or assumptions of a hyper-prior for these parameters. Three such prior classes that we develop below incorporate the equality assumptions in Figure 3.2.
Hierarchical Uniform Prior (HUP)  The HUP assumes that the non-zero probabilities \( \pi_{\alpha} \) are all equal (Figure 2(a)). Specifically, for a model \( M \in \mathcal{M} \) it is assumed that \( \pi_{\alpha} = \pi \) for all \( \alpha \in \Upsilon(M) \cup C(M) \). A complete Bayesian specification of the HUP is completed by assuming a prior distribution for \( \pi \). The choice of \( \pi \sim Beta(a, b) \) produces

\[
\pi^{HUP}(M|\mathcal{M}, a, b) = \frac{B(|\Upsilon(M)| + a, |C(M)| + b)}{B(a, b)}, \tag{3}
\]

where \( B \) is the beta function. Setting \( a = b = 1 \) gives the particular value of

\[
\pi^{HUP}(M|\mathcal{M}, a = 1, b = 1) = \frac{1}{|\Upsilon(M)| + |C(M)| + 1} \left( \frac{|\Upsilon(M)| + |C(M)|}{|\Upsilon(M)|} \right)^{-1}. \tag{4}
\]

The HUP assigns equal probabilities to all models for which the sets of nodes \( \Upsilon(M) \) and \( C(M) \) have the same cardinality. This prior provides a combinatorial penalization, but essentially fails to account for the hierarchical structure of the model space. An additional penalization for model complexity can be incorporated into the HUP by changing the values of \( a \) and \( b \). Because \( \pi_{\alpha} = \pi \) for all \( \alpha \), this penalization can only depend on some aspect of the entire graph of \( M_F \). One such penalization is to take \( a = 1 \) and \( b = |\Upsilon(M_F)| \).

Hierarchical Independence Prior (HIP)  The HIP assumes that there are no equality constraints among the non-zero \( \pi_{\alpha} \) (Figure 2(b)). Each non-zero \( \pi_{\alpha} \) is given its own prior, which is assumed to be a Beta distribution with parameters \( a_{\alpha} \) and \( b_{\alpha} \). Thus, the prior
probability of $M$ under the HIP is

\[
\pi^{HIP}(M|M, a, b) = \left( \prod_{\alpha \in \Upsilon(M)} \frac{a_{\alpha}}{a_{\alpha} + b_{\alpha}} \right) \left( \prod_{\alpha \in C(M)} \frac{b_{\alpha}}{a_{\alpha} + b_{\alpha}} \right)
\]

(5)

where the product over the empty set is taken to be one. Because the $\pi_{\alpha}$ are totally independent, any choice of $a_{\alpha}$ and $b_{\alpha}$ is equivalent to choosing a probability of success $\pi_{\alpha}$ for a given $\alpha$. Setting $a_{\alpha} = b_{\alpha} = 1$ for all $\alpha \in \Upsilon(M) \cup C(M)$ gives the particular value of

\[
\pi^{HIP}(M|M, a = 1, b = 1) = \left( \frac{1}{2} \right)^{|\Upsilon(M)| + |C(M)|}.
\]

(6)

While the prior with this choice of hyper-parameters accounts for the hierarchical structure of the model space, it essentially provides no penalization for combinatorial complexity at different levels of the hierarchy. This can be observed by considering a model space with main effects only; the exponent in equation (6) is the same for every model in the space because each node is either in the model or in the children set.

An additional penalization for model complexity can be incorporated into the HIP. Because each $\gamma^j$ is conditioned on $\gamma^{<j}$ in the prior construction, the $a_{\alpha}$ and $b_{\alpha}$ for $\alpha$ of order $j$ can be conditioned on $\gamma^{<j}$. One such additional penalization utilizes the number of nodes of order $j$ that could be added to produce a WFM conditioned on the inclusion vector $\gamma^{<j}$, which is denoted as $ch_j(\gamma^{<j})$. Choosing $a_{\alpha} = 1$ and $b_{\alpha}(M) = ch_j(\gamma^{<j})$ is equivalent to choosing a probability of success $\pi_{\alpha} = (1 + ch_j(\gamma^{<j}))^{-1}$.

Hierarchical Order Prior (HOP) A compromise between complete equality and complete independence of the $\pi_{\alpha}$ is to assume equality between the $\pi_{\alpha}$ of a given order and independence across the different orders, as shown in Figure 2(c). Define $\Upsilon_j(M) = \{\alpha \in \Upsilon(M) : \text{order}(\alpha) = j\}$ and $C_j(M) = \{\alpha \in C(M) : \text{order}(\alpha) = j\}$. The HOP assumes that $\pi_{\alpha} = \pi^{(j)}$ for all $\alpha \in \Upsilon_j(M) \cup C_j(M)$. Assuming that $\pi^{(j)} \sim Beta(a_j, b_j)$ provides a prior
probability of

\[
\pi^{\text{HOP}}(M|M, a, b) = \prod_{j=J_{\min}}^{J_{\max}} \frac{B(|Y_j(M)| + a_j, |C_j(M)| + b_j)}{B(a_j, b_j)}
\]  

(7)

The specific choice of \(a_j = b_j = 1\) for all \(j\) gives a value of

\[
\pi^{\text{HOP}}(M|M, a = 1, b = 1) = \prod_j \left[ \frac{1}{|Y_j(M)| + |C_j(M)| + 1} \left( \frac{|Y_j(M)| + |C_j(M)|}{|Y_j(M)|} \right)^{-1} \right]^{-1}
\]  

(8)

and produces a hierarchical version of the Scott and Berger (2010) multiplicity correction.

An additional complexity penalization can be incorporated into the HOP in a similar fashion to the HIP. The number of possible nodes of order \(j\) that could be added while maintaining a WFM is given by \(ch_j(M) = ch_j(\gamma^{<j}(M)) = |Y_j(M) \cup C_j(M)|\). Using \(a_j = 1\) and \(b_j(M) = ch_j(M)\) produces a hierarchical version of the complexity penalizing prior introduced in Wilson et al. (2010).

### 3.2 Choice of prior structure and hyper-parameters

Each form of the priors introduced in Section 3.1 defines a whole family of model priors, characterized by the probability distribution for the inclusion probabilities \(\pi_M\). For the sake of simplicity, this paper focuses on those arising from Beta-Binomial distributions and emphasizes particular choices of hyper-parameters which can be specified automatically. As there is an infinite number of ways in which the hyper-parameters can be specified, we focus on the default choice \(a = b = 1\) as well as the complexity penalizations described in Section 3.1. The second alternative is referred to as \(a = 1, b = ch\), where \(b = ch\) has slightly different interpretations depending on the prior structure. Accordingly, \(b = ch\) is given by \(b_j(M) = b_\alpha(M) = ch_j(M) = |Y_j(M) \cup C_j(M)|\) for the HOP and HIP, where \(j = \text{order}(\alpha)\), while \(b = ch\) denotes that \(b = |Y(M_F)|\) for the HUP. Figures 3.3 and 3.4 illustrate the prior behavior for two model spaces. In both cases, the base model \(M_B\) is taken to be the intercept only model and \(M_F\) is the DAG shown. The priors considered treat model complexity differently, and some general features can be seen from these examples.
First, compare the choices of HIP, HUP, and HOP for \((a, b) = (1, 1)\). The HIP induces a complexity penalization that only accounts for the order of the terms in the model. This is best exhibited by the model space in Figure 3.3. Models including \(x_1\) and \(x_2\), models 6 through 13, are given the same prior probability and no penalization is incurred for the
inclusion of any or all of the quadratic terms. The HUP induces a penalization for model complexity that does not adequately penalize models for including additional terms. In both Figures 3.3 and 3.4, models including all of the terms are given at least as much probability as any model containing any non-empty set of terms. This lack of penalization of the full model is induced by the combinatorial simplicity of the the full model, which produces a model space distribution that favors the base and full models. Similar behavior is observed with the HOP. As models become more complex, they are appropriately penalized for their size. However, after a sufficient number of nodes are added, the number of possible models of that particular size is considerably reduced, and thus combinatorial complexity is negligible on the largest models. This is best exhibited in Figure 3.4 where the HOP places more mass on the full model than on any model containing a single order one node.

In contrast, if \((a, b) = (1, ch)\) then all three priors produce strong penalization as models become more complex, both in terms of the number and order of the nodes contained in the model. For all of the priors, adding a node \(\alpha\) to a model \(M\) to form \(M'\) produces \(p(M) \geq p(M')\). However, there are clear differences between the priors. The HIP penalizes the full model the most, with the HOP penalizing it the least, and the HUP lying in-between. At face value, the HOP creates the most compelling penalization in terms of model complexity. In Figure 3.4, the penalization of the HOP is the least dramatic, producing prior odds of 20 for \(M_B\) versus \(M_F\) as opposed to the HUP and HIP, which produce prior odds of 40 and 54, respectively. Similarly, the prior odds in Figure 3.3 are 60, 180, and 256 for the HOP, HUP, and HIP, respectively.

