Benchmark Priors Revisited: On Adaptive Shrinkage and the Supermodel Effect in Bayesian Model Averaging*

Martin Feldkircher†
Oesterreichische Nationalbank

Stefan Zeugner‡
ECARES, Université libre de Bruxelles

May 16, 2012

Abstract

Predominant in the Bayesian Model Averaging literature, default prior choices fixing Zellner's g tend to concentrate posterior mass on a tiny set of models. The paper demonstrates this *supermodel effect* and proposes to address it by a *hyper-g prior*, whose data-dependent shrinkage adapts posterior model distributions to data quality.

This paper complements existing work by demonstrating an asymptotically consistent specification of the hyper-g prior, and its interpretation as a goodness-of-fit indicator. Moreover, this paper highlights the similarities between hyper-g and 'Empirical Bayes priors', and introduces closed-form hyper-g expressions that are essential for computationally feasible, fully Bayesian analysis. A simulation experiment illustrates the robustness of the hyper-g prior versus popular fixed prior settings. Finally, the merits of the hyper-g prior both in terms of predictive performance and stability of posterior results are demonstrated using a prominent growth data set with four different vintages of Penn World Table income data.

Keywords: Bayesian model averaging, hyper-g prior, shrinkage factor, Zellner's g prior, model uncertainty.

JEL Classifications: C11, C15, C21, C52, O50.

*The opinions in this paper are those of the authors and do not necessarily coincide with those of the Oesterreichische Nationalbank. We would like to thank Aart Kraay, Jesús Crespo Cuaresma, Gernot Doppelhofer, Domenico Giannone, Robert Kollmann, and Eduardo Ley for helpful comments.

†Oesterreichische Nationalbank, Otto-Wagner-Platz 3, 1090 Vienna, Austria. E-mail: martin.feldkircher@oenb.at

‡ULB, CP 139, 44 Avenue Jeanne, 1050 Brussels, Belgium. E-mail: stefan.zeugner@ulb.ac.be

1 Introduction

Statistical inference that neglects model uncertainty leads to overstated confidence in statistical estimates, as has been amply demonstrated since the seminal contributions by Raftery (1995) and Hoeting et al. (1999). *Bayesian Model Averaging* (BMA) tackles such model uncertainty directly by basing inference on a weighted average of all potential covariate combinations, or 'models'. In a Bayesian setting, these weights arise naturally as posterior model probabilities that correspond to the classical likelihood concept. Relying on this framework, numerous authors (e.g., Raftery, 1995; Fernández et al., 2001a; Liang et al., 2008) have demonstrated that BMA outperforms other strategies in terms of predictive ability. Virtually all of them have so far concentrated on linear models with model-specific inference based on the 'Normal-Gamma' coefficient prior with *Zellner's g* (Zellner, 1986). This prior structure has proven popular in BMA, since it leads to simple closed-form expressions of posterior statistics and because it reduces prior elicitation to the choice of a single scalar hyperparameter g. This *shrinkage* parameter determines how far a model's coefficients are shrunk toward zero: High values for g are meant to embody weak prior knowledge and correspond to model estimators that are close to least squares results. In contrast, low values for g imply less reliance on the data and posterior coefficients closer to their prior values (zero). Crucially, the parameter g shapes the weights of models in BMA. The exact specification of g is subject to intense debate,[1] but the use of a constant hyperparameter g as such has been less frequently criticized.

This paper demonstrates that the practical advantages of Zellner's g come at a serious cost: g exerts non-negligible influence on posterior inference since it governs how posterior mass is spread over the models. For given data, high values of g concentrate posterior mass on few models, which runs the risk of overfitting. In contrast, small values of g spread posterior model probabilities (PMPs) more evenly among the models (irrespective of model sizes and size penalty terms). Posterior statistics, in particular PMPs and the covariates' posterior inclusion probabilities are thus notoriously sensitive to the value of the g prior. In other words, the researcher's prior on g determines how much posterior mass is attributed to the few best-performing models – regardless of whether these have been generating the data.[2] In this paper, we establish the conditions for this g-induced concentration of posterior mass (which we will henceforth refer to as the *supermodel effect*). While crucial in terms of prior sensitivity, this feature went more or less unnoticed in previous simulation studies that focused on 'asymptotic consistency': In order to uncover a single 'true' model in Monte Carlo simulations with a weak noise component, such exercises profit from a large value for g that induces posterior mass to concentrate on the best-performing model.

The proper Bayesian approach to address this problem is to introduce a non-degenerate *hyperprior distribution on g*, and thus 'let the data choose'. Such a flexible prior allows for shrinking the estimated coefficients more toward zero under models with a large noise component,[3] i.e., inducing data-dependent shrinkage. Only few papers have applied such hyperpriors in BMA so far: Among them are Strachan and van Dijk (2004), Cui and George (2008), Liang et al. (2008), and Ley

[1] The literature on the optimal choice of g (e.g., Liang et al., 2008; Ley and Steel, 2011; Hoeting et al., 1999; Fernández et al., 2001a; Eicher et al., 2011) has concentrated on two theoretical considerations: First, asymptotic consistency, i.e. the choice of g such that BMA asymptotically uncovers 'the true model'. However, from a Bayesian viewpoint, many models might be 'true', in the sense that they are generating the data examined. Second, the specification of g was studied in terms of its virtues as a model size penalty term to favor parsimonious models. From a Bayesian perspective, however, such preferences on parameter size should rather be considered in the formulation of model priors, which constitute a crucial component of BMA.

[2] Note that this effect not only poses a risk for model averaging, but also for model selection. For instance, a too large value for g might lead to underestimating the model uncertainty related to selecting a particular model. Moreover, it might increase the risk of selecting the 'wrong' model, cf. discussion in section 3.

[3] I.e. by up- or downweighting the prior beliefs on coefficients β.

and Steel (2011). These authors have proposed various hyperprior specifications in response to theoretical issues other than the problem described above,[4] but none of them has focused on the expected properties of flexible priors under small samples. Most of these hyperpriors have in common that their respective statistics are not available in closed form, thus forcing the researcher to resort to MCMC sampling. The contribution by Liang et al. (2008) differs in providing a closed-form solution – which is so computationally demanding that they implement it via analytical approximations. In this paper, we provide algebraic transformations of the Liang et al. (2008) prior that allow for a sound and accurate numerical application at minimal computational cost. We use this augmented hyper-g prior to show that the model-specific advantages of the hyper-g prior also extend to inference under model uncertainty, as it is not exposed to the supermodel effect a priori. The hyper-g prior adjusts the distribution of posterior mass in dependence of the information provided by the data. Thus if noise dominates the data, PMPs under the hyper-g prior will be distributed more evenly, whereas in the case of minor noise, posterior mass will be concentrated even more than under fixed settings with large values for g.

Based on the above considerations, the contribution of our paper is fivefold: First, we show that fixed coefficient priors may introduce too much or too few shrinkage into individual models, but also have an even stronger impact on the concentration of model probabilities in BMA.[5] We demonstrate the supermodel effect analytically, in a simulation exercise and an empirical application. Second, we propose a particular prior framework that reconciles the Liang et al. (2008) hyperprior with asymptotic consistency, and provide closed-form representations for important posterior quantities. Third, we show further properties of the hyper-g prior: We demonstrate how its posterior statistics are analytically related to the strand of 'Empirical Bayes' priors, and why their results hardly differ under data-sets with a weak noise component. Moreover, we show the relationship of the hyper-g to the familiar OLS F-statistic as a measure of goodness-of-fit. Fourth, we show how the superior robustness of the hyper-g prior addresses the supermodel effect under a simulation exercise with both a simple and a complex data-generating process. Our results show that under noisy data the hyper-g prior dilutes posterior mass among models whereas the popular fixed priors incorrectly favor one (wrong) model. We examine the forecasting properties of various settings for g by means of a simulation study, which points to superior predictive ability of the hyper-g prior under varying signal-to-noise ratios. Finally, our empirical exercise illustrates this advantage in the context of growth regressions. We show how far the BMA parameter instability over revisions of growth data (as found by Ciccone and Jarociński (2010)) is due the supermodel effect, and that the hyper-g prior can reduce it to a great extent.

The remainder of this study is organized as follows: the next section briefly reiterates the concept of Bayesian model averaging and its most popular prior settings. Section 3 sketches the reasons underlying the supermodel effect and provides formal conditions for its presence. Section 4 introduces the hyper-g prior, outlines further posterior statistics and properties, and introduces an implementation strategy of practical relevance. Section 5 presents a simulation exercise that examines the supermodel effect inherent to traditional priors and highlights the predictive performance of flexible priors. The following section demonstrates the sensitivity of posterior results to the choice of g by means of an empirical application to a prominent growth data set. Section 7 concludes the paper.

[4] See footnote 16 for a discussion of these papers' motivations.

[5] The detrimental effect of fixed priors on robustness and the advantage of hyperpriors have not only been noted in Bayesian regression-type models, but also other Bayesian frameworks (see, e.g., Giannone et al. (2012) for a similar motivation in the case of Bayesian VARs). The impact on model averaging, however, is novel, to the best of our knowledge.

2 Bayesian Model Averaging under Zellner's g prior

This section summarizes the popular set-up of Bayesian model averaging (BMA) under the natural conjugate framework with Zellner's g prior and reviews the prior settings that have resurfaced most often in the literature so far. Consider the canonical regression problem of sample size N with the dependent variable in the $N \times 1$ vector y, X_s an $N \times k_s$ design matrix of covariates, and ε an N-dimensional vector of residuals in the following, linear model M_s:

$$y = \mathbf{1}\alpha_s + X_s\beta_s + \varepsilon$$

Here α_s denotes the (scalar) intercept, and β_s the the $k_s \times 1$-vector of unrestricted regression coefficients. The residuals are assumed to be normally IID with variance σ^2, i.e. $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Note that X_s can be assumed to be centered ($X_s'\mathbf{1} = \mathbf{0}$) without loss of generality, as this will only affect the posterior distribution of the constant α_s. Bayesian Model Averaging deals with uncertainty about the model M_s by drawing on the model-specific inference presented above. In the generic linear BMA problem, model uncertainty focuses on the choice of covariates X_s, which may be drawn from a set of K potential regressors. This induces 2^K unique covariate combinations, as represented by the model candidate space $\mathcal{M} = \{M_1, M_2, \ldots, M_{2^K}\}$ (cf. Hoeting et al., 1999, for a more detailed account).

The Bayesian framework calls for specifying a prior distribution on the model's parameters α, β_s, and σ^2. The bulk of the BMA literature (e.g., Raftery, 1995; Chipman et al., 2001), favors the natural-conjugate approach, or its variant outlined by Fernández et al. (2001a): In order to represent lack of information over constant and variance, place improper priors on constant $p(\alpha) \propto 1$ and variance $p(\sigma) \propto \sigma^{-1}$.[6] The prior on coefficients β is assumed to be normal and potentially allows for model-specific elicitation of prior expected value and coefficient covariance. However, the explicit formulation of these hyperparameters is difficult to perform given the many combinations possible in model selection problems. Virtually all linear BMA applications have thus opted for a common uninformative prior centered at zero, with the variance structure given by *Zellner's g prior* (Zellner, 1986):

$$\beta_s|\sigma^2, M_s, g \;\sim\; N\big(0, \sigma^2 g (X_s'X_s)^{-1}\big)$$

This prior assumes the coefficient covariance to be proportional to the posterior covariance expression $(X_s'X_s)^{-1}$ that arises from the sample, with the scalar g determining how certain the researcher is in centering the prior coefficient distribution at zero. Apart from offering computational efficiency, Zellner's g thus reduces the elicitation of the covariance structure to choosing the scalar g. Employing Bayes' theorem via

$$p(\beta_s|y, X_s, M_s) = \int_0^\infty p(\beta_s|M_s, y, X_s, \sigma^2)dp(\sigma^2)$$

yields the posterior coefficient distribution as k_s-variate student-t[7] with expected value $E(\beta_s|y, X, M_s, g) = \frac{g}{1+g}\hat{\beta}_s$, where $\hat{\beta}_s$ denotes the standard OLS estimator for M_s.[8] Note that the posterior expected

[6] Note that the specification for α and σ departs from earlier tradition which typically elicited proper priors for the two parameters. However, both choices do not affect the crucial posterior statistics: The improper prior on the constant allows for an easy disentanglement of the constant with respect to the other coefficients. In contrast to the traditional Gamma-priors, the improper prior on σ offers the advantage of being invariant under scale transformations (Fernández et al., 2001a, p. 391)

[7] Note that this posterior distribution requires $N > 2$.

[8] The posterior variance of β_s is $\frac{\breve{y}'\breve{y}}{N-2}\left(1 - \frac{g}{1+g}R_s^2\right)\frac{g}{1+g}(X_s'X_s)^{-1}$, where $\breve{y} = y - \mathbf{1}\bar{y}$ denotes the centered response vector.

value is a convex combination of its OLS estimator and the prior expected value (zero) weighted by the shrinkage factor $\frac{g}{1+g}$. The larger the shrinkage factor, the more importance is attributed to sample data rather than to prior information.

For its use in BMA, the main advantage of Zellner's g is that it yields a closed-form expression for the marginal likelihood of M_s:[9]

$$p(y|M_s, g) \propto (1+g)^{-\frac{k_s}{2}} \left(1 - \frac{g}{1+g} R_s^2\right)^{-\frac{N-1}{2}} \tag{1}$$

with k_s denoting the number of covariates included in model M_s and R_s^2 its OLS R-squared. This marginal likelihood is crucial in determining the posterior model probability that arises from Bayes' theorem $p(M_s|y, X, g) \propto p(y|M_s, X, g)p(M_s)$ as an update of a prior model probability $p(M_s)$:

$$p(M_s|y, X, g) = \frac{p(y|M_s, X, g)p(M_s)}{\sum_{j=1}^{2^K} p(y|M_j, X, g)p(M_j)} \tag{2}$$

Multiplied with a normalization constant, these posterior model probabilities serve as model weights in Bayesian model averaging. In this vein, the marginal posterior distribution of any statistic Θ may be obtained as a mixture over posterior model probabilities:[10]

$$p(\Theta|y, X, g) = \sum_{j=1}^{2^K} p(\Theta|y, X, M_j)p(M_j|y, X, g)$$

This property is particularly useful in computing the posterior moments of the coefficient vector β as a weighted average over all models.[11] Likewise, posterior inclusion probabilities (PIPs), used for assessing the importance of single covariates, are obtained as the sum of probabilities for all models in which the covariate is included.

In view of equation (2), BMA inference hinges on posterior model probabilities and, in turn, on two important prior specifications: the model priors $p(M_s)$ and Zellner's g prior for the coefficients: The Bayesian framework calls for defining prior model probabilities $p(M_j)$ for all models contained in the model space $j \in \{1, 2, \ldots, 2^K\}$. While advocates of purism may call for subjective prior specification of $p(M_s)$, the number of model candidates renders this virtually infeasible. Consequently, most authors have relied on the uniform model prior $p(M_s) = 2^{-K}$, whereas several (Brown et al., 1998; Sala-i-Martin et al., 2004; Ley and Steel, 2009) have proposed to specify model priors in dependence of average model size k_s, typically in such a way that prior elicitation is reduced to choosing the prior expected model size.