### 3.3 Posterior sensitivity to the choice of prior

To determine how the proposed priors are adjusting the posterior probabilities to account for multiplicity, a simple simulation was performed. The goal of this exercise is to understand how the posterior analysis is affected by the priors proposed in response to increasing complexity. First, the priors are compared as the number of main effects \(p\) grows. Second, they are compared as the depth of the hierarchy – captured by the order \(J_{\text{max}}^M\) – increases. Finally, posterior behavior is observed in unstructured (all terms are order one) and irregular
WFM spaces (polynomial hierarchy in only some predictors). This section only outlines the results of the experiments; detailed results are provided in Section A of the Supplementary Materials.

The results were extracted for the median probability model (MPM), which includes all terms whose marginal posterior inclusion probability, $p_\alpha = \sum_{M \in \mathcal{M}} \gamma_\alpha(M)p(M|y, \mathcal{M})$, is greater than 0.5. The posteriors were calculated for the proposed priors as well as the equal probability prior (EPP) on $\mathcal{M}$. For all prior structures, both the default hyper-parameters $a = b = 1$ and the penalizing choice of $a = 1$ and $b = ch$ were considered. The results, for the different combinations of $M_F$ and $M_T$ incorporated in the analysis, were obtained from 100 random replicates of covariates and response.

The MPM was compared to the true model $M_T$ to assess posterior sensitivity to the prior in terms of the mean number of true and false positive terms. The true positives (TP) are defined as those nodes $\alpha \in M_T$ such that $p_\alpha > 0.5$. With the false positives (FP), three different cutoffs are considered for $p_\alpha$, elucidating the adjustment for multiplicity induced by the model priors. These cutoffs are 0.10, 0.20, and 0.50 for $\alpha \notin M_T$. These results provide insight about the influence of the prior on the marginal posterior inclusion probabilities.

**Growing number of predictors, fixed polynomial degree** As the number of main effects in $M_F$ grows, considering the lower FP cutoffs makes the number of false positives increase, both for the EPP and the hierarchical priors with $(a, b) = (1, 1)$. Conversely, the hierarchical priors with $(a, b) = (1, ch)$ produced Median Probability Models with a significantly lower number of FPs even with lower thresholds for the FP. In these simulations all priors performed similarly in terms of the number of TPs, with slightly lower values for the priors with $(a, b) = (1, ch)$.

**Growing polynomial degree, fixed number of predictors** When the complexity is increased by making the order of $M_F$ larger, the inability of the EPP to adjust the inclusion posteriors for multiplicity becomes more pronounced; the EPP becomes less and less efficient at removing false positives when the FP cutoff is low. Among the priors with $a = b = 1$, as the order increases, the HIP does the best job at filtering out the FPs. Using the 50% false
positive cutoff some false positives are included, both for the EPP and for all the priors with
\( a = b = 1 \), indicating that the default hyper-parameters are not the best option to control
FPs. The seven covariates in the true model have high posterior inclusion probabilities under
these four priors. In contrast, any of the \( a = 1 \) and \( b = ch \) priors dramatically improve
upon their \( a = b = 1 \) counterparts, consistently assigning low inclusion probabilities for the
majority of the false positive terms, even for low cutoffs. As the order of the polynomial
surface increases the difference in performance between these priors and either the EPP or
their default versions becomes even more clear. At the 50% cutoff, the hierarchical priors
with complexity penalization exhibit very low false positive rates. The true positive rate
decreases slightly for the priors, but not to an alarming degree.

**Other model spaces**  The first example is a model space with 18 main effects only. Two
true models are used in combination with each model space, one “large” and one “small”
true model. When \(| M_T | = 16 \), the HUP(1, 1) and HOP(1, 1) do poorly at controlling false
positives, even at the 50% cutoff. In contrast, the HIP(1, 1) and EPP, with the 50% cutoff
identifies the true positives and no false positives. This result, however, does not imply that
the EPP is better at controlling false positives, it has to be considered in light of the size
of the true model which contains 16 out of the 18 terms in \( M_F \). The \( a = 1 \) and \( b = ch \)
priors show dramatically different behavior. The HIP controls the number of false positives
well, but fails to identify the true coefficients at the 50% cutoff. In contrast, the HOP
identifies all of the true positives and has a small false positive rate for the 50% cutoff.
When \(| M_T | = 4 \), most terms in the full model are truly zero. The EPP includes at least
one false positive in approximately 50% of the randomly sampled datasets. On the other
hand, the HUP(1, 1) and HOP(1, 1) obtain on average a lower number of false positives
than the EPP and HIP(1, 1). Furthermore, the hierarchical priors with \( (a = 1, b = ch) \)
are substantially better than the EPP (and the choice of \( a = b = 1 \)) at controlling false
positives and capturing all TPs using the marginal posterior inclusion probabilities. The
two examples suggest that the HOP(1, ch) is the best choice for model selection when the
number of terms available at a given degree is large.
The second example includes a full quadratic surface of order 2, but additionally includes six terms for which only main effects are considered; the true models have $|M_T| = 10$ and $|M_T| = 6$. The HIP($1, 1$) and EPP again behave quite similarly, incorporating a large number of false positives for the 0.1 cutoff; at 0.5 cutoff some false positives are still included. The HUP($1, 1$) and HOP($1, 1$) behave similarly with a slightly higher false positive rate at the 50% cutoff. In terms of TP, the EPP and $a = b = 1$ priors always include all the coefficients in $M_T$, regardless of its size. On the other hand, the ability of the $a = 1, b = ch$ priors to control for false positives is markedly better than that of the EPP and the hierarchical priors with choice of $a = 1 = b = 1$. At the 50% cutoff, these priors identify all of the true positives and true negatives. Once again, these examples point to the hierarchical priors with additional penalization for complexity as being good default priors on the model space.

### 4 Random Walks on the Model Space

When the model space $\mathcal{M}$ is too large to enumerate, a stochastic procedure can be used to find models with high posterior probability. In particular, an MCMC algorithm — in particular, the Metropolis-Hastings (MH) algorithm — can be utilized to generate a dependent sample of models from the model posterior. The structure of the model space $\mathcal{M}$ both presents difficulties and provides clues on how to build algorithms to explore it. There are several ways to implement such an MH algorithm, three of which are outlined in this section and are based on local, global, and intermediate steps. Combining the different strategies (Tierney, 1994), e.g., by mixing Markov chain kernels, will allow the MH algorithm to explore the model space thoroughly and relatively quickly. In addition, because the method is a balanced random walk, one can employ either renormalization or visit frequency to compute posterior quantities of interest, such as model posterior probabilities; the sampler is uniformly ergodic since the model space is finite (Garcia-Donato and Martinez-Beneito, 2013). We describe the algorithms succinctly here and more details can be found in Section B of the Supplementary Materials.
Global Jump In order to keep the chain from getting stuck in a local mode, the global jump kernel generates independently at random a model from the prior distribution. The MH correction is the ratio of Bayes’ factors for the proposed and current model.

Local Jump Given a current model $M$, the set $\mathcal{E}(M) \cup \mathcal{C}(M)$ contains the nodes whose inclusion can be changed while maintaining a WFM. The local jump kernel implements a stochastic forwards-backwards proposal kernel conditioned on $M$. For each $\alpha \in \mathcal{E}(M)$ the model $M_{\alpha}$ is defined to be the model $M \setminus \{\alpha\}$ and for each $\alpha \in \mathcal{C}(M)$ it is defined to be $M \cup \{\alpha\}$. The proposal kernel is supported only on the set of models \{\$M_{\alpha} : \alpha \in \mathcal{E}(M) \cup \mathcal{C}(M)\}$ and the proposal probabilities are a mixture of the models’ relative posterior probabilities and uniform proposal kernel. The standard MH correction is applied.

Intermediate Jump The intermediate jump proposes a model by making proposals at each order. The algorithm either increases the order of the nodes being modified from $J_{\min}^M$ to $J_{\max}^M$ or decreases it. If increasing the order, then at a given order $k + 1$, a model $M'_{k+1}$ is proposed from the set \{\$M'_k \cup \{M'_{k,\alpha} : \alpha \in (\mathcal{E}(M'_k) \cup \mathcal{C}(M'_k)) \cap \Upsilon(M'_k)\}\} with proposal probabilities that are a mixture of the models’ relative posterior probabilities and uniform proposal kernel. Decreasing the order is the reverse kernel of increasing the order and defined analogously. The MH correction is calculated by accounting for the posterior probabilities of the proposed model ($M'_{J_{\max}^M}$ or $M'_{J_{\min}^M}$ if increasing or decreasing order, respectively) and current model $M$. The proposal probability is calculated by taking into account the conditional nature in which the proposed model was constructed.

5 Simulation Study

To study the operating characteristics of the proposed priors, a simulation experiment was designed with three goals. First, the priors are characterized by how the posterior distributions are affected by the sample size and the signal-to-noise ratio (SNR). Second, given the SNR level, the influence of the allocation of the signal across the terms in the model is investigated. Third, performance is assessed when the true model has special points in the
scale (McCullagh and Nelder, 1989), i.e., when the true model has coefficients equal to zero for some lower-order terms in the polynomial hierarchy.

With these goals in mind, sets of predictors and responses are generated under various experimental conditions. The model space is defined with $M_B$ being the intercept-only model and $M_F$ being the complete order-four polynomial surface in five main effects that has with 126 nodes. The entries of the matrix of main effects are generated as independent standard normal. The response vectors are drawn from the $n$-variate normal distribution as $y \sim N_n (Z_{M_T}(X)\beta, \gamma, I_n)$, where $M_T$ is the true model and $I_n$ is the $n \times n$ identity matrix.