In addition to model priors, the choice of Zellner's g prior crucially affects marginal likelihoods $p(y|M_s, X, g)$ and thus PMPs. Its discussion so far has focused on two considerations:

[9] Note that although the term $((y-\bar{y})'(y-\bar{y}))^{-\frac{N-1}{2}}$ is constant over models, it is frequently included in the marginal likelihood expression, such as in Fernández et al. (2001a) – while others, such as Liang et al. (2008) omit it.

[10] Note that the concept lined out in equation (2), and its implication for posterior statistics, is of course not limited to linear models alone. However, the bulk of the empirical BMA literature focuses on linear models using the g-prior described in this section. The purpose of this paper is to discuss the g-prior's consequences for linear models, and thus it refrains from discussing the implications of similar priors for more complex models.

[11] Note that we have retained the improper priors for α and σ as common to all models.

- Consistency: The choice of g such that posterior model probabilities asymptotically uncover 'the true model' M_T, i.e. $p(M_T|y, X, g) \to 1$ as $N \to \infty$
- The importance of g as a penalty term enforcing parameter parsimony (the factor $(1+g)^{\frac{k_j-k_s}{2}}$ in (2))

Both issues have been reviewed by Fernández et al. (2001a): With respect to consistency, they prove that a choice of $g = w(N)$ such that $\lim_{N\to\infty} w(N) = \infty$ and $\lim_{N\to\infty} \frac{w'(N)}{w(N)} = 0$ ensures consistency as it was mentioned above. Still, consistency leaves open the exact specification of g. Over the course of more than a decade, various 'automatic' or 'default' specifications have been put forward (e.g., Fernández et al., 2001a; Eicher et al., 2011) that typically specify g according to sample size N. Note that the bulk of the literature concentrates on priors *fixing* g in such a way that the penalty term $(1 + g)^{-\frac{k_s}{2}}$ in (2) asymptotically mimics popular information criteria.[12] In particular, two settings for g resurface steadily in the literature: The Unit Information Prior (g-UIP) corresponds to $g = N$. Through its dependence on sample size it is a consistent prior and draws on the notion that the 'amount of information' contained in the prior equals the amount of information in one observation (Kass and Wasserman, 1995). Fernández et al. (2001a, p.424) demonstrate that as $N \to \infty$ the log of the Bayes factor for two models approaches the ratio of their Bayesian information criteria. Secondly, setting $g = K^2$ (g-RIC) calibrates the posterior model probability to asymptotically match the risk inflation criterion proposed by Foster and George (1994). Based on an extensive study of various specifications for g, Fernández et al. (2001a) recommend the 'benchmark' prior which bridges the g-UIP and the g-RIC by setting $g = \max(N, K^2)$.

3 The Supermodel Effect

The advantages of Zellner's g prior have fostered its widespread use in BMA, even though its exact specification is still subject to debate (as highlighted by the previous section). In general, g determines the tightness of the prior distribution on coefficients β around their prior expected value zero: Large g implies a diffuse prior distribution, i.e. the researcher is very uncertain about the prior expected value and relies heavily on the data. Small g means a prior that is more tightly centered at zero,[13] and leaves less scope to the data to determine the coefficients. In this sense, it is evident why most g specification schemes aim to set this hyperparameter according to data quality: A high signal-to-noise ratio warrants strong reliance on the data and thus a high g, while an important noise component should be met with a low g.

In practice, the choice of g can have considerable consequences for the robustness of BMA results. For instance, with very noisy data, a large g could attribute too much weight to results that are mainly driven by a particular realization of the error term. Such a case may lead to situations as in Ciccone and Jarociński (2010), who show that BMA under the 'benchmark' specification from Fernández et al. (2001a) produces results that differ strikingly over small revisions to the

[12] Information criteria are a widely used approach for model selection and are conceptually similar to the marginal likelihoods that arise from incorporating model uncertainty in a Bayesian framework (Indeed, several popular information criteria (IC) can be derived from such a setting, such as Schwarz (1978)). Drawing on this similarity, 'frequentist' model averaging techniques rely on IC in order to obtain 'posterior' model weights (see Claeskens and Hjort, 2008, for an overview). Due to their numerical connections, empirical results under IC-based linear model averaging are usually quite similar to BMA with g-priors that mimick IC. In this paper, we therefore forgo the explicit discussion of IC-based model averaging techniques, and concentrate on their Bayesian analogues.

[13] In general, small g implies centering at the prior (expected values of coefficients). Here, we follow the bulk of the literature in presupposing coefficient priors to be centered at zero.

response variable. Such robustness problems can arise from a too loose g that focuses posterior model mass on too few 'supermodels'. In this section, we demonstrate that g is positively linked with the concentration of posterior model probabilities – and that this *supermodel effect* matters to empirical practice. A look at the role of the shrinkage factor $\frac{g}{1+g}$ in the marginal likelihood from (1) provides some intuition:

$$p(y|M_s, g) \propto \underbrace{(1 - \tfrac{g}{1+g})^{\frac{k_s}{2}}}_{A} \underbrace{\left(1 - \tfrac{g}{1+g} R_s^2\right)^{-\frac{N-1}{2}}}_{B} \tag{3}$$

The shrinkage factor $\frac{g}{1+g}$ affects marginal likelihood (and thus posterior model probability) via a size penalty term (A) and a model fit term (B). The term (A) shapes the distribution of posterior mass between different model sizes, while term (B) determines the concentration of PMP within models of the same size k_s. Among models of the same size k_s, larger $\frac{g}{1+g}$ will increase the relative posterior weight of the models with the largest R_s^2. The term (B) thus implies a direct positive link between the shrinkage factor and relative PMP concentration among models of the same size k_s.

The term (A) has stimulated most of the debate on the g prior through its virtues as a size penalty term that could mimic well-founded information criteria. But from a Bayesian viewpoint, size penalty represents prior preferences on model size that should be fused into the formulation of the model prior rather than a coefficient prior. Instead, one should consider that the term (A) can reinforce the link between g and PMP concentration: Increasing the shrinkage factor $\frac{g}{1+g}$ strengthens the size penalty and skews the posterior model size distribution to smaller (more parsimonious) models. When most of the posterior mass focuses on models below the size of $K/2$ (which typically applies to empirical exercises), strengthening the size penalty means concentrating mass on model sizes that comprise fewer models to choose from. In this sense, the term (A) contributes to a positive link between g and PMP concentration.

The interplay of posterior model size distribution (from term (A)) and relative PMP concentration among models of a given size (from term (B)) is not straightforward, and can depend on data set characteristics and the particular value of g. Proposition 1 therefore formally pins down the conditions for the supermodel effect:

PROPOSITION 1 *For linear BMA with a fixed common Zellner's g prior, a given realization of* (y, X)*, and any model prior that does not depend of g, the following holds: The cumulative posterior probability of the best r models have a non-negative derivative with respect to g if* $E_r(k|y, X) + \eta < E(k|y, X)$*, where* $E(k|y, X)$ *represents posterior model size and* $E_r(k|y, X)$ *expected posterior model size of the best r models; with* $\eta > 0$ *vanishing as* $N \to \infty$ *or* $g \to \infty$*.*

Thus an increase in g will increase the concentration of PMP on the most important models (by PMP), as long as their average model size is somewhat smaller than the overall posterior model size. Conversely, the most important models could only lose PMP with increasing g if their model sizes are relatively large. In any case, the cumulative PMP of the most important models will increase almost sure over the domain of g.[14] Proposition 1 also holds implications for the special case of model selection (as opposed to model averaging): Increasing g attributes increasingly (perceived) posterior importance to the model which the largest PMP (if its model size is smaller than average model size). However, at some point the conditions of Proposition 1 will not be met locally, and the PMP of this best-performing model will be overtaken by the PMP of a model with fewer parameters.

[14]This result obtains trivially from the fact that the PMP for the null model tends to 1 as g approaches infinity, and that the null model will be the most important model for all g greater than some finite $\bar{g}$.

Thus, model selection with high g values under fixed priors risk selecting a too small model (that is possibly not even nested in the ' true' model), and underestimating the model uncertainty around it.

In order to illustrate the supermodel effect, consider the BMA results from the growth data set as in section 6, for different values of g and under uniform model priors.[15] Figure 1 (top panel) exhibits the posterior model size as well as the cumulative posterior probability of the best 2,000 models for varying values of g. The results show that increasing g leads to a marked reduction in posterior model size, while increasing the concentration of posterior model probabilities. The rate of increase in the PMP of the best 2,000 models certainly depends on the characteristics of the data set and might vary locally, but the exercise illustrates that the concentration of PMP broadly intensifies with increasing g. In particular, this effect applies to ranges of g that are typically used in the empirics of economic growth.

The previous discussion has shown that term (B) in equation (3) directly links g positively with PMP concentration. In contrast, term (A) might have an ambiguous effect, which also underlies the qualifying conditions in Proposition 1. In order to disentangle the two effects in the empirical exercise, we neutralize the term (A) by the following model prior:

$$p(M_s) = \frac{(1+g)^{\frac{k_s}{2}}}{(\sqrt{1+g}+1)^K}$$

Such a model prior exactly cancels the size penalty term (A) in the marginal likelihood (3). The impact for differing values of g is shown in Figure 1 (bottom panel). When size penalty is neutralized, higher g leads to increased posterior model size, as posterior model mass then concentrates more on the least parsimonious model which have the largest R^2. As expected, the cumulative PMP of the best 2,000 models increases with g, although at a considerably lower pace than under the uniform model priors discussed above. We therefore conclude that for the growth data examined in this paper, the supermodel effect has a sizable impact, and is mainly driven by the size penalty term $(1-\frac{g}{1+g})^{\frac{k_s}{2}}$.

Note that the supermodel effect not only has implications for the skewness of the PMP distribution, but also for the regressors' posterior inclusion probabilities (PIPs). The term (B) in (3) establishes a direct positive link between g and the concentration of PIPs, as larger g leads to more concentration in the relative PMPs for each model size. The term (A) can reinforce this effect: Note that the variance of R^2 for models of size k is by definition (weakly) greater than the variance of R^2 for models of size $K-k$. This implies that relative PMP concentration among models of the same size k is more intense for smaller k. Increasing g therefore tends to skew the relative distribution of the PIPs. In view of Proposition 1 and the above illustration we therefore conclude that g is positively linked not only with the concentration of PMPs, but also of posterior inclusion probabilities.

4 Flexible priors: The Hyper-g Prior

The previous section has demonstrated the problems with fixed g priors: First, it might be insensitive to assume a common parameter for all models considered, and second, eliciting the right parameter value strongly risks over- or under-identification with respect to posterior model and inclusion probabilities. Introducing a flexible hyper-prior on g, in contrast, would allow to update

[15] Results are from BMA estimations with uniform model priors of the Sala-i-Martin et al. (2004) data set with growth and initial income according to the Penn World Tables revision 6.3. Figure 1 displays the result for 24 different values of g, each estimated from MCMC sampling with 200,000 burn-in draws and 2,000,000 subsequent iterations.

prior beliefs according to data quality, and thus mitigate the risks from choosing a prior value for g.

In principle, the concerns raised in the previous section might be addressed by virtually any flexible hyper-prior setting for g. Based on different motivations, recent contributions have introduced several candidates that could potentially be used to that end (cf. Ley and Steel, 2011, for an overview). However, most of them do not yield closed form solutions for posterior statistics of interests. Instead, those have to be obtained by numerical sampling techniques, which complicates the computationally demanding task of evaluating models in BMA. Among the proposed candidates g, only the hyper-g prior by Liang et al. (2008) stands out as fairly flexible prior distribution that allows for closed-form posterior statistics. We therefore concentrate on the latter approach to demonstrate the properties of hyper-priors vs. fixed priors.

Liang et al. (2008) introduce two priors motivated on theoretical grounds,[16] among them the closed-form hyper-g prior. While they ingeniously outline the basic features of the hyper-g prior, their posterior expressions involve ratios of hypergeometric functions, which are difficult to evaluate computationally. For feasible computational implementation, the authors thus resort to Laplace approximations – an approach that risks numerical inaccuracies, in particular with respect to the mentioned ratios. This section therefore introduces algebraic transformations of these expressions that yield accurate statistics at low computational cost. Moreover it complements Liang et al. (2008) by establishing additional, common posterior expressions, in particular with respect to second moments of posterior parameters and predictive distributions. Finally we proceed to show some properties of the hyper-g prior, in particular how it be may reconciled with consistency in the sense of Fernández et al. (2001a), its asymptotic equivalence to the Empirical Bayes (EBL) prior, and the relationship between its posterior statistics and the OLS F-statistic.

4.1 The hyper-g prior and its posterior statistics

The hyper-g prior for g translates into a Beta prior on the shrinkage factor $\frac{g}{1+g}$ that is common to all models (Liang et al., 2008, p. 415):

$$\frac{g}{1+g} \sim \text{Beta}\left(1, \frac{a}{2} - 1\right)$$

i.e. $\frac{g}{1+g}$ is Beta distributed with $E(\frac{g}{1+g}) = \frac{2}{a}$.[17] The elicitation of g is therefore supplanted by the choice of the hyperparameter $a \in (2, \infty)$: $a = 4$ renders the prior distribution of $\frac{g}{1+g}$ uniform, while moving a close to 2 concentrates the prior mass on the shrinkage factor close to 1. Conversely, any $a > 4$ tends to concentrate prior mass near 0. Liang et al. (2008) therefore omit those cases and concentrate on $a \in (2, 4]$ – a strategy we will follow in this study.

An integral representation for the Gaussian hypergeometric function ${}_2F_1(a, b, c, z)$ allows for straightforwardly establishing the model-specific posterior distribution of the shrinkage factor (Abramowitz and Stegun, 1972, p.563).

[16] Liang et al. (2008) motivate their paper with two 'paradoxes' that arise with constant g. First, they raise a BMA formulation of 'Bartlett's paradox' stating that if $g \to \infty$ for fixed N and K, the Bayes Factor $B(M_s : M_0)$ of any model with respect to the null model eventually goes to zero. Second, they refer to an 'information paradox' stating that for fixed N and K, if the R-squared of model M_s converges to unity, its Bayes factor with respect to any other fit-wise inferior model does not go to infinity. Moreover, both arguments bite only in the case when N and K are kept constant: Bartlett's paradox in this case may be less relevant as typical specifications for g require it to rise in line with N. The 'information paradox' does not contradict the standard consistency argument that requires the respective Bayes Factor to converge to infinity only when N tends likewise to infinity. See the comment by Zellner (2008) for a more detailed discussion.

[17] Note that this is equivalent to putting the following prior on g: $p(g) = \frac{a-2}{2}(1+g)^{-\frac{a}{2}}$.

$$p(\tfrac{g}{1+g}|y, X_s, M_s) = \frac{k_s + a - 2}{2_2F_1\left(\frac{N-1}{2}, 1, \frac{k_s+a}{2}, R_s^2\right)} \left(1 - \tfrac{g}{1+g}\right)^{\frac{k_s+a-4}{2}} \left(1 - \tfrac{g}{1+g} R_s^2\right)^{-\frac{N-1}{2}} \tag{4}$$

The kernel of the shrinkage factor's posterior distribution mimics the expression for marginal model likelihood (1).[18] It thus skews the posterior density towards values close to one as the parameters N or R_S^2 increase. In contrast, the shrinkage factor density concentrates closer to zero with increasing parameter size k_s or the hyperparameter a. In other words, the posterior density adapts to a model's marginal likelihood, rewarding good fit with increasing the shrinkage factor towards one (i.e., emulating maximum -likelihood estimates), while punishing parameter size with shrinking posterior estimates towards zero. In this sense, the posterior density of $\frac{g}{1+g}|y, X_s, M_s$ follows a behavior similar to information criteria or the OLS F-statistic.