The sample sizes considered are $n \in \{130, 260, 1040\}$, which ensures that $Z_{M_F}(X)$ is of full rank. The cardinality of this model space is $|M| > 1.2 \times 10^{22}$, which makes enumeration of all models unfeasible. Since the value of the $2k$-th moment of the standard normal distribution increases with $k = 1, 2, \ldots$, higher-order terms by construction have a larger variance than their ancestors. As such, assuming equal coefficients, higher-order terms necessarily contain more “signal” than the terms from which they inherit. Once a higher-order term is selected, its entire ancestry is also included. Therefore, to prevent the simulation results from being overly optimistic (because of the larger signals from the higher-order terms) sphering is used to calculate meaningful values of the coefficients, ensuring that the signal is of the magnitude intended in any given direction. Given the results of the simulations from Section 3.3, only the HOP with $a = 1, b = ch$ is considered, with the EPP included for comparison.

The total number of combinations of SNR, sample size, regression coefficient values, and nodes in $M_T$ amounts to 108 different scenarios. Each scenario was run with 100 independently generated datasets, and the mean behavior of the samples was observed. The results presented in this section correspond to the median probability model (MPM) from each of the 108 simulation scenarios considered. Figure 5.2 shows the comparison between the two priors for the mean number of true positive (TP) and false positive (FP) terms. Although some of the scenarios consider true models that are not well-formulated, the smallest well-formulated model that stems from $M_T$ is always the one shown in Figure 5.1.

The results are summarized in Figure 5.2. Each point on the horizontal axis corresponds to the average for a given set of simulation conditions. Only labels for the SNR and sample
size are included for clarity, but the results are also shown for the different values of the regression coefficients and the different true models considered. Additional details about the procedure and other results are included in the Section C of the Supplementary Materials.

**SNR and sample size effect** As expected, small sample sizes conditioned upon a small SNR impair the ability of the algorithm to detect true coefficients with both the EPP and HOP(1, ch), with this effect more notorious when using the latter prior. However, considering the mean number of TPs jointly with number of FPs, it is clear that although the number of TPs is specially low with HOP(1, ch), most of the few predictors that are discovered in fact belong to the true model. In comparison to the results with EPP in terms of FPs the HOP(1, ch) does better, and even more so when both the sample size and the SNR are smallest. Finally, when either the SNR or the sample size is large the performance in terms of TPs is similar between both priors, but the number of FPs are somewhat lower with the HOP.

**Coefficient magnitude** Three ways to allocate the amount of signal across predictors are considered. For the first choice, all coefficients contain the same amount of signal regardless of their order; in the second, each order-one coefficient contains twice as much signal as any
Figure 5.2: Average true positives (TP) and average false positives (FP) in all simulated scenarios for the median probability model with EPP and HOP(1, ch).

order-two coefficient, and four times as much as any order-three coefficient; and third, each order-one coefficient contains a half as much signal as any order-two coefficient, and a quarter of what any order-three coefficient has. These choices are denoted by $\beta^{(1)} = c(1_{o1}, 1_{o2}, 1_{o3})$, $\beta^{(2)} = c(1_{o1}, 0.5_{o2}, 0.25_{o3})$, and $\beta^{(3)} = c(0.25_{o1}, 0.5_{o2}, 1_{o3})$, respectively. In Figure 5.2 the first 4 scenarios correspond to simulations with $\beta^{(1)}$, the next four use $\beta^{(2)}$, the next four correspond to $\beta^{(3)}$, and then the values are cycled in the same way. The results show that scenarios using either $\beta^{(1)}$ and $\beta^{(3)}$ behave similarly, contrasting with the negative impact of allocating the most signal in the order-one terms through $\beta^{(2)}$. In Figure 5.2 the effect of using $\beta^{(2)}$ is evident, as it corresponds to the lowest values for the TPs regardless of the sample size, the SNR and the prior used. This is an intuitive result since giving more signal to higher-order terms makes it easier to detect higher-order terms, and consequently by strong heredity, the algorithm will also select the corresponding lower-order terms included in the true model.
Special points on the scale  Four true models were considered. First, the model from Figure 5.1 ($M_T^1$); second, the model without the order-one terms ($M_T^2$); third, the model without order-two terms ($M_T^3$); and four, the model without $x_1^2$ and $x_2x_5$ ($M_T^4$). The last three are clearly not well-formulated. In Figure 5.2, the leftmost point on the horizontal axis corresponds to scenarios with $M_T^1$, the next point is for scenarios with $M_T^2$, followed by those with $M_T^3$, then with $M_T^4$, then $M_T^1$, etc. In comparison to the EPP, the HOP($1, ch$) tightly controls the inclusion of FPs by choosing smaller models, at the expense of also reducing the TP count, especially when there is more uncertainty about the true model (i.e., SNR=0.25). For both prior structures, the results indicate that at low SNR levels the presence of special points has no apparent impact as the selection behavior is similar between the four models in terms of both the TP and FP. An interesting observation is that the effect of having special points on the scale is vastly magnified whenever the coefficients that assign more weight to order-one terms ($\beta^{(2)}$) are used.

6  Method comparison: Ozone data analysis

This section uses the ozone data from Breiman and Friedman (1985) which has been extensively analyzed in the literature, to assess the performance of our procedures. We follow the analysis of Liang et al. (2008), who investigated hyper g-priors. After removing observations with missing values, 330 observations remain; these include daily measurements of maximum ozone concentration near Los Angeles and eight meteorological variables. A subset of 165 observations was sampled uniformly at random and used for variable selection; the remaining data were used for validation. The eight meteorological variables, interactions and square terms are used as predictors, which results in a full model $M_F$ with 44 potential predictors. The base model $M_B$ is taken to be the intercept-only model. The model space contains approximately $71 \times 10^9$ models, and computation of all model posterior probabilities is infeasible.

The HOP, HUP, and HIP with $a = 1$ and $b = ch$ as well as the EPP are considered for comparison purposes. To obtain Bayes factors, four different mixtures of g-priors are
utilized: intrinsic priors (IP), hyper-g (HG) priors (Liang et al., 2008) with hyper-parameters $\alpha = 2, \beta = 1$ and $\alpha = \beta = 1$, and Zellner-Siow (ZS) priors (Zellner and Siow, 1980) (for details, see Section D of the Supplementary Materials). The results were extracted for the highest posterior probability (HPM) and median posterior probability (MPM) models. Additionally, the model is estimated using the R package hierNet (Bien et al., 2013), to compare model selection results with those obtained using the hierarchical LASSO (Bien et al., 2013) restricted to well-formulated models by imposing the strong heredity constraint. The procedures were assessed through their predictive accuracy on the validation dataset.

Among all models, the one that yields the smallest root mean squared error (RMSE) is the median probability model obtained using the HOP and EPP with the ZS prior, and also using the HOP with both of HG priors considered (Table 6.1). The HOP model with the intrinsic prior has all the terms contained in the lowest RMSE model, with the exception of $dpg^2$, which has a relatively high marginal inclusion probability of 46%. This disparity between the IP and other mixtures of $g$-priors is explained by the fact that the IP induces less posterior shrinkage than the ZS and HG priors. The MPMs obtained through the HUP and HIP are nested in the best model, suggesting that these model space priors penalize complexity too much and result in false negatives. Consideration of these MPMs suggests that the HOP is best at producing true positives while controlling for false positives.

The highest probability models, regardless of the parameter prior used, coincide for each of the model priors considered. For the HIP, HOP, and HUP, these models include 5 of the terms contained in the best MPM. The HPM obtained from the EPP nests the best MPM model, adding the term $hum \ast dpg$ whose posterior probability is between 1% and 16% under HIP, HOP, and HUP and most likely represents a false positive (see Section E of the Supplementary Materials). Finally, the model obtained from the hierarchical LASSO (HN) is the largest model and produces the second to largest RMSE. All of the terms contained in any of the other models, except for $vh$, are nested within the hierarchical LASSO model, and many of the terms which are exclusive to this model, receive extremely low marginal inclusion probabilities under any of the model priors and parameter priors considered under Bayesian model selection.
<table>
<thead>
<tr>
<th>Type</th>
<th>BF</th>
<th>Prior</th>
<th>Model</th>
<th>R^2</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPM</td>
<td>mix-g</td>
<td>EPP</td>
<td>hum, dpg, ibt, hum^2, hum * dpg, hum * ibt, dpg^2, ibt</td>
<td>0.8054</td>
<td>4.2739</td>
</tr>
<tr>
<td></td>
<td>mix-g</td>
<td>HIP</td>
<td>hum, ibt, hum^2, hum * ibt, ibt^2</td>
<td>0.7740</td>
<td>4.3396</td>
</tr>
<tr>
<td></td>
<td>mix-g</td>
<td>HOP</td>
<td>hum, ibt, hum^2, hum * ibt, ibt^2</td>
<td>0.7740</td>
<td>4.3396</td>
</tr>
<tr>
<td></td>
<td>mix-g</td>
<td>HUP</td>
<td>hum, dpg, ibt, hum * ibt, ibt^2</td>
<td>0.7767</td>
<td>4.3508</td>
</tr>
<tr>
<td>MPM</td>
<td>IP</td>
<td>EPP</td>
<td>hum, dpg, ibt, hum^2, hum * dpg, hum * ibt, dpg^2, ibt</td>
<td>0.8054</td>
<td>4.2739</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>HIP</td>
<td>hum, ibt, hum^2, hum * ibt, ibt^2</td>
<td>0.7740</td>
<td>4.3396</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>HOP</td>
<td>hum, dpg, ibt, hum^2, hum * ibt, ibt^2</td>
<td>0.7848</td>
<td>4.3175</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>HUP</td>
<td>hum, dpg, ibt, hum * ibt, ibt^2</td>
<td>0.7767</td>
<td>4.3508</td>
</tr>
<tr>
<td></td>
<td>ZS</td>
<td>EPP</td>
<td>hum, dpg, ibt, hum^2, hum * ibt, dpg^2, ibt^2</td>
<td>0.7896</td>
<td>4.2518</td>
</tr>
<tr>
<td></td>
<td>ZS</td>
<td>HIP</td>
<td>hum, ibt, hum * ibt, ibt^2</td>
<td>0.7525</td>
<td>4.3505</td>
</tr>
<tr>
<td></td>
<td>ZS</td>
<td>HOP</td>
<td>hum, dpg, ibt, hum^2, hum * ibt, dpg^2, ibt^2</td>
<td>0.7896</td>
<td>4.2518</td>
</tr>
<tr>
<td></td>
<td>ZS</td>
<td>HUP</td>
<td>hum, dpg, ibt, hum * ibt, ibt^2</td>
<td>0.7767</td>
<td>4.3508</td>
</tr>
<tr>
<td></td>
<td>HG_{11}</td>
<td>EPP</td>
<td>vh, hum, dpg, ibt, hum^2, hum * ibt, dpg^2</td>
<td>0.7701</td>
<td>4.3049</td>
</tr>
<tr>
<td></td>
<td>HG_{11}</td>
<td>HIP</td>
<td>hum, ibt, hum * ibt, ibt^2</td>
<td>0.7525</td>
<td>4.3505</td>
</tr>
<tr>
<td></td>
<td>HG_{11}</td>
<td>HOP</td>
<td>hum, dpg, ibt, hum^2, hum * ibt, dpg^2, ibt^2</td>
<td>0.7896</td>
<td>4.2518</td>
</tr>
<tr>
<td></td>
<td>HG_{11}</td>
<td>HUP</td>
<td>hum, dpg, ibt, hum * ibt, ibt^2</td>
<td>0.7767</td>
<td>4.3508</td>
</tr>
<tr>
<td></td>
<td>HG_{21}</td>
<td>EPP</td>
<td>hum, dpg, ibt, hum^2, hum * ibt, dpg^2</td>
<td>0.7701</td>
<td>4.3037</td>
</tr>
<tr>
<td></td>
<td>HG_{21}</td>
<td>HIP</td>
<td>hum, dpg, ibt, hum * ibt, ibt^2</td>
<td>0.7767</td>
<td>4.3508</td>
</tr>
<tr>
<td></td>
<td>HG_{21}</td>
<td>HOP</td>
<td>hum, dpg, ibt, hum^2, hum * ibt, dpg^2, ibt^2</td>
<td>0.7896</td>
<td>4.2518</td>
</tr>
<tr>
<td></td>
<td>HG_{21}</td>
<td>HUP</td>
<td>hum, dpg, ibt, hum * ibt</td>
<td>0.7526</td>
<td>4.4036</td>
</tr>
<tr>
<td></td>
<td>HierNet</td>
<td></td>
<td>hum, temp, ibh, dpg, ibt, vis, hum^2, hum * ibt, temp^2, temp * ibt, dpg^2</td>
<td>0.7651</td>
<td>4.3680</td>
</tr>
</tbody>
</table>

Table 6.1: Highest probability (HPM) and median probability (MPM) models for different mixtures of g-priors for parameters and model priors. mix-g: all mixtures of g-priors, IP: intrinsic prior, ZN: Zellner-Siow prior, HG_{11}: hyper-g prior \( \alpha = \beta = 1 \), HG_{21}: hyper-g prior \( \alpha = 2, \beta = 1 \). EPP: Equal model probability prior, HIP: hierarchical independence prior, HOP: hierarchical order prior, HUP: hierarchical uniform prior.

### 7 Discussion

In this paper, we investigated prior structures for well-formulated models and developed random walk samplers to draw from the model space. The key to understanding prior distributions on the space of WFM s is the hierarchical nature of the model space itself. The prior distributions described in this paper take advantage of that hierarchy in two ways. First,
conditional independence and immediate inheritance are used to develop the HOP, HIP, and HUP structures discussed in Section 3. Second, the conditional nature of the priors allows for the direct incorporation of complexity penalizations. Of the priors proposed, the HOP using the hyperparameter choice \((1, ch)\) provides the best control of false positives while maintaining a reasonable true positive rate. Thus, this prior is recommended as the default prior on the space of WFM.

Software necessary to carry out a Metropolis-Hastings random walk on the space of WFM is provided in the R package \texttt{varSelectIP}. It implements a number of local priors for the regression coefficients, including the intrinsic prior, Zellner-Siow prior, and hyper \(g\)-priors used in this paper. In addition, the software supports the computation of credible sets for each regression coefficient conditioned on the selected model as well as under model averaging.

\textbf{References}


A Posterior sensitivity simulation details

To determine how the proposed priors are adjusting the posterior probabilities to account for multiplicity, a simple simulation was performed. The goal of this exercise was to understand how the posterior analysis is affected by the priors proposed in response to increasing complexity. First, the priors are compared as the number of main effects $p$ grows. Second, they are compared as the depth of the hierarchy increases, or in other words, as the order $J^M_{\text{max}}$, increases.

The quality of a node is characterized by its marginal posterior inclusion probabilities $p_\alpha = \sum_{M \in \mathcal{M}} I(\alpha \in M)p(M|y, \mathcal{M})$ for $\alpha \in \mathcal{M}_F$. These posteriors were obtained for the proposed priors as well as the equal probability prior (EPP) on $\mathcal{M}$. For all prior structures, both the default hyper-parameters $a = b = 1$ and the penalizing choice of $a = 1$ and $b = c h$ were considered. The results, for the different combinations of $M_F$ and $M_T$ incorporated in the analysis, were obtained from 100 random replications (i.e., generating at random 100 matrices of main effects and responses). The simulation proceeds as follows:

1. Randomly generate main effects matrices $X = (x_1, \ldots, x_{18})$ for $x_i \sim \mathcal{N}(0, I_n)$ and error vectors $\epsilon \sim \mathcal{N}(0, I_n)$, for $n = 60$.

2. Setting all coefficient values equal to one, calculate $y = Z M_T \beta + \epsilon$ for the true models given by:

   - $M_{T1} = x_1, x_2, x_3, x_1^2, x_1 x_2, x_2^2, x_2 x_3$ with $|M_{T1}| = 7$
   - $M_{T2} = x_1, x_2, \ldots, x_{16}$ with $|M_{T2}| = 16$
   - $M_{T3} = x_1, x_2, x_3, x_4$ with $|M_{T3}| = 4$
   - $M_{T4} = x_1, x_2, \ldots, x_8, x_1^2, x_3 x_4$ with $|M_{T4}| = 10$
   - $M_{T5} = x_1, x_2, x_3, x_4, x_1^2, x_3 x_4$ with $|M_{T5}| = 6$

3. In all simulations, the base model $M_B$ is the intercept-only model. The notation $(x_1 + \ldots + x_p)^d$ is used to represent the full order-$d$ polynomial response surface in $p$ main effects. The model spaces, characterized by their corresponding full model $M_F$ are presented in Table A.1, as well as the true models used in each case.

4. Enumerate the model spaces and calculate $p(M|y, \mathcal{M})$ for all $M \in \mathcal{M}$ using the EPP, HUP, HIP, and HOP, the latter two each with the two sets of hyper-parameters.

5. Count the number of true positives and false positives in each $\mathcal{M}$ for the different priors.
The true positives (TP) are defined as those nodes $\alpha \in M_T$ such that $p_\alpha > 0.5$. With the false positives (FP), three different cutoffs are considered for $p_\alpha$, elucidating the adjustment for multiplicity induced by the model priors. These cutoffs are 0.10, 0.20, and 0.50 for $\alpha \notin M_T$. The results from this exercise provide insight about the influence of the prior on the marginal posterior inclusion probabilities. In Table A.1, the model spaces considered are described in terms of the number of models they contain, and in terms of the number of nodes of $M_F$, the full model that defines the DAG for $\mathcal{M}$.

### Growing the number of main effects, holding the polynomial degree fixed

This simulation investigates the posterior behavior as the number of covariates grows for a polynomial surface of degree two. The true model is assumed to be $M_{T1}$ and has 7 polynomial terms. The false positive and true positive rates are displayed in Table A.2.

First, focus on the posterior when $(a, b) = (1, 1)$. As $p$ increases and the cutoff is low, the number of false positives increases for the EPP as well as the hierarchical priors, though less dramatically for the latter. All of the priors identify all of the true positives. The false positive rate for the 50% cutoff is less than one for all four prior structures, with the HIP exhibiting the smallest false positive rate.