The integration constant of (4) is a Gaussian hypergeometric function, which consequently also turns up in the expression for marginal likelihood of model M_s (cf. Liang et al., 2008, equation (17)):

$$p(y|X_s, M_s) \propto (\breve{y}'\breve{y})^{-\frac{N-1}{2}} \; \tfrac{a-2}{k_s+a-2} \; {}_2F_1\left(\tfrac{N-1}{2}, 1, \tfrac{k_s+a}{2}, R_s^2\right) \tag{5}$$

While this expression differs from the expression for marginal likelihood under fixed g (1), it displays similar behaviour with respect to parameters. Its partial derivatives with respect to parameters correspond to the ones in (1). Building on equation (5), Liang et al. (2008, equation (19)) proceed by expressing the posterior expected value of the shrinkage factor $\mathrm{E}\left(\frac{g}{1+g}\middle| y, X_s, M_s\right)$ as a ratio of two hypergeometric functions. The expression is primarily relevant for the expected value of the response:[19]

$$E(y|X_s, M_s) = \mathbf{1}E(\alpha_s|X_s, M_s) + \mathrm{E}\left(\tfrac{g}{1+g}\middle| y, X_s, M_s\right) X_s\hat{\beta}_s \tag{6}$$

with $\hat{\beta}_s$ denoting the estimated OLS coefficient for model M_s. Equation (6) highlights the importance of the shrinkage factor, as the hyper-g prior allows for model-specific, data-adaptive shrinkage as opposed to fixing the value for the shrinkage factor a priori.

The posterior statistics outlined so far suffice for the analysis in Liang et al. (2008). However, fully Bayesian inference requires several more expressions, notably with respect to second moments. Therefore, we introduce in equations (7)-(A.2) the moments of the shrinkage factor and the coefficients, as well as the posterior distribution of coefficients (For completeness, the posterior predictive distribution is provided in the appendix). Note that straightforward integration characterizes all of these posterior moments as fractions of differing hypergeometric functions. However, they may all be expressed as functions of a single scalar $F_s^* \equiv {}_2F_1(\frac{N-1}{2}, 1, \frac{k_s+a}{2}, R_s^2)$ using Gauss' relations for contiguous hypergeometric functions (Abramowitz and Stegun, 1972, p.563). Let $\bar{N} \equiv N - 3$ and $\bar{\theta}_s \equiv k_s + a - 2$ represent collected terms. Tedious, but straightforward algebra then yields the following results for posterior moments (as long as $R_s^2 \in (0, 1)$):[20]

[18] Note that due to this feature, the mode of density (4) is the local Empirical Bayes prior (EBL) from section 4.2.

[19] Note that $\mathrm{E}(\beta_s|y, X_s, M_s) = \mathrm{E}\left(\frac{g}{1+g}\middle| y, X_s, M_s\right)\hat{\beta}_s$.

[20] In case $R_s^2 = 0$ (in particular for the null model), the respective quantities are $\mathrm{E}\left(\frac{g}{1+g}\middle| y, X_s, M_s\right) = \frac{2}{k_s+a}$, and $\mathrm{Cov}(\beta_s|y, X_s, M_s) = \frac{2}{k_s+a} \; \frac{\breve{y}'\breve{y}}{N-2}(X'X)^{-1}$

$$\mathrm{E}\left(\frac{g}{1+g}\middle| y, X_s, M_s\right) = \frac{1}{R_s^2(\bar{N}-\bar{\theta}_s)}\left(\frac{\bar{\theta}_s}{F_s^*} - \bar{\theta}_s + \bar{N}R_s^2\right) \tag{7}$$

$$\begin{aligned}\mathrm{Cov}(\beta|y, X_s, M_s) = \frac{\breve{y}'\breve{y}}{N-2}(X'X)^{-1}\,\frac{\bar{N}}{(\bar{N}-\bar{\theta}_s-1)^2-1}\frac{1-R_s^2}{R_s^2}\times \\ \times\left(\left(1+\frac{2}{\bar{N}}\frac{R_s^2}{1-R_s^2}\right)\frac{\bar{\theta}_s}{F_s^*} + ((\bar{N}-2)R_s^2 - \bar{\theta}_s)\right)\end{aligned} \tag{8}$$

Note that the equations above all contain the term $\bar{\theta}_s/F_s^*$.[21] So for each model's statistics, a hypergeometric function (or its Laplace approximation) has to be computed only once, which benefits numerical implementation in terms of computational burden.[22] In the BMA implementation used in the next section, the speed loss of hyper-g vs. a fixed-g setting is around 30%. Similarly cumbersome algebra also establishes the higher moments of the shrinkage factor and the moments of the predictive distribution as simple transformations of $\bar{\theta}_s/F_s^*$. Section A.1 in the appendix presents these terms for reference, as well as a closed-form expression for the posterior coefficient density.

4.2 Properties of the hyper-g prior

The hyperparameter a can be trimmed to capture prior beliefs on the shrinkage factor in the associated Beta distribution. It is straightforward, for instance, to specify the prior beliefs such that the expected shrinkage factor matches the expressions of popular fixed g priors. In general, most popular settings for g can thus be emulated by $a = 2 + 2/w(N)$, with $w(N) > 0$, $w'(N) > 0$ and $\lim_{N\to\infty} w(N) = \infty$, thus positioning the prior expected value at $E(\frac{g}{1+g}) = \frac{w(N)}{1+w(N)}$. Setting a in dependence of sample size has the appealing virtue of ensuring 'consistency' in the sense of Fernández et al. (2001a, p.6).[23] By the same mechanism as in the corresponding fixed settings, the weight of the prior vanishes with increasing sample size and thus lets the posterior probability of a 'true' model $p(M_T|y)$ tend to to unity. Note that this applies to any true model; a proof is provided in section A.2 in the appendix.
In this light, we concentrate on the following specifications for adaptive shrinkage priors:

- *HG-UIP*: $a = 2 + \frac{2}{N}$ corresponds to the 'g-UIP'-shrinkage factor with $E\big(\frac{g}{1+g}\big) = \frac{N}{1+N}$. Then 95% of the prior mass on the shrinkage factor is contained in the interval $[1 - 0.95^N, 1]$.
- *HG-RIC*: $a = 2 + \frac{2}{K^2}$ corresponds to 'g-RIC'-shrinkage with $E\big(\frac{g}{1+g}\big) = \frac{K^2}{1+K^2}$. In this case 95% of the prior mass is contained in the interval $[1 - 0.95^{K^2}, 1]$. Akin to Fernández et al. (2001a), such a setting will be asymptotically consistent by choosing $w(N) = \max(N, K^2)$.
- *Empirical Bayes – Local (EBL):* $g_s = \arg\max_g p(y|M_s, X, g)$. Authors such as George and Foster (2000) or Hansen and Yu (2001) advocate an 'Empirical Bayes' approach by using

[21] Note that $\bar{\theta}_s/F_s^*$ is just $2/(a-2)$ times the integration constant of $p(g|y, X_s, M_s)$ or $\bar{\theta}_s/F_s^* = \frac{a-2}{BF(M_s:M_0)}$ where $BF(M_s : M_0)$ is the null-based Bayes Factor for model M_s.

[22] Note that with respect to equations (7) and (A.1) it is straightforward to derive the corresponding expressions for $E(g|y, X_s, M_s)$ and $E(g^2|y, X_s, M_s)$. However, $E(g|y, X_s, M_s)$ will only be finite for $k_s + a > 4$ and $E(g^2|y, X_s, M_s)$ only for $k_s + a > 6$. We therefore concentrate on the posterior moments of the shrinkage factor.

[23] Consistency does not directly apply to the g-RIC prior outlined below. However, throughout the following sections, g-RIC is in practice identical with the g-BRIC prior (as always $K^2 > N$). Since the latter qualifies for consistency, the notion may be extended to g-RIC, at least in our case.

information contained in the data (y, X) to elicit g. The latter provide a theoretical underpinning for doing so locally, i.e. separately for each model. In the formulation given in Liang et al. (2008), this corresponds to $g_s = \max(0, F_s - 1)$ where F_s is the standard F-statistic for M_s, with $F_s = \frac{R_s^2(N-1-k_s)}{(1-R_s^2)k_s}$. Note that this formulation frequently raises objections, since it is not necessarily consistent and the data-dependency of g runs counter the intuition of a prior.

Similarly, other specifications akin to 'classic' g formulations could be implemented – as long as they depend on N as defined above, in order to retain asymptotic consistency. Henceforth, we will refer to such elicited hyper-g priors as 'consistent hyper-g priors'. However, as posterior expressions are quite insensitive to the value of a, and most of these formulations will lead to a close to 2, the resulting posterior statistics will be virtually identical. We therefore limit our attention to the two specifications above.

Equations (7)-(8) reveal a certain resemblance to the respective posterior statistics under the 'Empirical Bayes - Local' (EBL) approach, whose posterior statistics depend on the OLS F-statistic. This feature is not surprising, as many Bayesian posterior statistics under a well-defined, non-degenerate prior asymptotically converge to their maximum-likelihood equivalent. Section A.3 in the appendix shows that also the posterior model probabilities under EBL and consistent hyper-g priors (5) converge asymptotically, given that the 'true' model is not the null model (with zero covariates). But the similarities between EBL and hyper-g priors also extend to small samples: The main difference between their posterior expressions is the term $\bar{\theta}_s/F_s^*$, which guarantees non-negativity for the hyper-g statistics. Considering that the models associated with very low $\bar{\theta}_s/F_s^*$ (and thus high PMP) are disproportionally weighted into model averaging, this term thus virtually disappears from model-averaged statistics, if the data is not completely dominated by noise.[24] Consequently, the posterior statistics under both types of flexible priors will be very similar under any sample with a decent signal-to-noise ratio. However, compared to hyper-g, the EBL setting has two major drawbacks: First, it is not a prior in the classical sense, as it draws on the dependent variable. Second, it cannot be established whether the EBL setting is consistent if the true model is the null model. Nonetheless, due to its computational simplicity, the EBL prior can serve as a reasonable approximation to (and shares its asymptotic properties with) the hyper-g prior under data with a small noise component.

In view of the resemblance between the hyper-g prior results and the OLS- F-statistic, the posterior distribution of the shrinkage factor $\frac{g}{1+g}$ could be interpreted in terms of goodness-of-fit: Equation (7) presents its model-specific expected value as close to $1 - 1/\hat{F}_s$, where $\hat{F}_s$ represents an adjusted OLS F-statistic for the model M_s: $\hat{F}_s = \frac{R_s^2(\bar{N}-\bar{\theta}_s)}{(1-R_s^2)\bar{\theta}_s}$. Larger values of the shrinkage factor hence correspond to more variance explained by the model M_s. The model-averaged expected value of the shrinkage factor $E(\frac{g}{1+g}|y, X)$ may be interpreted likewise. If $K + a < N + 1$,the following inequality will hold asymptotically under a consistent hyper-g prior as N tends to infinity – in small samples, it will hold as well except in cases of very low data quality (cf. section A.4).

$$\frac{(1 - p(M_0|y, X))^2}{1 - E(\frac{g}{1+g}|y, X)} \leq E\left(\frac{\bar{N} - \bar{\theta}}{\bar{\theta}} \frac{R_s^2}{1 - R_s^2}\middle| y, X\right) \equiv E(\hat{F}|y, X) \tag{9}$$

The model-weighted average of adjusted F-statistics thus establishes an upper bound for the posterior shrinkage factor. Consequently, the shrinkage factor can be related to goodness-of-fit in the data (y, X). The term involving the probability of the null model $p(M_0|y, X)$ is necessary, since

[24] Note that ${}_2F_1(\frac{N-1}{2}, 1, \frac{k_s+a}{2}, R_s^2)$ increases rapidly as R_s^2 increases. The term $\bar{\theta}_s/F_s^*$ could thus noticeably affect model-averaged posterior moments only in case the data examined offers a very low signal-to-noise ratio.

the F-statistic of a model M_s will in general move in line with $E(\frac{g}{1+g}|y, X, M_s)$, except in case of the null model (there, the posterior conforms to the prior $\frac{2}{a}$).

A similar argumentation allows for relating posterior shrinkage to the F-statistic of the full data sample: If $N \to \infty$ under a consistent hyper-g prior, then the F-statistic of the full model with K regressors will form an upper bound for a function of posterior shrinkage (inequality (10)). Note that this inequality will also hold in small samples as long as there are some posterior model probabilities considerably larger than the one of the null model.[25]

$$\frac{1}{1 - E(\frac{g}{1+g}|y, X)} \leq \frac{R_F^2}{(1 - R_F^2)} \frac{(N - E(k|y, X) - a - 1)}{(E(k|y, X) + a - 2)} \tag{10}$$

Here, R_F^2 denotes the OLS R-squared of the full model, and $E(k|y, X)$ is the expected posterior model size. The right-hand side thus constitutes an adjusted F-statistic that relates R_F^2 with 'effective parameter size' $E(k|y, X) + a - 2$. Note that this adjusted F-statistic is (almost sure) larger than the F-statistic of the full model, which illustrates the estimation advantage of shrinkage methods versus OLS. It is thus straightforward to express shrinkage as a function bounded by the unadjusted OLS F-statistic, which allows for applying likelihood-ratio significance tests in a classic sense. The relationship between the F-test and and information criteria thus implies that the posterior expected shrinkage factor is an indicator for goodness-of-fit of the model average that behaves similar to information criteria.

5 A simulation exercise

In this section we carry out a simulation study that empirically investigates the supermodel effect and assesses the predictive performance of selected prior structures. We group this broadly into *fixed* prior settings, as discussed in section 2, as opposed to model-specific adaptive *flexible* g priors (as in section 4.1). In the following, we concentrate on the 8 prior structures given in Table 1.

Fixed Prior Settings	
g-RIC	Risk inflation criterion, $g = K^2$.
g-UIP	Unit information prior, $g = N$.
g-E($\frac{g}{1+g}|y$)	$\frac{g}{1+g}$ is set to the posterior mean under the HG-4 prior (i.e. $E(\frac{g}{1+g}|y)$).

Flexible Prior Settings	
EBL	Local empirical Bayes estimate of g.
HG-3	Hyper-g prior with $a = 3$.
HG-4	Hyper-g prior with $a = 4$.
HG-RIC	Hyper-g prior with $a = 2 + 2/K^2$.
HG-UIP	Hyper-g prior with $a = 2 + 2/N$.

Table 1: Definition of Prior Settings.

The first two fixed settings correspond to what Fernández et al. (2001a) coined the 'benchmark'

[25] Even though this inequality will hold in virtually all relevant cases for small samples, it may not hold in case the dependent-covariate correlation is less than expected under a null hypothesis of no relation. As a rule of thumb, $R_F^2 > \frac{K+a-2}{N-3}$ is sufficient for (10) to hold in any case. Please refer to section A.4 in the appendix for further details.

prior and is widely used in applied work.[26] In macroeconomic studies such as (e.g., Fernández et al., 2001b), their recommendation usually results in the g-RIC prior. The implied (large) value for g under g-RIC is expected to have two consequences: first g-RIC will favor parsimonious models, and second posterior mass will be concentrated on a small set of models.[27] The unit information prior and the g-E($\frac{g}{1+g}|y$) complete the set up for fixed prior structures on g. For the latter we impose $\frac{g}{1+g}$ a priori to equate the (model weighted) posterior mean of ($\frac{g}{1+g}|y$) under the HG-4 setting (the hyperprior with $a = 4$). We have chosen this particular prior structure to exemplify the impact of adaptive shrinkage: both the HG-4 and g-E($\frac{g}{1+g}|y$) priors share the same average shrinkage factor and thus should yield a similar posterior model size distribution. However, posterior results are expected to seriously differ regarding the relative concentration of PMPs. In keeping g constant, the g-E($\frac{g}{1+g}|y$) setting will favor models which have a comparably small posterior support under the HG-4 prior. The differing results between g-E($\frac{g}{1+g}|y$) and HG-4 thus illustrate the impact of model-specific shrinkage as opposed to adapting the aggregate shrinkage factor to the data.

The flexible prior structures with model specific, data-dependent shrinkage divide into local empirical Bayes (EBL) estimates and the hyper-g prior corresponding to a fully Bayesian approach. One strength in placing a prior on g lies in the fact that we can incorporate our prior beliefs following the rules of Bayesian statistics[28] via the hyperparameter a. For the simulation study, we devise four different values for a: HG-3 (a=3) corresponds to a prior expected shrinkage factor of $\frac{2}{3}$, whereas HG-4 (a=4) corresponds to a flat prior over the shrinkage factor. We contrast these two settings with two consistent priors that are calibrated to match the g-RIC and g-UIP prior structure (HG-RIC, HG-UIP). I.e., the prior expected value of the shrinkage factor $E(\frac{g}{1+g})$ conforms to the shrinkage factors induced by g-RIC ($g = K^2$) or g-UIP ($g = N$).

Data-wise, we employ two different settings, where the first set-up 'A' follows Fernández et al. (2001a). Each Monte Carlo run draws ten potential explanatory variables $(\mathbf{x}_1, \ldots, \mathbf{x}_{10})$ with $N = 100$ observations from a standard normal distribution for each covariate. Five more variables are generated by multiplying the first five regressors with the vector $(0.3, 0.5, 0.7, 0.9, 1.1)$ in order to induce a correlation structure among the covariates. Note that this correlation structure hampers uncovering the data-generating model under short samples.

The second set-up 'B' is more demanding since the data-generating process cannot be traced back to a single model. This is more in line with the Bayesian model averaging approach, whose question is not whether the preferred model is perfectly true, but whether under the assumed model(s) the observed data is a plausible outcome.[29] The data-generating process is composed of five partially nested models with unequal model weights imposed. This creates a 'hierarchy' of models, with y_4 and y_5 dominating the remaining models in terms of explained variation.[30]

Posterior inference under the different prior structures will be examined with varying signal-to-noise ratios. In particular we conduct the simulation study for four levels of noise:[31] $\sigma = 1/2, \sigma = 1, \sigma = 2.5, \sigma = 5$. For each value of σ and each setting, results are computed as averages from 50 Monte Carlo draws. The relatively low number of covariates $K = 15$ allows for easily enumerating the full

[26] See, for instance, Fernández et al. (2001b), Masanjala and Papageorgiou (2008) as well as Koop and Potter (2003).

[27] Note that this feature facilitates quick convergence of stochastic search algorithms such as the MC^3 to the target distribution.

[28] See Laud and Ibrahim (1995) for a model selection approach designing information criteria that allow for the input of prior knowledge.

[29] See, for instance, Gelman et al. (1995).

[30] Note that setup 'B' is observationally equivalent to generating data from a single, complicated model.

[31] Note that in this setting, varying σ has an effect that is similar to varying the number of observations N. We therefore leave N constant.

Setup 'A': $y = 4 + 2\mathbf{x}_1 - \mathbf{x}_5 + 1.5\mathbf{x}_7 + \mathbf{x}_{11} + 0.5\mathbf{x}_{13} + \sigma\varepsilon$

Setup 'B':
$$
\begin{aligned}
y &= 0.2y_1 + 0.1y_2 + 0.1y_3 + 0.3y_4 + 0.3y_5 \\
y_1 &= 4 + 2\mathbf{x}_1 - \mathbf{x}_5 + 1.5\mathbf{x}_7 + \mathbf{x}_{11} + 0.5\mathbf{x}_{13} + \sigma\varepsilon \\
y_2 &= 4 + 4\mathbf{x}_1 - \mathbf{x}_5 + 1.5\mathbf{x}_2 + \mathbf{x}_8 + 0.5\mathbf{x}_{11} + \sigma\varepsilon \\
y_3 &= 4 + 1\mathbf{x}_5 - \mathbf{x}_7 + 1.5\mathbf{x}_3 + \mathbf{x}_9 + 0.5\mathbf{x}_6 + \sigma\varepsilon \\
y_4 &= 4 + 2\mathbf{x}_1 - \mathbf{x}_2 + 1.5\mathbf{x}_4 + \mathbf{x}_7 + 0.5\mathbf{x}_6 + \sigma\varepsilon \\
y_5 &= 4 + 2\mathbf{x}_1 - \mathbf{x}_{10} + 1.5\mathbf{x}_{11} + \mathbf{x}_{12} + 0.5\mathbf{x}_{13} - 2\mathbf{x}_{14} + \sigma\varepsilon
\end{aligned}
$$

model space of 2^K models. This guarantees that the differences of results for the competing priors do not arise from variation due to stochastic search.

Empirical research frequently focuses on the posterior inclusion probabilities (PIPs) of the variables entering the analysis and the posterior moments of the related coefficients. Table 2 and 3 highlight PIPs for setting 'A': Under a small degree of noise ($\sigma = 1/2$ and $\sigma = 1$) results do not differ considerably between fixed and data-dependent priors for g. Under the $\sigma = 2.5$ setting, the PIPs of the coefficients from the the data-generating model exhibit differences in magnitude but still lead to the same interpretation. Results change when looking at the $\sigma = 5$ case. Posterior mass under the flexible priors is spread more evenly than under fixed g priors. The g-RIC prior shows strong support for the first variable, with a large PIP for β_1 of approximately 0.8. The remaining variables receive negligible posterior support, tempting the researcher to believe that the data-generating process is solely driven by the first variable. In contrast, flexible priors still 'identify' all variables. As expected, mass is spread more evenly, and over larger models,[32] which results in a high share of covariates with PIP close to 0.5 and thus reflects the serious degree of noise in the data. Note, however, that under high noise, the g-E($\frac{g}{1+g}|y$) prior (which is data-adaptive but not model-specific) is more prone to misidentifying variables (by PIP) than the hyper-g priors. This result suggests that while adapting shrinkage to data quality is crucial, it is also important to allow for model-specific adjustment of the shrinkage factor.

In addition to PIPs, the posterior model probability of the data-generating model can be of interest to examine consistency properties in the sense of Fernández et al. (2001a). Tables 6 and 7 show summary statistics for its posterior model probability (under setting 'A'). In line with asymptotic consistency, more information in the data leads the hyperprior to uncover the data-generating process with highest precision, whereas increasing noise deteriorates the selection ability of BMA for all settings. The ratio of the posterior model probability for the data-generating process to the one with highest PMP is given in Table 7. The results show that in situations described by higher degrees of noise in the data all specifications favor a model different from the one generating the data.

While flexible priors fail to uncover the data-generating model (as do the fixed priors) the assigned PMP for the best (and wrong) model is considerably smaller than under fixed priors. Hence, the degree of uncertainty is reflected in the evenness of posterior mass distribution. Figures 2 and 3 exemplify the differences in PMP concentration for the 8 priors. The first figure shows the cumulative posterior mass of the 100 best models under the four signal-to-noise settings. From these figures and Table 6, it becomes evident that flexible priors uncover the data-generating model with highest precision *and* concentrate most mass on this model(s) in situations characterized by a high degree of information in the data. This means that under the flexible priors, posterior

[32] Note that the sum of PIPs equals posterior model size. Therefore, if posterior mass concentrates on larger models due to noise, the PIPs will not discriminate much among covariates but will exhibit high absolute values. It is therefore more insightful to compare the relative differences in PIPs rather than their absolute values.

expected shrinkage $E(\frac{g}{1+g}|y, X)$ is larger than the constant factors $\frac{g}{1+g}$ under the fixed priors. As noise increases, the flexible priors distribute mass more evenly among explanatory variables which reflects the rise in uncertainty. In contrast, fixed g priors are not capable of adjusting posterior mass distribution to uncertainty inherent in the data. This inability to adapt limits the merits of Bayesian model averaging under fixed g priors with respect to robust inference and predictive ability. Figure 3 uses a QQ-plot to compare for PMPs under the various prior settings with the g-RIC prior. For all data-dependent priors, differences to fixed priors increase with noise as expected.[33]

Under setting 'B', the employed models should rather be understood as approximations, while uncovering a 'true' model is of minor importance. The results exemplify once again the supermodel effect behavior of fixed prior settings illustrated in Figures 4 and 5. Small degrees of noise trigger a concentration of posterior mass under the hyper-g prior and the empirical Bayes approach. An increase in noise is reflected in a wider spread of posterior mass among models under flexible priors, whereas fixed priors still concentrate on a small number of models. Moreover, note how close the hyper-g results are to the Empirical Bayes prior under both settings 'A' and 'B', which are particularly striking when noise is small. This illustrates that both concepts are not only related asymptotically, but also lead to similar conclusions under small samples (that are characterized by a noise component that is not too large).

Finally, we examine the robustness of the various g-prior settings via their performance in out-of-sample prediction. Results from a prediction exercise are expected to vary considerably between fixed and flexible prior settings, since the latter incorporate data adaptive shrinkage. Akin to Liang et al. (2008) we randomly split the data from settings 'A' and 'B' into 70 estimation and 30 out-of-sample observations. We then calculate the root mean squared error (RMSE) of the forecasts for the 30 out-of-sample data points, averaged over 50 Monte Carlo steps. The RMSE statistics shown in Table 8 are normalized with respect to forecasting results under the g-RIC prior. Thus values below 1 indicate better predictive accuracy of the respective prior structure than under the g-RIC prior. The top panel of Table 8 shows mixed results for setting 'A'. As expected, the g-RIC prior with its focus on parsimonious models excels in nearly all signal-to-noise settings, concentrating on a single (and luckily the correct data-generating) model. In the $\sigma = 1/2$ case, however, the flexible priors concentrate mass even more tightly than does the g-RIC and consequently yield better predictions in terms of RMSE. At intermediate noise levels, g-RIC outperforms the other priors by greater margins by exploiting the comparative advantage that the data-generating process is composed of a single model. This shows *how the supermodel effect could be exploited* to achieve superior predictive performance under the following conditions: If a researcher has prior knowledge that the data is generated by a simple model with few covariates, and if the noise component is neither too weak such that flexible priors would outperform any fixed priors, nor too strong such that a fixed prior setting would hardly identify the one 'true' model as the best-performing one, then a high g-parameter will concentrate more mass on the 'correct' model than flexible prior settings. In view of this conditions, the supermodel effect can be exploited best in a typical simulation set-up. However, such an approach can be dangerous under more complex datasets, and in any case when such peculiar prior knowledge is not given. Simulation setting 'B' illustrates the merits of flexible priors when the lack of ideal conditions turns the supermodel effect into a disadvantage for fixed g-priors: The predictive performance of flexible priors dominates throughout nearly all signal-to-noise settings. Especially the HG-3 prior and the empirical Bayes approach demonstrate superior predictive abilities with the latter one outperforming the g-RIC prior for all signal-to-noise setups. This contrasts with typical simulation exercises in the literature, as their data-generating processes emanates from a parsimonious single models, which plays in favor of (large) fixed priors because of the supermodel effect. Given their focus on large models, it is surprising how well hyper-g priors

[33] We have omitted results from the HG-3 setting, since results are very similar to that of HG-4.

perform in this prediction exercise. The results suggest that imposing a model prior that favors parsimonious models yields even better predictive performance under a flexible prior.[34]

6 (Un)stable growth determinants

Emanating from the pioneering work of Sala-i-Martin et al. (2004),[35] numerous studies have employed model averaging techniques in the empirics of economic growth. For comparability, we follow this approach and investigate the effect of the prior settings from section 5 on inference in a cross-country growth data set. A vast part of the empirical studies use international income data provided by the Penn World Tables (PWT). This data base publishes GDP data adjusted for purchasing power parities (PPP), which is essential for conducting country comparison studies. The core purpose of the PWT is collecting prices for the same or similar goods in different countries. Gathering prices is carried out on an irregular basis with each 'generation' of the table complying with a different round of price collection (Johnson et al., 2009). The methodologies employed by PWT to derive PPP-adjusted GDP data - and in particular growth rates thereof - have been frequently criticized as being plagued by considerable measurement error. Johnson et al. (2009) carry out a replication exercise by re-estimating selected growth equations employed in the literature using different versions of PWT. They show that estimates vary markedly across different versions of the PWT. In particular, growth studies using high frequency data (e.g. annual as opposed to long run averages) are especially prone to the measurement error inherent to the PWT methodology. Johnson et al. (2009) conclude that in order to make significant policy conclusions empirical results should be robust to PWT revisions.

Ciccone and Jarociński (2010) investigated the impact of PWT revisions on Bayesian model averaging results. They use three different versions of the Penn World Table income data (PWT 6.0, PWT 6.1 and PWT 6.2) and show that the identification of 'robust' determinants - as measured by the respective PIP - varies tremendously among data revisions. Feldkircher and Zeugner (2012) argue that this instability is partially rooted in the prior setup chosen by Ciccone and Jarociński (2010): a uniform prior on the model space is coupled with the g-RIC prior, with the latter being characterized by the 'supermodel effect'.

In what follows we build on this example and estimate cross-country growth regressions for an extended set of four PWT revisions:

$$\Delta y^j = \alpha + \gamma y_0^j + \vec{\beta}_s X_s + \varepsilon, \tag{11}$$

Here, Δy^j denotes the average annual growth of income per capita over the period from 1960 to 1996 for $N = 75$ countries, α the intercept, ε the error term, and $\mathbf{X}_s = (\mathbf{x}_1 \dots \mathbf{x}_s)$ a matrix whose columns represent a subset s of explanatory variables. Initial income is denoted by y_0^j and is the only explanatory variable that changes with PWT vintages. The potential growth determinants (whose combinations are represented by $\mathbf{X}_s$) are the ones originally put forward in Sala-i-Martin et al. (2004) and employed in Ciccone and Jarociński (2010). These 66 variables comprise measures for factor accumulation and convergence (as implied by the Solow growth model), human capital, institutional environment, and socio-geographical determinants. The estimation is carried out for each of the four considered PWT revisions, indexed by $j \in \{\text{PWT 6.0}, \text{PWT 6.1}, \text{PWT 6.2 and PWT 6.3}\}$. Note that all of them stem from the same PWT generation and are thus based on the same raw

[34] Tentative results available from the authors on request.