With the second choice of hyper-parameters, $(1, c h)$, the improvement of the hierarchical priors over the EPP is dramatic, and the difference in performance is more pronounced as $p$ increases. These also considerably outperform the priors using the default hyper-parameters, $a = b = 1$ in terms of the false positives. Regarding the number true positives, all priors discovered the 7 true predictors in $M_{T1}$ for most of the 100 random samples of data, with only minor differences observed between either of the priors considered. That being said, the means for the priors with $a = 1, b = c h$ are slightly lower for the true positives. With a 50% cutoff, the hierarchical priors keep a tight control on the number of false positives but in doing so discard true positives with slightly higher frequency.

| $M_F$ | $|M_F|$ | $|\mathcal{M}|$ | $M_T$ used |
|-------|--------|-------------|------------|
| $(x_1 + x_2 + x_3)^2$ | 9 | 95 | $M_{T1}$ |
| $(x_1 + \ldots + x_4)^2$ | 14 | 1,337 | $M_{T1}$ |
| $(x_1 + \ldots + x_5)^2$ | 20 | 38,619 | $M_{T1}$ |

| $M_F$ | $|M_F|$ | $|\mathcal{M}|$ | $M_T$ used |
|-------|--------|-------------|------------|
| $(x_1 + x_2 + x_3)^2$ | 9 | 95 | $M_{T1}$ |
| $(x_1 + x_2 + x_3)^3$ | 19 | 2,497 | $M_{T1}$ |
| $(x_1 + x_2 + x_3)^4$ | 34 | 161,421 | $M_{T1}$ |

| $M_F$ | $|M_F|$ | $|\mathcal{M}|$ | $M_T$ used |
|-------|--------|-------------|------------|
| $x_1 + x_2 + \ldots + x_{18}$ | 18 | 262,144 | $M_{T2}, M_{T3}$ |
| $(x_1 + x_2 \ldots + x_4)^2 + x_5 + x_6 + \ldots + x_{10}$ | 20 | 85,568 | $M_{T4}, M_{T5}$ |

Table A.1: Characterization of the full models $M_F$ and corresponding model spaces $\mathcal{M}$ considered in simulations.
Table A.2: Mean number of false and true positives in 100 randomly generated datasets as the number of main effects increases.

| Cutoff     | $|M_T|$ | $M_F$     | EPP | $a = b = 1$ | $a = b = ch$ |
|------------|-------|-----------|-----|-------------|--------------|
|            |       |           |     | HIP | HUP | HOP | HIP | HUP | HOP |
| FP (>0.10) | 1.78  | 1.78      | 2.00| 2.00| 0.11 | 1.31 | 1.06|  
| FP (>0.20) | 0.43  | 0.43      | 2.00| 1.98| 0.01 | 0.28 | 0.24|  
| FP (>0.50) | 0.04  | 0.04      | 0.97| 0.36| 0.00 | 0.03 | 0.02|  
| TP (>0.50) | 7.00  | 7.00      | 7.00| 7.00| 6.97 | 6.99 | 6.99|  
| FP (>0.10) | 3.62  | 1.94      | 2.33| 2.45| 0.10 | 0.63 | 1.07|  
| FP (>0.20) | 1.60  | 0.47      | 2.17| 2.15| 0.01 | 0.17 | 0.24|  
| FP (>0.50) | 0.25  | 0.06      | 0.35| 0.36| 0.00 | 0.02 | 0.02|  
| TP (>0.50) | 7.00  | 7.00      | 7.00| 7.00| 6.97 | 6.99 | 6.99|  
| FP (>0.10) | 6.00  | 2.16      | 2.60| 2.55| 0.12 | 0.43 | 1.15|  
| FP (>0.20) | 2.91  | 0.55      | 2.13| 2.18| 0.02 | 0.19 | 0.27|  
| FP (>0.50) | 0.66  | 0.11      | 0.25| 0.37| 0.00 | 0.03 | 0.01|  
| TP (>0.50) | 7.00  | 7.00      | 7.00| 7.00| 6.97 | 6.99 | 6.99|  

Growing polynomial degree, holding the number of main effects fixed

For these examples, the true model is once again $M_{T1}$. When the complexity is increased by making the order of $M_F$ larger (see Table A.3), the inability of the EPP to adjust the inclusion posteriors for multiplicity becomes more pronounced; the EPP becomes less and less efficient at removing false positives when the FP cutoff is low. Among the priors with $a = b = 1$, as the order increases, the HIP does the best job at filtering out the false positives. Using the 0.5 false positive cutoff some false positives are included, both for the EPP and for all the priors with $a = b = 1$, indicating that the default hyper-parameters might not be the best option to control FP. The seven covariates in the true model all obtain a high inclusion posterior probability both with the EPP and the $a = b = 1$ priors.

In contrast, any of the $a = 1$ and $b = ch$ priors dramatically improve upon their $a = b = 1$ counterparts, consistently assigning low inclusion probabilities for the majority of the false positive terms, even for low cutoffs. As the order of the polynomial surface increases the difference in performance between these priors and either the EPP or their default versions becomes even more clear. At the 50% cutoff, the hierarchical priors with complexity penalization exhibit very low false positive rates. The true positive rate decreases slightly for the priors, but not to an alarming degree.

Other model spaces

This part of the analysis considers model spaces that do not correspond to full polynomial degree response surfaces (see Table A.4). The first example is a model space with main
effects only. The second example includes a full quadratic surface of order 2, but in addition includes six terms for which only main effects are to be modeled. Two true models are used in combination with each model space, to observe how the posterior probabilities vary under the influence of the different priors for “large” and “small” true models.

By construction, in model spaces with main effects only HIP(1, 1) and EPP are equivalent, as are HOP(a, b) and HUP(a, b). This accounts for the similarities observed among the results for the first two cases presented in Table A.4, where the model space corresponds to a full model with 18 main effects, and the true models are a model with 16 and 4 main effects, respectively. When the number of true coefficients is large, the HUP(1, 1) and HOP(1, 1) do poorly at controlling false positives, even at the 50% cutoff. In contrast, the HIP (and thus the EPP), with the 50% cutoff identifies the true positives and no false positives. This result, however, does not imply that the EPP is better at controlling false positives, it has to be considered in light of the size of the true model which contains 16 out of the 18 nodes in MF. The a = 1 and b = ch priors show dramatically different behavior. The HIP controls the number of false positives well, but fails to identify the true coefficients at the 50% cutoff. In contrast, the HOP identifies all of the true positives and has a small false positive rate for the 50% cutoff.

If the number of true positives is small, most terms in the full model are truly zero. The EPP includes least one false positive in approximately 50% of the randomly sampled datasets. On the other hand, the HUP(1, 1) provides some control for multiplicity, obtaining on average a lower number of false positives than the EPP. Furthermore, the proposed hierarchical priors with a = 1, b = ch are substantially better than the EPP (and the choice of a = b = 1) at controlling false positives and capturing all true positives using the marginal posterior inclusion probabilities. The two examples suggest that the HOP(1, ch) is the best default
choice for model selection when the number of terms available at a given degree is large.

The third and fourth examples in Table A.4 consider the same irregular model space, with data generated from $M_{T4}$ with ten terms and $M_{T5}$ with six terms. HIP(1, 1) and EPP again behave quite similarly, incorporating a large number of false positives for the 0.1 cutoff; at 0.5 cutoff some false positives are still included. The HUP(1, 1) and HOP(1, 1) behave similarly with a slightly higher false positive rate at the 50% cutoff. In terms of the true positives, the EPP and $a = b = 1$ priors always include all the coefficients in $M_{T4}$ and $M_{T5}$, respectively. On the other hand, the ability of the $a = 1, b = ch$ priors to control for false positives is markedly better than that of the EPP and the hierarchical priors with choice of $a = 1 = b = 1$. At the 50% cutoff, these priors identify all of the true positives and true negatives. Once again, these examples point to the hierarchical priors with additional penalization for complexity as being good default priors on the model space.