[35] See also Crespo Cuaresma and Doppelhofer (2007) and Sala-i-Martin (1997) for traditional approaches to model averaging as opposed to Fernández et al. (2001b) and Eicher et al. (2011) for Bayesian strategies.

price data. In this setting, we examine the effect of small perturbations to income data on posterior results under different g priors. The correlation of the dependent variable between vintages ranges from 0.92 to 0.97 and thus the revisions can be considered as reasonably small perturbations. In order to represent loose prior expectations about model size, all estimations are based on the Ley and Steel (2009) binomial-beta model prior anchored at an expected model size of $K/2$ variables.[36]

We compare the stability of the posterior inclusion probabilities over revisions by the ratio of maximum over minimum PIP for a variable: $\text{Max/Min}_k = \max(PIP(X_k^j))/\min(PIP(X_k^j))$, with $k = 1, \ldots, 67$ and $j = 1, \ldots, 4$ denoting the four PWT data sets. Note that for all PWT vintages, the g-RIC prior complies with the 'benchmark' prior put forward in Fernández et al. (2001a). Moreover, note that in Ciccone and Jarociński (2010) the country sample - and thus the number of observations - changes from revision to revision. Feldkircher and Zeugner (2012) show that conditioning on the same observations throughout the revisions reduces the instability of posterior results. We therefore opt for holding samples constant since we are interested in the part of PIP instability that is caused by employing the fixed g-RIC prior (the benchmark prior). Table 10 summarizes the results. Employing a flexible prior decreases PIP variation to a great extent, regardless of which PWT vintages are considered: The overall Max/Min ratios under a flexible prior are 67% smaller than under the fixed g-RIC prior.

This suggests that posterior results of the fixed and flexible priors will also differ considerably in qualitative terms. To further examine the robustness of growth determinants, Table 11 lists the posterior inclusion probabilities per variable under the g-RIC and the hyper-g (UIP) prior for all four PWT vintages. Explanatory variables are labeled 'robust' when they exceed a PIP of 0.50. This threshold can be motivated from a predictive stance (Barbieri and Berger, 2003) as well as from an intuitive perspective. Under the g-RIC prior, only a proxy for human capital (primary schooling in 1960) can be considered as robust in all four PWT vintages. Results are unstable for regional dummy variables such as a dummy for East Asia, Tropical Area, and Latin America. Moreover, no clear-cut results emerge on the effect of initial GDP on international differences in income. This is particularly worrisome from the viewpoint of economic theory and casts doubts on the results under the g-RIC setting. The hyper-g (UIP) prior, in contrast, identifies several variables as robust. Both initial GDP and primary schooling display impressive posterior support ($PIP \geq 90\%$) and their posterior coefficients are well in line with economic theory.[37] Furthermore, the regional dummy for Africa, and variables for the share of Confucian population receive robust posterior support over all PWT revisions, which provides strong empirical evidence in the vein of Johnson et al. (2009). Finally, the proxies for fertility and for Buddhism show considerable posterior support in three out of the four PWT vintages. The robustness of regional dummy variables points to heterogeneous growth dynamics in the examined country set.

The difference in posterior results between flexible and fixed priors is rooted in how posterior mass is distributed among covariates and models. For the PWT data set, this can be seen best when comparing the posterior expected model size. While a model for economic growth should be expected to contain 3 to 5 explanatory variables according to the g-RIC prior, the posterior model size for flexible priors lies in the range of 12 to 14 regressors. Accordingly, the corresponding shrinkage factors of the flexible priors are considerably lower (across all revisions) than the value implied by the fixed g-RIC prior (see the bottom of Table 11). These differences become even more striking when considering the implied posterior mean for g, which is 4448 under the g-RIC compared to 23 - 26 under the hyper-g (UIP) prior. The smaller shrinkage factors (and implied values for g) under the flexible priors indicate noise to prevail more strongly under PWT data sets than

[36] Note that Ciccone and Jarociński (2010) have used a uniform prior on the model space. Results under the uniform prior are available from the authors upon request. See also Feldkircher and Zeugner (2012).

[37] Results are available from the authors upon request.

inherently assumed under g-RIC. With that in mind, it comes as no surprise that the flexible priors push the PIPs of most variables towards the 50% threshold, which points to empirical evidence neither in favor nor against their inclusion in models of economic growth. The inconclusiveness that plagues the empirical growth literature is thus more correctly mirrored in the posterior results of the flexible priors, whereas the benchmark prior is prone to the risk of over-fitting the PWT data sets.

Figure 6 (upper left panel) illustrates the problems arising from fixed g priors in the (PWT 6.1) data set. The concentration of posterior model probabilities under flexible priors is considerably lower than under the g-UIP setting, and is far less than under the g-RIC setting. In view of the discussion in section 3, this points to the supermodel effect severely affecting the posterior statistics in this popular data set. This characteristic is also mirrored in the corresponding posterior inclusion probabilities (Figure 6, upper right panel). Under fixed-g settings (particularly g-RIC), the PIPs are more skewed than under flexible priors, which again illustrates that fixed-g priors risk discriminating more among covariates than seems justified by the data at hand. Moreover, figure 6 displays virtually identical results for the various hyper-g settings, which illustrates the feature that their differing prior expectations do not have too much impact on posterior statistics. Finally, their results are indistinguishable from the Empirical Bayes prior, which further illustrates how quickly hyper-g and EBL prior results converge in small samples.

The above results illustrate the advantages of flexible g priors in BMA inference. However, it is not clear whether these advantages carry over to robustness in predictive performance: On the one hand, the instability of of fixed priors such as the g-RIC and g-UIP should deteriorate forecast performance. On the other hand, it is a stylized fact in the forecasting literature that parsimonious models outperform saturated models in terms of forecasting quality. It is thus not clear a priori which prior setting excels in forecasting economic growth. We therefore conduct a forecast evaluation in which we randomly split the observations into a training sample of 56 countries and a 'hold-out' sample of 19 countries. The training sample is used to estimate the models for forecasting the remaining observations of the hold-out sample. The corresponding root mean square error (RMSE) is calculated over 30 random sample splits. Table 8 summarizes the results: All flexible prior settings outperform the fixed priors g-RIC and g-UIP. In particular, the predictive performance of the hyper-g (UIP) prior is superior to the other priors in most of the PWT revisions.

Finally, note that the posterior expected shrinkage factor, an indicator for goodness-of-fit, did not steadily increase from PWT 6.0 to PWT 6.3. We notice that goodness-of-fit and data quality need not to go hand in hand.[38] However, we think a goodness-of-fit measure might be a reasonable indicator to get a first impression about how the data revision progressed. The sharp drop of the shrinkage factor from PWT 6.1 to PWT 6.2 might further stimulate the debate whether newer is always better in the context of PWT revisions.[39]

7 Concluding remarks

The widespread use of Zellner's g prior in linear BMA rests on two convenient features: it provides closed-form solutions and reduces the complexity of prior elicitation to the scalar g. Consequently, theoretical considerations have mostly focused on the choice of g, in particular in view of its

[38] An increase in goodness-of-fit might be driven by stronger correlation of measurement error in both, explanatory variables and dependent variable.

[39] See also Johnson et al. (2009).

virtues as a penalty term for model size. This study departs from earlier literature in bringing forward two arguments that have been overlooked so far: First, model size considerations should be disentangled from the primary purpose of g (which is scaling coefficient covariance) and rather be fused into the formulation of model priors. The elicitation of g should thus not interfere with prior desiderata on model size. Second, we demonstrate that fixing g to arbitrary values may have unintended consequences on posterior model probabilities: The higher g, the more tightly posterior mass will concentrate on the few best-performing 'supermodels' – regardless of model sizes, number of observations, or signal-to-noise ratios. Ultimately, a large value for g will favor a single model, thereby emulating model selection rather than model averaging. As previous studies predominantly have assessed BMA performance on simulated data generated by a single model, they tended to favor g-specifications with large values of g that effectively *select* the right model. However, in empirical practice a large g runs the risk of putting to much posterior weight on a single model. We demonstrate that the popular prior suggested by Fernández et al. (2001a) is particularly prone to this behavior.

As it is virtually impossible to specify the 'right' value for g under unknown variance, we propose to put a prior distribution on this parameter instead: Such a hyperprior allows for data-adaptive and model-specific shrinkage, thus adjusting the impact of prior beliefs to data quality. In discriminating among models only as far as data quality allows, a prior on g thus offers a remedy for the supermodel effect. In this manner, we focus on a particular *hyper-g* prior, whose formulation offers three main advantages: First, it admits closed form solutions for almost any quantity of interest, thereby facilitating implementation. Second, its hyperparameter allows for formulating prior beliefs on coefficient variance, but without incurring the risk of unintended consequences on posterior model mass. Third, we demonstrate that the hyper-g prior can be reconciled with BMA consistency. We complement the existing literature on the hyper-g prior by providing additional posterior expressions that allow for fully Bayesian inference, as well as for sound numerical implementation. Moreover, we demonstrate that its posterior statistics can be considered as a goodness-of-fit indicator, and show why its results are closely related to those of the Empirical Bayes g-prior.

A simulation exercise contrasts various formulations of fixed and hyper-g priors. The fixed g-priors perform well when the data-generating process rests on a single model that is part of the candidate model space – but so does the hyper-g prior. The virtues of flexible prior structures become evident in more complex settings: Flexible priors outperform fixed g settings in terms of forecasting accuracy and exhibit a more stable structure of posterior model and inclusion probabilities over varying degrees of noise in the data.

The final section illustrates these conclusions by investigating fixed and flexible priors under different revisions of an economic growth data set. The results demonstrate that fixing g has a detrimental effect on the stability of posterior results. While fixed g-priors list initial income among the most unstable growth determinants, the estimates from the flexible priors are well in line with economic theory: Both conditional income convergence and human capital are identified as robustly related to income growth, along with a handful of regional dummies. In contrast to fixed settings, these results are insensitive to data revisions. Concluding, the hyper-g prior offers a sound, fully Bayesian approach that features the virtues of prior input and predictive gains without incurring the risk of mis-specification.

References

Abramowitz, M. and Stegun, I. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards Applied Mathematics Series 55. Tenth Printing.

Barbieri, M. M. and Berger, J. O. (2003). Optimal Predictive Model Selection. *Ann. Statist.*, 32:870–897.

Brown, P., Vannucci, M., and Fearn, T. (1998). Multivariate Bayesian Variable Selection and Prediction. *Journal of the Royal Statistical Society B*, 60:627–641.

Chipman, H., George, E., and McCulloch, R. (2001). *The Practical Implementation of Bayesian Model Selection*. Institute of Mathematical Statistics Lecture Notes-Monograph Series Vol. 38. Beachwood, Ohio.

Ciccone, A. and Jarociński, M. (2010). Determinants of Economic Growth: Will Data Tell? *American Economic Journal: Macroeconomics*, 2:222–46.

Claeskens, G. and Hjort, N. (2008). *Model Selection and Model Averaging*. Cambridge University Press.

Crespo Cuaresma, J. and Doppelhofer, G. (2007). Nonlinearities in Cross-Country Growth Regressions: A Bayesian Averaging of Thresholds (BAT) Approach. *Journal of Macroeconomics*, 29:541–554.

Cui, W. and George, E. (2008). Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference*, 138:4:888–900.

Eicher, T. S., Papageorgiou, C., and Raftery, A. E. (2011). Default priors and predictive performance in bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics*, 26(1):30–55.

Eklund, J. and Karlsson, S. (2007). Forecast Combination and Model Averaging using Predictive Measures. *Econometric Reviews*, 26:329–362.

Feldkircher, M. and Zeugner, S. (2012). The Impact of Data Revisions on the Robustness of Growth Determinants – A Note on 'Determinants of Economic Growth. Will Data Tell?'. *Journal of Applied Econometrics*, (forthcoming).

Fernández, C., Ley, E., and Steel, M. F. (2001a). Benchmark Priors for Bayesian Model Averaging. *Journal of Econometrics*, 100:381–427.

Fernández, C., Ley, E., and Steel, M. F. (2001b). Model Uncertainty in Cross-Country Growth Regressions. *Journal of Applied Econometrics*, 16:563–576.

Foster, D. P. and George, E. I. (1994). The Risk Inflation Criterion for Multiple Regression. *The Annals of Statistics*, 22:1947–1975.

Gelfand, A. and Dey, D. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 501–514.

Gelman, A., Carlin, J. B., Stern, S. H., and Rubin, B. D. (1995). *Bayesian Data Analysis*. Chapman & Hall.

George, E. and Foster, D. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747.

Giannone, D., Lenza, M., and Primiceri, G. (2012). Prior selection for vector autoregressions. *CEPR Discussion Paper*, 8755.

Guptar, A. K. and Nagar, D. K. (2000). *Matrix Variate Distributions*. Chapman & Hall, CRC, Monographs and Surveys in Pure and Applied Mathematics, 104.

Hansen, M. and Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14, No. 4:382–417.

Johnson, S., Larson, W., Papageorgiou, C., and Subramanian, A. (2009). Is Newer Better? Penn World Table Revisions and Their Impact on Growth Estimates. *Center for Global Development, Working Paper 191*.

Kass, R. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90:928–934.

Koop, G. and Potter, S. (2003). Forecasting in Large Macroeconomic Panels Using Bayesian Model Averaging . *FRB NY Staff Report*, 163.

Laud, P. and Ibrahim, J. (1995). Predictive model selection. *Journal of the Royal Statistical Society*, Series B 57:247–262.

Ley, E. and Steel, M. F. (2009). On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regressions. *Journal of Applied Econometrics*, 24(4):651–674.

Ley, E. and Steel, M. F. (2011). Mixtures of g-priors for bayesian model averaging with economic applications. Policy Research Working Paper 5732, World Bank.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g Priors for Bayesian Variable Selection. *Journal of the American Statistical Association*, 103:410–423.

Masanjala, W. H. and Papageorgiou, C. (2008). Rough and Lonely Road to Prosperity: A Reexamination of the Sources of Growth in Africa Using Bayesian Model Averaging. *Journal of Applied Econometrics*, 23:671–682.

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, 25:111–163.

Sala-i-Martin, X. (1997). I Just Ran 2 Million Regressions. *American Economic Review*, 87:178–183.

Sala-i-Martin, X., Doppelhofer, G., and Miller, R. I. (2004). Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach. *American Economic Review*, 94:813–835.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6(2):461–464.

Strachan, R. and van Dijk, H. (2004). Exceptions to Bartletts Paradox. *Keele Economic Research Papers*.

Zellner, A. (1986). *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, chapter On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions. North-Holland: Amsterdam.

Zellner, A. (2008). Comments on Mixtures of g-priors for Bayesian Variable Selection by F. Liang, R. Paulo, G. Molina, MA Clyde and JO Berger.

A Technical appendix

A.1 Posterior statistics of the hyper-g prior: Further results

Joining tedious algebra with Gauss' contiguous relations for hypergeometric functions(Abramowitz and Stegun, 1972, p.563) allows to establish important posterior expressions of the hyper-g prior in closed form, on top of the ones provided in section 4.1. It is important to note that the posterior moments of coefficients, the shrinkage factor, and the predictive distribution all arise as simple transformations of the posterior model likelihood in equation (5). Their implementation thus bears virtually no computational cost. The following equations present the most important posterior statistics (keeping the notation from section 4.1). Equation (A.1) expresses the posterior second moment of the shrinkage factor if $(R_s^2 \in (0,1))$:[40]

$$\begin{aligned} \mathrm{E}\left(\left(\frac{g}{1+g}\right)^2 \middle| y, X_s, M_s\right) &= \frac{1}{(R_s^2)^2(\bar{N}-\bar{\theta}_s)(\bar{N}-(\bar{\theta}_s+2))} \times \\ \times \left(\left((\bar{N}-2)R_s^2-(\bar{\theta}_s+2)\right)\frac{\bar{\theta}_s}{F_s^*} + \left(\bar{N}R_s^2-\bar{\theta}_s\right)^2 - 2\left(\bar{N}(R_s^2)^2-\bar{\theta}_s\right)\right) \end{aligned} \tag{A.1}$$

In addition to posterior moments, the posterior distribution of coefficients $\beta_s|y, M_s$ can also be established in closed form, but as ratio of two hypergeometric functions:

$$\begin{aligned} p(\beta_s|y,X_s,M_s) &= \int_0^\infty p(\beta_s|y,X_s,M_s,g)p(g|y,X_s,M_s)dg = \\ = \frac{\Gamma\left(\frac{N-1+k_s}{2}\right)\Gamma\left(\frac{k_s+a}{2}\right)\frac{N-1}{2}}{\Gamma\left(\frac{N-1+k+a}{2}\right)} &\frac{\sqrt{|X_s'X_s|}}{\pi^{\frac{k_s}{2}}}(\breve{y}'\breve{y})^{\frac{N-1}{2}}(\beta_s'X_s'X_s\beta_s)^{-\frac{N-1+k_s}{2}} \times \\ \times \frac{{}_2F_1\left(\frac{N-1+k_s}{2}, \frac{N-1}{2}+1, \frac{N-1+k_s+a}{2}, 1-\frac{(y-X_s\beta_s)'(y-X_s\beta_s)}{\beta_s'X_s'X_s\beta_s}\right)}{{}_2F_1\left(\frac{N-1}{2}, 1, \frac{k_s+a}{2}, R_s^2\right)} & \end{aligned} \tag{A.2}$$

Note that this expression is of close, though not perfect resemblance to a hypergeometric function distribution of type II.[41] The first two moments of this distribution are provided in section 4.1.

Similar algebra also allows for simplifying the predictive distribution and yields its moments in closed form: Consider using the data (X, y) to forecast the dependent variable $\hat{y}$ conditional on 'prediction' covariates $\hat{X}$. Let X be an $N \times k$ matrix, y be $N \times 1$, while $\hat{y}$ is $l \times 1$ and $\hat{X}$ $l \times k$. The posterior predictive distribution of $\hat{y}$ is then given as a a multivariate t-distribution of dimension l (Eklund and Karlsson, 2007, equation (A.15))[42]

$$\begin{aligned} \hat{y}|\hat{X}, X, y, g, M_s &\sim t_l(\bar{y} + s\hat{X}\hat{\beta}, \Sigma, N-1) \\ \text{where } \Sigma &= \left(I_l + s\hat{X}(X'X)^{-1}\hat{X}'\right)\frac{\breve{y}'\breve{y}}{N-1}(1-sR_s^2) \end{aligned}$$

[40] In case $R_s^2 = 0$ (which relates in particular to the null model), the expression for (A.1) is $\mathrm{E}\left(\left(\frac{g}{1+g}\right)^2 \middle| y, X_s, M_s\right) = \frac{8}{(k_s+a)(k_s+a+2)}$

[41] See Guptar and Nagar (2000) for the exact definition of the type II hypergeometric distribution.

[42] The slight differences with respect to Eklund and Karlsson (2007) are due to the fact that we employ an improper prior on beta variance σ and the constant.

Here, s denotes the shrinkage factor $s = \frac{g}{1+g}$, and R_s^2 the (centered) R-squared of y on X. $\bar{y}$ denotes an N-dimensional vector whose elements are the arithmetic mean of y, and $\breve{y} \equiv y - \bar{y}$ the centered response variable. Integrating the density function of $\hat{y}|\hat{X}, X, y, g, M_s$ with respect to the shrinkage factor yields the integrand of the following equation (after some rearrangement):

$$f(\hat{y}|\hat{X}, X, y, M_s) = \frac{\Gamma\left(\frac{N-1+l}{2}\right)}{\Gamma\left(\frac{N-1}{2}\right)\pi^{\frac{l}{2}}} \frac{\overbrace{(k+a-2)((\breve{y}'\breve{y})^{\frac{N-1}{2}})}^{\frac{a-2}{2}\frac{1}{p(y|X)}}}{2F_s^*} \times$$
$$\times \int_0^1 \left|I_l + s\hat{X}(X'X)^{-1}\hat{X}'\right|^{-\frac{1}{2}} (1-s)^{\frac{k+a-4}{2}} \times$$
$$\times \left(\breve{y}'\breve{y}(1-sR_s^2) + (\hat{y} - \bar{y} - s\hat{X}\hat{\beta})' \left(I_l + s\hat{X}(X'X)^{-1}\hat{X}'\right)^{-1} (\hat{y} - \bar{y} - s\hat{X}\hat{\beta})\right)^{-\frac{N-1+l}{2}} ds$$

To our knowledge, there is no closed-form solution to to the integral above, neither to its Laplace approximation. We therefore recommend to resort to numerical integration. Nonetheless, it is possible to obtain the predictive variance, i.e. the squared predictive standard error, as:

$$Var(\hat{y}|y, X, \hat{X}, M_s) = \hat{X}Var(\beta|y, M_s)\hat{X}' +$$
$$+ \frac{\breve{y}'\breve{y}}{N-3}\frac{N+l}{N} I_l\left(\frac{(N-3)(1-R_s^2)}{N-1-k-a} - \frac{k+a-2}{N-1-a-k}/F_s^*\right)$$

A.2 Consistency of the hyper-g prior

Fernández et al. (2001a) define asymptotic 'consistency' as follows: Consider that only Model M_s is true, while all other models $M_j \neq M_s$ are not true. Consistency then requires:

$$\operatorname*{plim}_{n\to\infty} p(M_s|y, X_s) = 1 \quad \text{and} \qquad \operatorname*{plim}_{n\to\infty} p(M_j|y, X_s) = 0 \quad \forall M_j \neq M_s$$

Liang et al. (2008, Appendix B) have proven the above for the hyper-g prior except for the case where the true model M_s is the null model M_0. They stop short their proof because in this case the Bayes factor $B(M_j : M_0)$ is (Liang et al., 2008, p.423):

$$\frac{p(M_j|y, X_s)}{p(M_0|y)} \geq \int_0^\infty (1+g)^{-\frac{k_j}{2}}\, p(g)dg \tag{A.3}$$

Moreover they state that if the above integral vanishes as $N \to \infty$, then consistency is ensured. Applying the hyper-g setting transforms the right-hand side in (A.3) into the following (by $a > 2$):

$$\int_0^\infty (1+g)^{-\frac{k_j}{2}}\, p(g)dg = \frac{a-2}{2}\int_0^1 (1+g)^{-\frac{k_j+a}{2}} dg = \frac{a-2}{k_j+a-2}$$

If $a = 2 + w(N)$ with $w(N) > 0$ and $\lim_{N\to\infty} w(N) = 0$, then the integral vanishes and thus concludes the proof.

A.3 Relationship between hyper-g and Empirical Bayes prior

It is well established in Bayesian statistics that under any non-degenerate prior, Bayesian regression results asymptotically approach their maximum-likelihood equivalent with increasing sample

size. Against this backdrop it is not surprising that the results under the Empirical Bayes prior in section 5 are close to the hyper-g settings. This sections outlines why posterior model probabilities under consistent hyper-g priors and the Empirical Bayes are close in small samples and converge asymptotically. As a byproduct, this section demonstrates the asymptotic consistency of the Empirical Bayes prior if the ' true' model is not the null model. The results in this section are based on Laplace approximations[43], cf. Gelfand and Dey (1994) for their asymptotic properties in the context of Bayesian model selection.
Consider the familiar form of the Laplace approximation, where $\hat{\theta}$ is the maximizer of the integrand's logarithm $h(\theta)$:

$$\int_{\Theta} \exp(h(\theta))d\theta \approx \sqrt{\frac{2\pi}{-h''(\hat{\theta})}} \exp h(\hat{\theta})$$

Consider in turn the Bayes Factor for the hyper-g prior formulation as in (5), between a model with k covariates, and the null model:

$$BF_h = \frac{a-2}{2} \int_0^{\infty} (1+g)^{\frac{N-1-k-a}{2}} (1+g(1-R^2))^{-\frac{N-1}{2}} dg$$

Letting

$$h(g) = \frac{1}{2}\left((N-1-k-a)\log(1+g) - (N-1)\log(1+(1-R^2)g)\right)$$

yields the maximizer:

$$\hat{g} = \max(\frac{R^2(N-1-k-a)}{(1-R^2)(k+a)} - 1, 0)$$

where $\hat{g} = 0$ if and only if $k + a \geq R^2(N-1)$. Liang et al. (2008, p.421) note the similarity to the local Empirical Bayes (EBL) estimator of g, but abstain from further investigating the issue.
The second derivative of $h(g)$ is given as

$$h''(g) = \frac{1}{2}\left(-\frac{N-1-k-a}{(1+g)^2} + \frac{(N-1)(1-R^2)^2}{(1+(1-R^2)g)^2}\right)$$

The Bayes factor under a hyper-g prior is thus approximately equal to:

$$BF_h \approx (a-2)\sqrt{\frac{\pi}{\frac{N-1-k-a}{(1+\hat{g})^2} - \frac{(N-1)(1-z)^2}{(1+(1-z)\hat{g})^2}}} (1+\hat{g})^{\frac{N-1-k-a}{2}} (1+\hat{g}(1-R^2))^{-\frac{N-1}{2}}$$

In case we have $\hat{g} > 0$, then algebraic manipulation of the expression above yields:

$$BF_h \approx (a-2)\sqrt{\pi}\sqrt{\frac{N-1}{(N-1-k-a)(k+a)}} \left(\frac{R^2}{(1-R^2)} \frac{N-1-k-a}{k+a}\right)^{-\frac{k+a-2}{2}} \left(\frac{(1-R^2)(N-1)}{N-1-k-a}\right)^{-\frac{N-1}{2}}$$

Now consider the equivalent null-based model Bayes Factor for the EBL approach which is:

$$BF_{EBL} = \left(\frac{R^2}{1-R^2} \frac{N-1-k}{k}\right)^{-\frac{k}{2}} \left(\frac{(1-R^2)(N-1)}{(N-1-k)}\right)^{-\frac{N-1}{2}}$$

in case if $k \leq R^2(N-1)$

[43]Note that due to perceived numerical difficulties, Liang et al. (2008) propose the use of a Laplace approximation for the posterior model likelihood under the hyper-g distribution (Liang et al. (2008, equation (17))). Depending on the data, Laplace approximations can be prone to substantial numerical inaccuracies in small samples. However, they may be useful for the purpose of this section which is mainly interested in asymptotic results.

Therefore, if $k + a \leq R^2(N-1)$:

$$BF_h \approx (a-2)\sqrt{\pi}\sqrt{\frac{N-1}{(k+a)(N-1-k-a)}}\left(\frac{(1-z)}{z}\frac{k+a}{N-1-k-a}\right)^{\frac{a-2}{2}}\left(\frac{k+a}{k}\right)^{\frac{k}{2}}\left(\frac{N-1-k-a}{N-1-k}\right)^{\frac{N-1-k}{2}} BF_{EBL}$$

So if $a \to 2$,[44] the hyper-g Bayes Factor is approximately equivalent to an EBL Bayes factor times a k-based model prior (that does not depend on R_s^2). Moreover, this model prior is bounded in a relatively narrow range: Note that

$$(1+a)^{-a/2} \leq \left(\frac{k+a}{k}\right)^{\frac{k}{2}}\left(\frac{N-1-k-a}{N-1-k}\right)^{\frac{N-1-k}{2}} < 1$$

The upper bound follows from the fact that $((k+a)/k)^{k/2} = (1+\frac{a}{2}/\frac{k}{2})^{\frac{k}{2}} < \exp(\frac{a}{2})$. Similarly $((N-1-k-a)/(N-1-k))^{\frac{N-1-k}{2}} < \exp(-\frac{a}{2})$. Setting $k=1$ and letting $N-1 \geq k+a+1$ performs the lower bound.[45] The effect of the term in square roots actually counters the impact of the latter term, as $\sqrt{\frac{4}{N-1}} \leq \sqrt{\frac{N-1}{(N-1-k-a)(k+a)}} \leq 1$ for $k+a < N-1$. The 'model prior' thus results in an upweighting of models with few or many coefficients, while intermediate model sizes are downweighted (a feature very similar to the model prior of Ley and Steel (2009)). Since the model prior does not depend on the level of N, it will lose importance as $N \to \infty$. In the limit, therefore, both EBL and consistent hyper-g will approach the same Bayes Factors between any model except the null model. If the true model is not the null model, then the posterior model probabilities under EBL will therefore approach those under a consistent hyper-g – which establishes the asymptotic consistency of the EBL prior in the sense of Fernández et al. (2001a), provided the true model is not the null model.

Moreover, note that even in small samples, the impact of the k-based 'model prior' is virtually negligible with respect to the importance of BF_{EBL}. Thus, at least as long as $R^2(N-1) > k+a$, BF_h is quite close to BF_{EBL}. And as long as the signal-to-noise ratio in the data is not too small, BMA posterior statistics will be disproportionally based on models with large PMPs (i.e., models with $(N-1)R^2 \gg k+a$). Any models with large differences between BF_h and BF_{EBL} will thus hardly affect posterior model probabilities. Finally, note that in this case models with high PMPs will display very hypergeometric terms $F_s^* \gg 1$, which renders the posterior moments from sectio 4.1 very close to their EBL equivalents. This effect explains why this paper's results under hyper-g and EBL are so close.

A.4 The shrinkage factor and goodness-of-fit

In order to demonstrate inequalities (9) and (10), consider a reformulation of the posterior expected value of the shrinkage factor (7), where p_s is shorthand for posterior model probability $p(M_s|y,X)$ of model M_s (and p_0 denotes the PMP of the null model).[46]

$$\mathrm{E}\left(\frac{g}{1+g}\middle| y, X\right) = \epsilon + \sum_{j=1}^{2^K} p_s \frac{\bar{N}R_s^2 - \bar{\theta}_s}{R_s^2(\bar{N}-\bar{\theta}_s)} + p_0\frac{2}{a} \tag{A.4}$$

$$\text{where} \quad \epsilon = \sum_{j=1}^{2^K} p_s \frac{\bar{\theta}_s}{F_s^* R_s^2(\bar{N}-\bar{\theta}_s)}$$

[44] Recall that any consistent hyper-g prior requires $a \to 2$ for $N \to \infty$.

[45] Note that if $k=0$, $BF_{EBL} = BF_h = 1$.

[46] Note that this formulation assumes R_s^2 for all models other than the null model to be strictly larger than zero – a notion we will follow throughout this section. However, the following inequalities can be easily generalized to the case of $R_s^2 = 0$ as long as the the full model has $R_F^2 > 0$.