| Cutoff  | $|M_T|$ | $M_F$ | EPP | $a = 1, b = 1$ | $a = 1, b = ch$ |
|---------|--------|-------|-----|---------------|----------------|
|         |        |       | HIP | HUP | HOP | HIP | HUP | HOP |
| FP($>0.10$) | 16 | $x_1 + x_2 + \ldots + x_{18}$ | 1.93 | 1.93 | 2.00 | 2.00 | 0.03 | 1.80 | 1.80 |
| FP($>0.20$) | | | 0.52 | 0.52 | 2.00 | 2.00 | 0.01 | 0.46 | 0.46 |
| FP($>0.50$) | | | 0.07 | 0.07 | 2.00 | 2.00 | 0.01 | 0.04 | 0.04 |
| TP($>0.50$) | $(M_{T2})$ | | 15.99 | 15.99 | 16.00 | 16.00 | 0.01 | 0.04 | 0.04 |
| FP($>0.10$) | 13.95 | 13.95 | 9.15 | 9.15 | 0.26 | 1.31 | 1.31 |
| FP($>0.20$) | 5.45 | 5.45 | 3.03 | 3.03 | 0.05 | 0.45 | 0.45 |
| FP($>0.50$) | 0.84 | 0.84 | 0.45 | 0.45 | 0.02 | 0.06 | 0.06 |
| TP($>0.50$) | $(M_{T3})$ | | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 |
| FP($>0.10$) | 9.73 | 9.71 | 10.00 | 5.60 | 0.34 | 2.33 | 2.20 |
| FP($>0.20$) | 2.65 | 2.65 | 8.73 | 3.05 | 0.12 | 0.74 | 0.69 |
| FP($>0.50$) | 0.35 | 0.35 | 1.36 | 1.68 | 0.02 | 0.11 | 0.12 |
| TP($>0.50$) | $(M_{T4})$ | | 10.00 | 10.00 | 10.00 | 9.99 | 9.94 | 9.98 | 9.99 |
| FP($>0.10$) | 13.52 | 13.52 | 11.06 | 9.94 | 0.44 | 1.63 | 1.96 |
| FP($>0.20$) | 4.22 | 4.21 | 3.60 | 5.01 | 0.15 | 0.48 | 0.68 |
| FP($>0.50$) | 0.53 | 0.53 | 0.57 | 0.75 | 0.01 | 0.08 | 0.11 |
| TP($>0.50$) | $(M_{T5})$ | | 6.00 | 6.00 | 6.00 | 6.00 | 5.99 | 5.99 | 5.99 |

Table A.4: Mean number of false and true positives in 100 randomly generated datasets with an unstructured or irregular model space.

### B Random walk algorithms

#### Global Jump

From the current state $M$, the global jump is performed by drawing a model $M'$ at random from the model space. This is achieved by beginning at the base model and increasing the order from $J_{M}^{\min}$ to the $J_{M}^{\max}$, the minimum and maximum orders of nodes in $\Upsilon(M_F) = M_F \setminus M_B$. At each order a set of nodes is selected at random from the prior.
conditioned on the nodes already in the model. The MH correction is

\[
\alpha = \left\{ 1, \frac{m(y|M', \mathcal{M})}{m(y|M, \mathcal{M})} \right\}
\]

**Local Jump** From the current state \(M\), the local jump is performed by drawing a model from the set of models \(L(M) = \{M_\alpha : \alpha \in \mathcal{E}(M) \cup \mathcal{C}(M)\}\) where \(M_\alpha\) is \(M \setminus \alpha\) for \(\alpha \in \mathcal{E}(M)\) and \(M \cup \{\alpha\}\) for \(\alpha \in \mathcal{C}(M)\). The proposal probabilities for the model are computed as a mixture of \(p(M'|y, \mathcal{M}, M' \in L(M))\) and the discrete uniform distribution. The proposal kernel is

\[
q(M'|y, \mathcal{M}, M' \in L(M)) = \frac{1}{2} \left( p(M'|y, \mathcal{M}, M' \in L(M)) + \frac{1}{|L(M)|} \right)
\]

This choice promotes moving to better models while maintaining a non-negligible probability of moving to any of the possible models. The MH correction is

\[
\alpha = \left\{ 1, \frac{m(y|M', \mathcal{M}) q(M|y, \mathcal{M}, M \in L(M'))}{m(y|M, \mathcal{M}) q(M'|y, \mathcal{M}, M' \in L(M'))} \right\}
\]

**Intermediate Jump** The intermediate jump is performed by increasing or decreasing the order of the nodes under consideration performing local proposals based on order. For a model \(M'\) define \(L_j(M') = \{M'\} \cup \{M_\alpha' : \alpha \in (\mathcal{E}(M') \cup \mathcal{C}(M')) \cap \Upsilon_j(M_F)\}\). From a state \(M\), the kernel chooses at random whether to increase or decrease the order. If \(M = M_F\) then decreasing the order is chosen with probability 1, and if \(M = M_B\) then increasing the order is chosen with probability 1, in all other cases the probability of increasing and decreasing order is \(\frac{1}{2}\). The proposal kernels are given by

**Increasing order proposal kernel.**

1. Set \(j = J_{\mathcal{M}}^{\text{min}} - 1\) and \(M'_j = M\).
2. Draw \(M'_{j+1}\) from \(q_{\text{inc},j+1}(M'|y, \mathcal{M}, M' \in L_{j+1}(M'_j))\), where
   \[
   q_{\text{inc},j+1}(M'|y, \mathcal{M}, M' \in L_{j+1}(M'_j)) = \frac{1}{2} \left( p(M'|y, \mathcal{M}, M' \in L_{j+1}(M'_j)) + \frac{1}{|L_{j+1}(M'_j)|} \right).
   \]
3. Set \(j = j + 1\).
4. If \(j < J_{\mathcal{M}}^{\text{max}}\) then return to 2. Otherwise proceed to 5.
5. Set \(M' = M'_{J_{\mathcal{M}}^{\text{max}}}\) and compute the proposal probability
   \[
   q_{\text{inc}}(M'|y, \mathcal{M}, M) = \prod_{j=J_{\mathcal{M}}^{\text{min}}-1}^{J_{\mathcal{M}}^{\text{max}}-1} q_{\text{inc},j+1}(M'_j|y, \mathcal{M}, M' \in L_{j+1}(M'_j))
   \]
Decreasing order proposal kernel.

1. Set \( j = J^\text{max}_\mathcal{M} + 1 \) and \( M'_j = M \).

2. Draw \( M'_{j-1} \) from \( q_{\text{dec},j-1}(M'|y, \mathcal{M}, M' \in L_{j-1}(M'_j)) \), where
\[
q_{\text{dec},j-1}(M'|y, \mathcal{M}, M' \in L_{j-1}(M'_j)) = \frac{1}{2} \left( p(M'|y, \mathcal{M}, M' \in L_{j-1}(M'_j)) + \frac{1}{|L_{j-1}(M'_j)|} \right).
\]

3. Set \( j = j - 1 \).

4. If \( j > J^\text{min}_\mathcal{M} \), then return to 2. Otherwise proceed to 5.

5. Set \( M' = M'^{\text{min}}_{J^\text{min}_\mathcal{M}} \) and compute the proposal probability
\[
q_{\text{dec}}(M'|y, \mathcal{M}, M) = \prod_{j=J^\text{max}_\mathcal{M}+1}^{J^\text{min}_\mathcal{M}+1} q_{\text{dec},j-1}(M'|y, \mathcal{M}, M' \in L_{j-1}(M'_j))
\] (10)

If increasing order is chosen, then the MH correction is given by
\[
\alpha = \min \left\{ 1, \left( 1 + \frac{I(M' = M_F)}{1 + I(M = M_B)} \right) q_{\text{dec}}(M|y, \mathcal{M}, M') \frac{p(M'|y, \mathcal{M})}{q_{\text{inc}}(M'|y, \mathcal{M}, M) \frac{p(M|y, \mathcal{M})}} \right\}
\] (11)

and similarly if decreasing order is chosen.

Other Local and Intermediate Kernels The local and intermediate kernels described here perform a kind of stochastic forwards-backwards selection. Each kernel \( q \) can be relaxed to allow more than one node to be turned on or off at each step, which could provide larger jumps for each of these kernels. The tradeoff is that number of proposed models for such jumps could be very large, precluding the use of posterior information in the construction of the proposal kernel.

C Simulation details

Briefly, the idea is to let \( Z_{M_T}(X)\beta_{M_T} = (QR)\beta_{M_T} = Q\eta_{M_T} \) (i.e. \( \beta_{M_T} = R^{-1}\eta_{M_T} \)) using the QR decomposition. As such, setting all values in \( \eta_{M_T} \) proportional to one, corresponds to distributing the signal in the model uniformly across all predictors, regardless of their order.

The (unconditional) variance of a single observation \( y_i \) is \( \text{var}(y_i) = \text{var}(E[y_i|z_i]) + E[\text{var}(y_i|z_i)] \), where \( z_i \) is the \( i \)-th row of the design matrix \( Z_{M_T} \). Hence, we take the signal to noise ratio for each observation to be
\[
\text{SNR}(\eta) = \eta^T_{M_T} R^{-T} \Sigma_z R \eta_{M_T} / \sigma^2,
\]
where \( \Sigma_z = \text{var}(z_i) \). We determine how the signal is distributed across predictors up to a proportionality constant, to be able to control simultaneously the signal to noise ratio.
Additionally, to investigate the ability of the model to capture correctly the hierarchical structure, we specify four different 0-1 vectors that determine the predictors in $M_T$, which generates the data in the different scenarios.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Values considered</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{SNR}(\eta_{M_T}) = k$</td>
<td>0.25, 1, 4</td>
</tr>
<tr>
<td>$\eta_{M_T}$</td>
<td>$(1, 1_3, 1_4, 1_2), (1, 1_3, 1_4, 1_2), (1, 1_3, 1_4, 1_2)$</td>
</tr>
<tr>
<td>$\gamma_{M_T}$</td>
<td>$(1, 1_3, 1_4, 1_2), (1, 1_3, 1_4, 0_2), (1, 1_3, 1_4, 1_2), (1, 1_3, 1_4, 1_2)$</td>
</tr>
<tr>
<td>$n$</td>
<td>130, 260, 1040</td>
</tr>
</tbody>
</table>

Table C.1: Experimental conditions

The results presented below are somewhat different from those found in the main body of the article in Section 5. These are extracted averaging the number of FP’s TP’s and model sizes, respectively, over the 100 independent runs and across the corresponding scenarios, for the 20 highest probability models.

**SNR and sample size effect**

In terms of the SNR and the sample size (Figure C.1), we observe that, as expected, small sample sizes conditioned upon a small SNR impair the ability of the algorithm to detect true coefficients with both the EPP and HOP$(1, ch)$, with this effect more notorious when using the latter prior. However, considering the mean number of true positives (TP) jointly with the mean model size, it is clear that although the sensitivity is low, most of the few predictors that are discovered belong to the true model. The results observed with SNR of 0.25 and a relatively small sample size are far from being impressive, however, real problems where the SNR is as low as 0.25 will yield many spurious associations under the EPP. The fact that the HOP$(1, ch)$ has a strong protection against false positive is commendable in itself. A SNR of 1 also represents a feeble relationship between the predictors and the response, nonetheless, the method captures approximately half of the true coefficients while including very few false positives. Following intuition, as either the sample size or the SNR increase, the algorithms performance is greatly enhanced. Either having a large sample size or a large SNR yields models that contain mostly true predictors. Additionally, HOP$(1, ch)$ provides a strong control over the number of false positives, therefore for high SNR or larger sample sizes, the number of predictors in the top 20 models is close to the size of the true model. In general, the EPP allows the detection of more TP’s while the HOP$(1, ch)$ provides a stronger control on the amount of FP’s included when considering small sample sizes combined with small SNRs. As either sample size or SNR grows the differences between the two priors become indistinct.
Coefficient magnitude

This part of the experiment explores the effect of how the signal is distributed across predictors. As mentioned before, sphering is used to assign the coefficients values in a manner that controls the amount of signal that goes into each coefficient. Three possible ways to allocate the signal are considered. First, each order-one coefficient contains twice as much signal as any order-two coefficient, and four times as much any as order-three coefficient; second, all coefficients contain the same amount of signal regardless of their order; and third, each order-one coefficient contains a half as much signal as any order-two coefficient, and a quarter of what any order-three coefficient has. In Figure C.2 these values are denoted by $\beta = c(1_{o1}, 0.5_{o2}, 0.25_{o3})$, $\beta = c(1_{o1}, 1_{o2}, 1_{o3})$ and $\beta = c(0.25_{o1}, 0.5_{o2}, 1_{o3})$, respectively.

Observe that the number of FP’s is invulnerable to how the SNR is distributed across predictors using the HOP(1, $ch$); conversely, when using the EPP the number of FP’s decreases as the SNR grows always being slightly higher than those obtained with the HOP. With either prior structure the algorithm performs better whenever all coefficients are equally weighted or when those for the order-three terms have higher weights. In these two cases
(i.e. with $\beta = c(1, 0.5, 0.25, 0.3)$ or $\beta = c(1, 1, 1, 1)$), the effect of the SNR appears to be similar. In contrast, when more weight is given to order one terms the algorithm yields slightly worse models at any SNR level. This is an intuitive result since giving more signal to higher order terms makes it easier to detect higher order terms, and consequently by strong heredity, the algorithm will also select the corresponding lower order terms included in the true model.

Figure C.2: SNR vs coefficient values: Average model size, average true positives and average false positives for all simulated scenarios by model ranking according to model posterior probabilities.

Special points on the scale

In Nelder (1998) the author argues that the conditions under which the weak-heredity principle can be used for model selection are so restrictive that the principle is commonly not valid in practice in this context. In addition, the author states that considering well-formulated models only, does not take into account the possible presence of special points on the scales of the predictors, that is, situations where omitting lower order terms is justified due to the nature of the data. However, it is our contention that every model has an underlying well-
formulated structure, whether or not some predictor has special points on its scale will be
determined through the estimation of the coefficients, once a valid well-formulated structure
has been chosen.

Figure C.3: SNR vs different true models $M_T$: Average model size, average true positives
and average false positives for all simulated scenarios by model ranking according to model
posterior probabilities.

To understand how the algorithm behaves whenever the true data generating mechanism
has zero-valued coefficients for some lower order terms in the hierarchy, four different true
models are considered. Three of them are not well-formulated, while the remaining one is
the WFM shown in Figure 5.1. The three models that have special points, correspond to
the same model $M_T$ from Figure 5.1 but have, respectively, zero-valued coefficients for all
the order-one terms, all the order-two terms, and for $x_2^2$ and $x_2x_5$.

As seen before, in comparison to the EPP, the HOP($1,ch$) tightly controls the inclusion
FPs by choosing smaller models, at the expense of also reducing the TP count, especially
when there is more uncertainty about the true model (i.e., SNR=0.25). For both prior
structures, the results in Figure C.3 indicate that at low SNR levels the presence of special
points has no apparent impact as the selection behavior is similar between the four models in terms of both the TP and FP. As the SNR increases, the TPs and the model size are affected for true models with zero-valued lower order terms. These differences, however, are not very large. Relatively smaller models are selected whenever some terms in the hierarchy are missing, but with high SNR which is where the differences are most pronounced, the predictors included are mostly true coefficients. The impact is almost imperceptible for the true model that lacks order one terms and the model with zero coefficients for $x_1^2$ and $x_2 x_5$, and is more visible for models without order two terms. This last result is expected due to strong-heredity; whenever the order-one coefficients are missing, the inclusion of order-two and order-three terms will force their selection, which is also the case when only a few order two terms have zero-valued coefficients. Conversely, when all order two predictors are removed, some order three predictors are not selected as their signal is attributed the order two terms have zero-valued coefficients. This is especially the case for the order three interaction term $x_1 x_2 x_5$, which depends on the inclusion of three order two terms $(x_1 x_2, x_1 x_5, x_2 x_5)$ in order for it to be included as well. This makes the inclusion of this term somewhat more challenging; the three order two interactions capture most of the variation of the polynomial terms that is present when the order three term is also included. However, special points on the scale commonly occur on a single or at most on a few covariates. A true data generating mechanism that removes all terms of a given order in the context of polynomial models is clearly not justified; here this was only done for comparison purposes.

D Scaled mixtures of $g$-priors

Intrinsic priors

In the Bayesian paradigm, a model $M$ in $\mathcal{M}$ is defined by a sampling density and a prior distribution. The sampling density associated with model $M$ is denoted by $f(y | \beta_M, \sigma_M^2, M)$, where $(\beta_M, \sigma_M^2)$ is a vector of model-specific unknown parameters. The prior for model $M$ and its corresponding set of parameters is $\pi(\beta_M, \sigma_M^2, M | \mathcal{M}) = \pi(\beta_M, \sigma_M^2 | M, \mathcal{M}) \cdot \pi(M | \mathcal{M})$. The priors $\pi(M | \mathcal{M})$ are developed in Section 3.

Objective local priors for the model parameters $(\beta_M, \sigma_M^2)$ are achieved through modifications and extensions of Zellner’s $g$-prior (Liang et al., 2008). An automatic choice of an objective prior is the intrinsic prior (Berger and Pericchi, 1996; Moreno et al., 1998). Because $M_B \subseteq M$ for all $M \in \mathcal{M}$, the intrinsic prior for $(\beta_M, \sigma_M^2)$ is defined as an expected posterior prior

$$\pi^I(\beta_M, \sigma_M^2 | M) = \int p^R(\beta_M, \sigma_M^2 | \tilde{y}, M) m^R(\tilde{y} | M_B) d\tilde{y},$$  \hspace{1cm} (12)$$

where $\tilde{y}$ is a minimal training sample for model $M$, $I$ denotes the intrinsic distributions and $R$ denotes distributions derived from the reference prior $\pi^R(\beta_M, \sigma_M^2 | M) = c_M \frac{d\beta_M d\sigma_M^2}{\sigma_M^2}$. In (12), $m^R(\tilde{y} | M) = \int \int f(\tilde{y} | \beta_M, \sigma_M^2, M) \pi^R(\beta_M, \sigma_M^2 | M) d\beta_M d\sigma_M^2$ is the reference marginal of $\tilde{y}$ under model $M$ and $p^R(\beta_M, \sigma_M^2 | \tilde{y}, M) = \frac{f(\tilde{y} | \beta_M, \sigma_M^2, M) \pi^R(\beta_M, \sigma_M^2 | M)}{m^R(\tilde{y} | M)}$ is the reference posterior density.
In the regression framework, the reference marginal $m^R$ is improper and produces improper intrinsic priors. However, the intrinsic Bayes factor of model $M$ to the base model $M_B$ is well-defined and given by

$$BF^I_{M,M_B}(y) = (1 - R^2_M)^{-\frac{n - |M_B|}{2}} \times$$

$$\int_0^1 \left( \frac{n + \sin^2(\frac{\pi \theta}{2}) \cdot (|M| + 1)}{n + \frac{\sin^2(\frac{\pi \theta}{2}) \cdot (|M| + 1)}{1 - R^2_M^M}} \right)^{\frac{n - |M|}{2}} \left( \frac{\sin^2(\frac{\pi \theta}{2}) \cdot (|M| + 1)}{n + \frac{\sin^2(\frac{\pi \theta}{2}) \cdot (|M| + 1)}{1 - R^2_M^M}} \right)^{\frac{|M| - |M_B|}{2}} d\theta,$$  \hspace{1cm} (13)

where $R^2_M$ is the coefficient of determination of model $M$ versus model $M_B$. The Bayes factor between two models $M$ and $M'$ is defined as $BF^I_{M,M'}(y) = BF^I_{M,M_B}(y)/BF^I_{M',M_B}(y)$. The “goodness” of the model $M$, based on the intrinsic priors, is given by its posterior probability

$$p^I(M|y,M) = \frac{BF^I_{M,M_B}(y) \pi(M|M)}{\sum_{M' \in \mathcal{M}} BF^I_{M',M_B}(y) \pi(M'|\mathcal{M})}.$$  \hspace{1cm} (14)

It has been shown that the system of intrinsic priors produces consistent model selection (Casella et al., 2009; Girón et al., 2010). In the context of well-formulated models, the true model $M_T$ is the smallest well-formulated model $M \in \mathcal{M}$ such that $\alpha \in M$ if $\beta_\alpha \neq 0$. If $M_T$ is the true model, then the posterior probability of model $M_T$, based on equation (14), converges to 1.