Equation (A.4) can be transformed into the following:

$$1 - \mathrm{E}\left(\tfrac{g}{1+g}\middle| y, X\right) + \epsilon - \tfrac{a-2}{a}p_0 = \sum_{j=1}^{2^K} p_s \frac{\bar{\theta}_s(1-R_s^2)}{(\bar{N}-\bar{\theta}_s)R_s^2} \tag{A.5}$$

The ϵ term is based on the expression $\bar{\theta}_s/F_s^*$ in (7), whose only role is to keep $E(\frac{g}{1+g}|X,y)$ non-negative in case of a 'bad' model, whereas it rapidly vanishes for models with higher signal-to-noise ratios. For any setting of the hyperparameter a, ϵ vanishes as $N \to \infty$ for fixed K as long as the null model is not the single 'true' model – but even in small samples, ϵ tends rapidly towards zero as data quality increases. Using a consistent hyper-g prior such as HG-UIP, ensures that ϵ will asymptotically vanish even when the true model is the null model, as in this case R_s^2 will vanish with order of magnitude $O_T(\bar{N})$: Since $F_s^* \geq 1$, the model-specific expression is bounded from below while consistency will drive p_s to zero.

Likewise, under a consistent prior the requirement that $a \to 2$ as $\bar{N} \to \infty$ will induce the term $p_0(a-2)$ to vanish asymptotically.

That said, in small samples with any viable signal-to-noise ratio, any models with very low PMP will hardly affect posterior results, and hence the expression ϵ will vanish as soon as there exist some models with posterior model probabilities considerably larger than p_0 (which therefore must have their $F_s^* \gg 1$).

Omitting the expression $\epsilon - \frac{a-2}{a}p_0$ from (A.5) directly leads to inequality (9): Define $\hat{F}_s \equiv \frac{(\bar{N}-\bar{\theta}_s)R_s^2}{\bar{\theta}_s(1-R_s^2)}$, then (9) follows from applying Jensen's inequality to (A.5).

A similar argument applies to the demonstration of inequality (10). As long as $K + a < N + 1$, the following will hold by Jensen's inequality:

$$\sum_{s=1}^{K^2} p_s\big(\tfrac{1}{\bar{N}-\bar{\theta}_s}\big) \geq \frac{(1-p_0)^2}{\sum_{s=1}^{K^2} p_s\big(\bar{N}-\bar{\theta}_s\big)}$$

where $\sum_{s=1}^{K^2} p_s$ denotes the model probability-weighted sum over all models except the null model. Multiply with $\bar{N}$ and subtract $(1-p_0)$ to obtain

$$\sum_{s=1}^{K^2} p_s\big(\tfrac{\bar{\theta}_s}{\bar{N}-\bar{\theta}_s}\big) \geq (1-p_0)\frac{\sum_{s=1}^{K^2} p_s\bar{\theta}_s}{\sum_{s=1}^{K^2} p_s\big(\bar{N}-\bar{\theta}_s\big)}$$

Moreover, since any nested model's R-squared R_s^2 cannot exceed the R-squared of the full model R_F^2, we have that $\frac{1-R_s^2}{R_s^2} \geq \frac{1-R_F^2}{R_F^2}$ and therefore:

$$\sum_{s=1}^{K^2} p_s\big(\tfrac{1-R_s^2}{R_s^2}\tfrac{\bar{\theta}_s}{\bar{N}-\bar{\theta}_s}\big) \geq (1-p_0)\frac{\sum_{s=1}^{K^2} p_s\bar{\theta}_s}{\sum_{s=1}^{K^2} p_s\big(\bar{N}-\bar{\theta}_s\big)}\frac{1-R_F^2}{R_F^2}$$

Now note that $\sum_{s=1}^{K^2} p_s\bar{\theta}_s = E(\bar{\theta}|y,X) - p_0(a-2)$. Consistency lets $p_0(a-2)$ rapidly vanish with $N \to \infty$, which ensures that the right-hand side of this inequality is weakly larger than its version in expected value terms $E(\bar{\theta}_s|y,X) \equiv \sum_{s=0}^{K^2} p_s\bar{\theta}_s$:

$$(1-p_0)\frac{\sum_{s=1}^{K^2} p_s\bar{\theta}_s}{\sum_{s=1}^{K^2} p_s\big(\bar{N}-\bar{\theta}_s\big)}\frac{1-R_F^2}{R_F^2} \geq \frac{E\big(\bar{\theta}_s\big|y,X\big)}{E\big(\bar{N}-\bar{\theta}_s\big|y,X\big)}\frac{1-R_F^2}{R_F^2}$$

This establishes an upper bound for the left-hand side of (A.5) (recall that $\bar{N} \equiv N-3$ and $E(\bar{\theta}_s|y,X) = E(k_s|y,X)+a-2$):

$$\mathrm{E}\left(\tfrac{g}{1+g}\Big|y,X\right) - \epsilon + \tfrac{a-2}{a}p_0 \leq 1 - \frac{E\big(\bar{\theta}_s\big|y,X\big)}{E\big(\bar{N}-\bar{\theta}_s\big|y,X\big)}\frac{1-R_F^2}{R_F^2}$$

Under a consistent hyper-g prior, the term $\epsilon - \frac{a-2}{a}p_0$ will vanish asymptotically as $N \to \infty$, which establishes inequality (10). How close $E\big(\frac{g}{1+g}|y,X\big)$ comes to this upper bound is mainly determined by the posterior variance of model size (the less variance, the closer), and by parsimoniousness of the model priors.Note that the term ϵ on the left-hand side might break the inequality (10) in peculiar small samples. However, this term tends to be very small: Numerical simulations of a null hypothesis with varying N, K, a and standard deviations have yielded no single instance in which $R_F^2 > \frac{K+a-2}{N}$ and $E(\frac{g}{1+g}|y,X)$ larger than the right-hand side above. Therefore, if $R_F^2 > \frac{K+a}{N}$[47], then the ϵ term can be safely omitted from the inequality above.

A.5 Proof of Proposition 1

Given (y,X), define log-likelihood of a model M_γ as $\ell_\gamma \equiv \frac{k}{2}\log(1-s) - \frac{N-1}{2}\log(1-sz_\gamma)$ where $s = \frac{g}{1+g}$ is shorthand for the shrinkage factor, and z_γ denotes the R-squared of model M_γ. Moreover, index models according to their rank at a given s. Denoting the model the prior model probability of model M_γ by m_γ, the PMP of the best r models is given as:

$$PMP_r^* = \sum_{i=1}^{r}\exp(l_i)m_i \Big/ \sum_{i=1}^{2^K+1}\exp(l_i)m_i$$

Let $E_r(\theta)$ denote the posterior average of a statistic θ_i weighted according to r posterior model probabilities:

$$E_r(\theta) = \sum_{i=1}^{r}\exp(l_i)m_i\theta_i \Big/ \sum_{i=1}^{r}\exp(l_i)m_i$$

Moreover, write the posterior average over all models as $E(\theta) = E_{2^K+1}(\theta)$. Then, to prove proposition 1, it suffices to find the conditions under which the sign of the following derivative is positive:

$$\frac{d\log PMP_r^*}{ds} = E_r\left(\frac{\partial\ell}{\partial s}\right) - E\left(\frac{\partial\ell}{\partial s}\right)$$

LEMMA 1 *The posterior probability of the best model PMP_1^* has a non-negative derivative with respect to s if $k^* + \eta < E(k|y,X)$, where $E(k|y,X)$ represents posterior parameter size, and k^* the parameter size the best model; with $\eta > 0$ vanishing as $N \to \infty$, $s \to \infty$.*

[47]Note that this threshold is just slightly higher than the expected value of R_F^2 under the classic null hypothesis of no significant variance explanation by a regression model. As a rule of thumb, if the standard F-statistic for the full model is 'significant' by at least 20%, then the inequality above is guaranteed to hold.

PROOF Let $\frac{dPMP_1^*}{ds} < 0$, which implies

$$E(k) - k^* < (N-1)\left(\frac{1-z^*}{1-sz^*} - E\left(\frac{1-z}{1-sz}\right)\right)$$

The right-hand side could be positive or negative – although $E(k) > k*$ will in may cases be associated with a positive right-hand side. However, for constant N and $s \to \infty$ the right hand side vanishes rapidly which already proves one part of lemma 1.

Now that since $\exp(\ell_r^*)m_r^* \geq \exp(\ell_i)m_i \; \forall \; i \geq r$ we have that

$$\frac{sz_i}{1-sz_i} \leq \frac{(1-s)^{\frac{k_r-k_i}{N-1}}(\frac{m_r}{m_i})^{\frac{2}{N-1}}}{1-sz_r} - 1 \qquad \forall \; i \geq r \tag{A.6}$$

Therefore, $\frac{dPMP_1^*}{ds} < 0$ implies the following inequality:

$$E(k) - k^* < (N-1)\left(\frac{1-s}{s(1-sz^*)}\left(E((1-s)^{\frac{k^*-k}{N-1}}(\frac{m^*}{m})^{\frac{2}{N-1}} - 1)\right)\right)$$

Whether the right-hand side is positive or negative, depends on the actual distribution of parameter size k. Nonetheless, if $N \to \infty$ under constant s, the $N-1$ term in the exponent will dominate the $N-1$ factor in front, and the right-hand side will go to zero. Moreover, if s is a monotonically increasing function of N, the right-hand side will vanish even more rapidly. Defining η as the right-hand side of inequality (A.6) concludes the proof of lemma 1.

PROOF of proposition 1

Let be $(N-1)\left(E_{r-1}(\frac{1-z}{1-sz}) - E(\frac{1-z}{1-sz})\right)$ be bounded by η_{r-1}. We will have $\frac{dPMP_r^*}{ds} < 0$ if and only if

$$E(k) - E_r(k^*) < (N-1)\left(E_r\left(\frac{1-z^*}{1-sz^*}\right) - E\left(\frac{1-z}{1-sz}\right)\right) \tag{A.7}$$

Denote posterior model probability of model M_i by $p_i = \exp(\ell_i)m_i / \sum_{j=1}^{2^K+1}\exp(\ell_j)m_j$. If $\frac{1-z_r}{1-sz_r} \leq E_{r-1}(\frac{1-z}{1-sz})$, then the right hand side in (A.7) will be bounded by η_{r-1}

If $\frac{1-z_r}{1-sz_r} > E_{r-1}(\frac{1-z}{1-sz})$ then the bound η_{r-1} will not hold necessarily, but:

$$(N-1)\left(E_r\left(\frac{1-z}{1-sz}\right) - E\left(\frac{1-z}{1-sz}\right)\right) < (N-1)(1-s)\sum_{i=r+1}^{2^K+1} p_i\left(\frac{z_i}{1-sz_i} - \frac{z_r}{1-sz_r}\right)$$

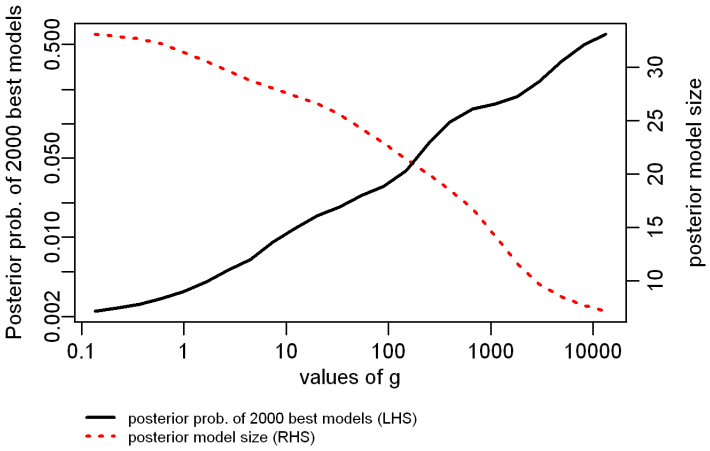
By inequality (A.6) the right hand-side is dominated by the right-hand side of the following inequality:

$$(N-1)\left(E_r\left(\frac{1-z}{1-sz}\right) - E\left(\frac{1-z}{1-sz}\right)\right) < (N-1)\frac{1-s}{s(1-sz_r)}\sum_{i=r+1}^{2^K+1} p_i\left((1-s)^{\frac{k_r-k_i}{N-1}}(\frac{m_r}{m_i})^{\frac{2}{N-1}} - 1\right)$$

A similar argument as before is applied: if $s \to \infty$ and/or $N \to \infty$, then the right-hand side will vanish. In case this right-hand side is greater than η_{r-1}, η_r might be redefined appropriately.

B Charts and Tables

Uniform model priors:



Model priors adjusted to neutralize size penalty:

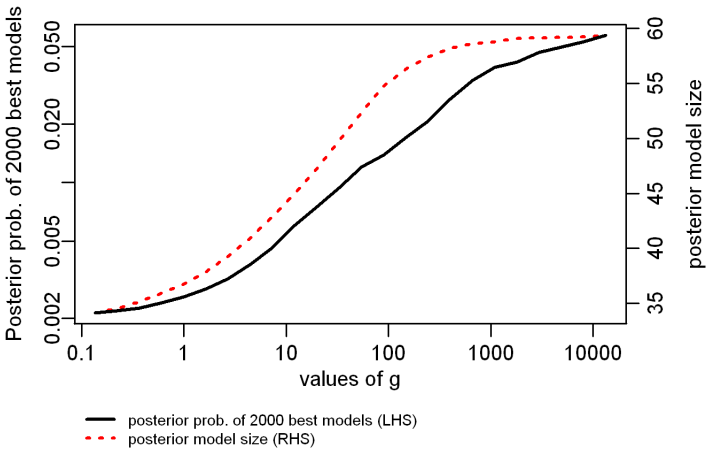


Figure 1: Illustration of the supermodel effect: sum of posterior model probabilities (PMP) for the best 2000 models (by PMP) and posterior model size for the PWT 6.3 growth data set (see section 6). Top panel shows results under uniform model priors. Bottom panel shows results under model priors that neutralize the size penalty from the g-prior (cf. section 3).

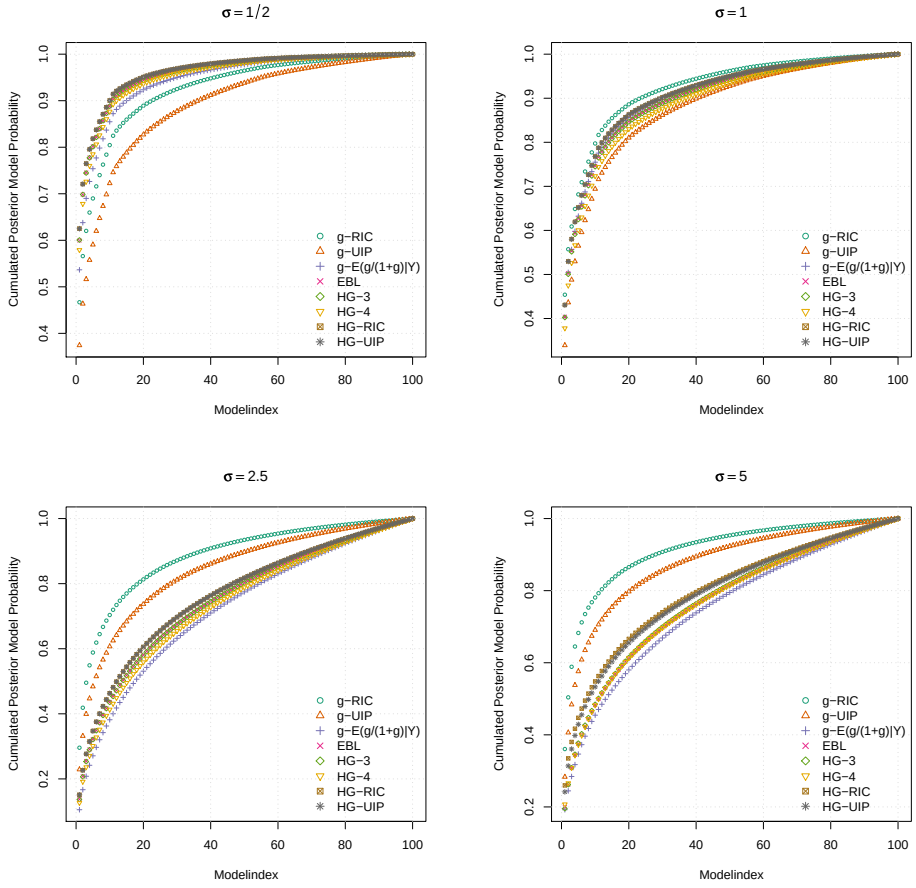


Figure 2: Cumulative posterior model probabilities for the best 100 models under Setting A. Top panel corresponds to a noise level of $\sigma = 1/2$ (left) and $\sigma = 1$ (right). Bottom panel corresponds to a ratio of $\sigma = 2.5$ (left) and $\sigma = 5$ (right).