**Other mixtures of g-priors**

Scaled mixtures of g-priors place reference prior on $(\beta_{M_B}, \sigma^2)$ and a multivariate normal distribution on $\beta$ in $M \setminus M_B$, which is normal with mean $0$ and precision matrix

$$q_M \frac{\mathbf{Z}'_M (\mathbf{I} - \mathbf{H}_0) \mathbf{Z}_M}{n \sigma^2},$$

where $H_0$ is the hat matrix associated with $Z_{MB}$. The prior is completed by a prior on $w$ and choice of scaling $q_M$, which is set at $|M| + 1$ to account for the minimal sample size of $M$. Under these assumptions the Bayes’ factor for $M$ to $M_B$ is given by

$$BF_{M,M_B}(y) = (1 - R^2_M)^{-\frac{n - |M_B|}{2}} \int \left( \frac{n + w(|M| + 1)}{n + \frac{w(|M| + 1)}{1 - R^2_M^M}} \right)^{\frac{n - |M|}{2}} \frac{w(|M| + 1)}{n + \frac{w(|M| + 1)}{1 - R^2_M^M}}^{\frac{|M| - |M_B|}{2}} \pi(w) dw$$

We consider the following priors on $w$. The intrinsic prior is $\pi(w) = Beta(w; 0.5, 0.5)$ which is only defined for $w \in (0, 1)$. A version of the Zellner-Siow prior is given by $w \sim Gamma(0.5, 0.5)$, which produces a multivariate Cauchy distribution on $\beta$. A family of hyper-g priors are defined by $\pi(w) \propto w^{-1/2}(\beta + w)^{-(\alpha + 1)/2}$, which have Cauchy-like tails but produce more shrinkage than the Cauchy prior.
E Supplementary information for the ozone data analysis

The covariates considered for the ozone data analysis match those used in Liang et al. (2008); these are displayed in table E.1 below.

Table E.1: Variables used in the analyses of the ozone contamination dataset

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ozone</td>
<td>Daily max 1hr-average ozone (ppm) at Upland, CA.</td>
</tr>
<tr>
<td>vh</td>
<td>500 millibar pressure height (m) at Vandenberg AFB.</td>
</tr>
<tr>
<td>wind</td>
<td>Wind speed (mph) at LAX.</td>
</tr>
<tr>
<td>hum</td>
<td>Humidity (%) at LAX.</td>
</tr>
<tr>
<td>temp</td>
<td>Temperature (F) measured at Sandburg, CA.</td>
</tr>
<tr>
<td>ibh</td>
<td>Inversion base height (ft) at LAX.</td>
</tr>
<tr>
<td>dpg</td>
<td>Pressure gradient (mm Hg) from LAX to Daggett, CA.</td>
</tr>
<tr>
<td>vis</td>
<td>Visibility (miles) measured at LAX.</td>
</tr>
<tr>
<td>ibt</td>
<td>Inversion base temperature (F) at LAX.</td>
</tr>
</tbody>
</table>

The marginal posterior inclusion probability corresponds to the probability of including a given term in the full model $M_F$ after summing over all models in the model space. For each node $\alpha \in M_F$, this probability is given by $p_\alpha = \sum_{M \in \mathcal{M}} \mathbb{I} (\alpha \in M) p(M|y, \mathcal{M})$. Given that in problems with a large model space, such as the one considered for the ozone concentration problem, enumeration of the entire space is not feasible. Thus, these probabilities are estimated summing over every model drawn by the random walk over the model space $\mathcal{M}$.

Given that there are in total 44 potential predictors, for convenience, in table E.2 below, we only display the marginal posterior probabilities for the terms included under at least one of the model priors considered (EPP, HIP, HUP, and HOP) for each of the parameter priors utilized (intrinsic priors, Zellner-Siow priors, Hyper-g(1,1), and Hyper-g(2,1))
### Table E.2: Marginal inclusion probabilities under each model prior of terms in at least one median model, using the different parameter priors: Intrinsic Priors (top left), Zellner-Siow (top right), Hyper-g(1,1) (bottom left), Hyper-g(2,1) (bottom right)

<table>
<thead>
<tr>
<th>Term</th>
<th>EPP</th>
<th>HIP</th>
<th>HUP</th>
<th>HOP</th>
</tr>
</thead>
<tbody>
<tr>
<td>hum</td>
<td>0.99</td>
<td>0.69</td>
<td>0.85</td>
<td>0.76</td>
</tr>
<tr>
<td>dpg</td>
<td>0.85</td>
<td>0.48</td>
<td>0.52</td>
<td>0.53</td>
</tr>
<tr>
<td>ibt</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>hum&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.76</td>
<td>0.51</td>
<td>0.43</td>
<td>0.62</td>
</tr>
<tr>
<td>hum*dp&lt;sup&gt;g&lt;/sup&gt;</td>
<td>0.55</td>
<td>0.02</td>
<td>0.03</td>
<td>0.17</td>
</tr>
<tr>
<td>hum*ib&lt;sup&gt;t&lt;/sup&gt;</td>
<td>0.98</td>
<td>0.69</td>
<td>0.84</td>
<td>0.75</td>
</tr>
<tr>
<td>dp&lt;sup&gt;g&lt;/sup&gt;&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.72</td>
<td>0.36</td>
<td>0.25</td>
<td>0.46</td>
</tr>
<tr>
<td>ib&lt;sup&gt;t&lt;/sup&gt;&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.59</td>
<td>0.78</td>
<td>0.57</td>
<td>0.81</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Term</th>
<th>EPP</th>
<th>HIP</th>
<th>HUP</th>
<th>HOP</th>
</tr>
</thead>
<tbody>
<tr>
<td>vh</td>
<td>0.54</td>
<td>0.05</td>
<td>0.10</td>
<td>0.11</td>
</tr>
<tr>
<td>hum</td>
<td>0.81</td>
<td>0.67</td>
<td>0.80</td>
<td>0.69</td>
</tr>
<tr>
<td>dpg</td>
<td>0.90</td>
<td>0.50</td>
<td>0.55</td>
<td>0.58</td>
</tr>
<tr>
<td>ibt</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>hum&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.61</td>
<td>0.49</td>
<td>0.40</td>
<td>0.57</td>
</tr>
<tr>
<td>hum*ib&lt;sup&gt;t&lt;/sup&gt;</td>
<td>0.78</td>
<td>0.66</td>
<td>0.78</td>
<td>0.68</td>
</tr>
<tr>
<td>dp&lt;sup&gt;g&lt;/sup&gt;&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.83</td>
<td>0.38</td>
<td>0.30</td>
<td>0.51</td>
</tr>
<tr>
<td>ib&lt;sup&gt;t&lt;/sup&gt;&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.49</td>
<td>0.76</td>
<td>0.54</td>
<td>0.77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Term</th>
<th>EPP</th>
<th>HIP</th>
<th>HUP</th>
<th>HOP</th>
</tr>
</thead>
<tbody>
<tr>
<td>vh</td>
<td>0.54</td>
<td>0.05</td>
<td>0.10</td>
<td>0.11</td>
</tr>
<tr>
<td>hum</td>
<td>0.81</td>
<td>0.67</td>
<td>0.80</td>
<td>0.69</td>
</tr>
<tr>
<td>dpg</td>
<td>0.90</td>
<td>0.50</td>
<td>0.55</td>
<td>0.58</td>
</tr>
<tr>
<td>ibt</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>hum&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.61</td>
<td>0.49</td>
<td>0.40</td>
<td>0.57</td>
</tr>
<tr>
<td>hum*ib&lt;sup&gt;t&lt;/sup&gt;</td>
<td>0.78</td>
<td>0.66</td>
<td>0.78</td>
<td>0.68</td>
</tr>
<tr>
<td>dp&lt;sup&gt;g&lt;/sup&gt;&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.83</td>
<td>0.38</td>
<td>0.30</td>
<td>0.51</td>
</tr>
<tr>
<td>ib&lt;sup&gt;t&lt;/sup&gt;&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.49</td>
<td>0.76</td>
<td>0.54</td>
<td>0.77</td>
</tr>
</tbody>
</table>

References