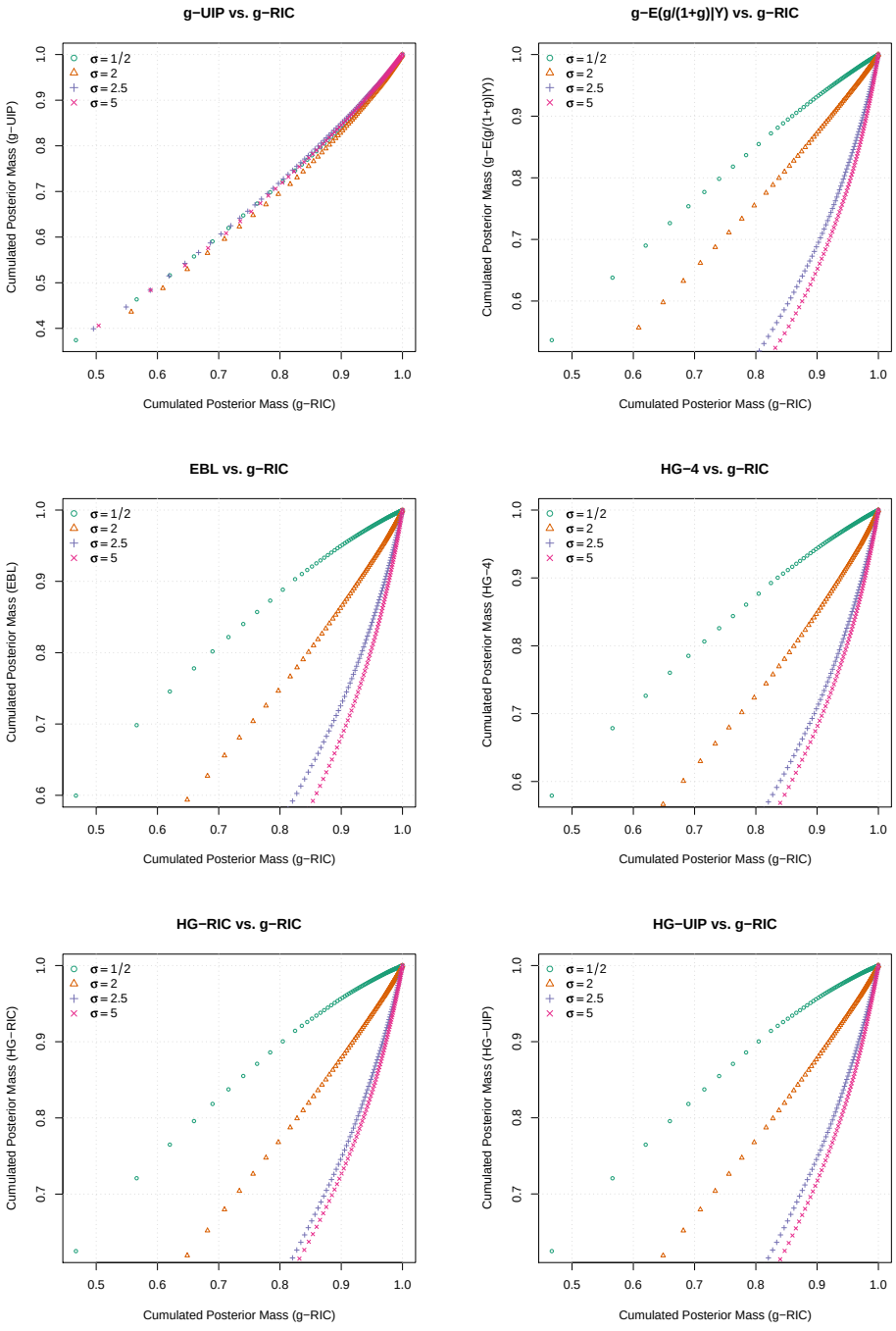


Figure 3: QQ-plot of cumulative posterior mass for different g-priors versus that of the g-RIC setting (Setting A, based on 50 Monte Carlo draws).

	signal-to-noise ratio of $\sigma = 1/2$								signal-to-noise ratio of $\sigma = 1$							
	g-RIC	g-UIP	g-E($\frac{g}{1+g}\|y$)	EBL	HG-3	HG-4	HG-RIC	HG-UIP	g-RIC	g-UIP	g-E($\frac{g}{1+g}\|y$)	EBL	HG-3	HG-4	HG-RIC	HG-UIP
$\boldsymbol{\beta_1}$	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
β_2	0.085	0.114	0.068	0.055	0.055	0.059	0.050	0.050	0.094	0.132	0.109	0.115	0.116	0.124	0.107	0.107
	(0.072)	(0.069)	(0.067)	(0.058)	(0.055)	(0.058)	(0.051)	(0.051)	(0.125)	(0.128)	(0.127)	(0.128)	(0.127)	(0.128)	(0.127)	(0.127)
β_3	0.083	0.112	0.066	0.054	0.052	0.056	0.047	0.047	0.085	0.125	0.101	0.108	0.109	0.117	0.099	0.099
	(0.067)	(0.061)	(0.064)	(0.055)	(0.046)	(0.049)	(0.043)	(0.043)	(0.073)	(0.084)	(0.078)	(0.081)	(0.080)	(0.083)	(0.077)	(0.077)
β_4	0.077	0.106	0.061	0.050	0.050	0.054	0.046	0.046	0.078	0.117	0.093	0.100	0.101	0.110	0.092	0.092
	(0.065)	(0.064)	(0.062)	(0.060)	(0.059)	(0.060)	(0.057)	(0.057)	(0.057)	(0.068)	(0.062)	(0.063)	(0.063)	(0.065)	(0.060)	(0.060)
$\boldsymbol{\beta_5}$	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
β_6	0.086	0.116	0.068	0.055	0.055	0.060	0.050	0.050	0.067	0.105	0.082	0.089	0.090	0.098	0.081	0.081
	(0.059)	(0.061)	(0.052)	(0.046)	(0.045)	(0.048)	(0.042)	(0.042)	(0.021)	(0.031)	(0.025)	(0.029)	(0.029)	(0.032)	(0.027)	(0.027)
$\boldsymbol{\beta_7}$	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
β_8	0.113	0.136	0.098	0.085	0.082	0.086	0.077	0.077	0.087	0.129	0.104	0.111	0.111	0.121	0.101	0.102
	(0.153)	(0.132)	(0.156)	(0.149)	(0.138)	(0.141)	(0.135)	(0.135)	(0.061)	(0.075)	(0.068)	(0.069)	(0.068)	(0.071)	(0.065)	(0.065)
β_9	0.084	0.114	0.066	0.054	0.054	0.058	0.049	0.049	0.083	0.125	0.099	0.107	0.108	0.117	0.098	0.098
	(0.066)	(0.070)	(0.058)	(0.055)	(0.054)	(0.057)	(0.051)	(0.051)	(0.049)	(0.066)	(0.057)	(0.065)	(0.064)	(0.068)	(0.060)	(0.060)
β_{10}	0.086	0.115	0.068	0.056	0.056	0.060	0.051	0.051	0.100	0.145	0.118	0.126	0.127	0.136	0.116	0.116
	(0.082)	(0.078)	(0.077)	(0.069)	(0.066)	(0.069)	(0.062)	(0.062)	(0.086)	(0.106)	(0.095)	(0.105)	(0.103)	(0.107)	(0.099)	(0.099)
$\boldsymbol{\beta_{11}}$	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
β_{12}	0.085	0.115	0.066	0.054	0.054	0.059	0.049	0.050	0.109	0.151	0.126	0.133	0.133	0.142	0.123	0.123
	(0.054)	(0.056)	(0.048)	(0.043)	(0.042)	(0.044)	(0.040)	(0.040)	(0.122)	(0.134)	(0.128)	(0.131)	(0.129)	(0.131)	(0.126)	(0.126)
$\boldsymbol{\beta_{13}}$	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.992**	**0.992**	**0.992**	**0.992**	**0.992**	**0.992**	**0.992**	**0.992**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.049)	(0.045)	(0.047)	(0.047)	(0.048)	(0.047)	(0.049)	(0.049)
β_{14}	0.079	0.109	0.063	0.052	0.052	0.055	0.047	0.047	0.097	0.136	0.112	0.120	0.120	0.129	0.111	0.111
	(0.069)	(0.065)	(0.067)	(0.064)	(0.063)	(0.064)	(0.061)	(0.061)	(0.124)	(0.130)	(0.127)	(0.129)	(0.127)	(0.129)	(0.125)	(0.125)
β_{15}	0.093	0.122	0.075	0.062	0.062	0.066	0.057	0.057	0.101	0.144	0.119	0.125	0.126	0.135	0.115	0.115
	(0.096)	(0.094)	(0.090)	(0.085)	(0.084)	(0.087)	(0.080)	(0.080)	(0.096)	(0.109)	(0.103)	(0.101)	(0.100)	(0.102)	(0.096)	(0.096)
$E(\frac{g}{1+g}\|Y)$	0.996	0.990	0.998	0.999	0.998	0.998	0.999	0.999	0.996	0.990	0.994	0.993	0.992	0.991	0.993	0.993

Table 2: Posterior Inclusion Probabilities for Setting A with standard deviations in parentheses. Left Panel corresponds to a signal-to-noise ratio of $\sigma = 1/2$, right panel to a ratio of $\sigma = 1$. Coefficients corresponding to variables of the data-generating model are highlighted in bold in the left column. PIP values exceeding 0.5 are highlighted in **bold**. Results are averaged over 50 Monte Carlo draws.

	signal-to-noise ratio of $\sigma = 2.5$								signal-to-noise ratio of $\sigma = 5$							
	g-RIC	g-UIP	g-E($\frac{g}{1+g}\|y$)	EBL	HG-3	HG-4	HG-RIC	HG-UIP	g-RIC	g-UIP	g-E($\frac{g}{1+g}\|y$)	EBL	HG-3	HG-4	HG-RIC	HG-UIP
$\boldsymbol{\beta_1}$	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.792**	**0.842**	**0.955**	**0.953**	**0.946**	**0.946**	**0.884**	**0.905**
	(0.002)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.257)	(0.208)	(0.066)	(0.068)	(0.077)	(0.075)	(0.207)	(0.171)
β_2	0.086	0.131	0.317	0.304	0.304	0.330	0.276	0.276	0.038	0.067	**0.512**	0.455	0.445	0.481	0.370	0.380
	(0.106)	(0.126)	(0.149)	(0.153)	(0.151)	(0.151)	(0.150)	(0.150)	(0.037)	(0.058)	(0.149)	(0.171)	(0.166)	(0.159)	(0.185)	(0.180)
β_3	0.094	0.141	0.331	0.316	0.317	0.343	0.289	0.289	0.050	0.081	**0.514**	0.460	0.450	0.486	0.376	0.386
	(0.064)	(0.077)	(0.109)	(0.115)	(0.113)	(0.114)	(0.111)	(0.111)	(0.052)	(0.074)	(0.160)	(0.180)	(0.176)	(0.167)	(0.197)	(0.192)
β_4	0.108	0.158	0.339	0.327	0.327	0.351	0.300	0.300	0.044	0.074	**0.503**	0.449	0.439	0.476	0.366	0.376
	(0.119)	(0.144)	(0.165)	(0.172)	(0.169)	(0.168)	(0.169)	(0.169)	(0.072)	(0.104)	(0.158)	(0.176)	(0.172)	(0.164)	(0.195)	(0.190)
$\boldsymbol{\beta_5}$	**0.796**	**0.826**	**0.871**	**0.873**	**0.870**	**0.875**	**0.865**	**0.865**	0.184	0.236	**0.617**	**0.579**	**0.567**	**0.594**	0.497	**0.509**
	(0.244)	(0.215)	(0.155)	(0.155)	(0.156)	(0.150)	(0.164)	(0.163)	(0.267)	(0.286)	(0.215)	(0.246)	(0.244)	(0.228)	(0.279)	(0.273)
β_6	0.062	0.103	0.282	0.272	0.273	0.298	0.245	0.245	0.048	0.080	**0.515**	0.463	0.453	0.488	0.380	0.390
	(0.044)	(0.068)	(0.115)	(0.118)	(0.117)	(0.119)	(0.113)	(0.113)	(0.070)	(0.100)	(0.185)	(0.208)	(0.204)	(0.194)	(0.226)	(0.221)
$\boldsymbol{\beta_7}$	**0.997**	**0.999**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.506**	**0.585**	**0.845**	**0.829**	**0.817**	**0.829**	**0.750**	**0.766**
	(0.014)	(0.005)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.342)	(0.340)	(0.183)	(0.208)	(0.212)	(0.195)	(0.284)	(0.268)
β_8	0.075	0.121	0.305	0.292	0.292	0.317	0.265	0.265	0.049	0.080	**0.509**	0.460	0.450	0.486	0.376	0.386
	(0.065)	(0.095)	(0.136)	(0.151)	(0.149)	(0.149)	(0.148)	(0.148)	(0.075)	(0.098)	(0.157)	(0.172)	(0.169)	(0.160)	(0.194)	(0.188)
β_9	0.131	0.187	0.369	0.356	0.356	0.380	0.329	0.330	0.045	0.077	**0.508**	0.455	0.445	0.482	0.372	0.382
	(0.161)	(0.199)	(0.217)	(0.219)	(0.216)	(0.213)	(0.218)	(0.218)	(0.061)	(0.094)	(0.165)	(0.176)	(0.173)	(0.165)	(0.198)	(0.192)
β_{10}	0.082	0.127	0.313	0.302	0.303	0.328	0.275	0.275	0.042	0.072	**0.520**	0.465	0.454	0.491	0.377	0.388
	(0.089)	(0.113)	(0.144)	(0.151)	(0.148)	(0.149)	(0.146)	(0.146)	(0.040)	(0.059)	(0.147)	(0.163)	(0.159)	(0.152)	(0.181)	(0.175)
$\boldsymbol{\beta_{11}}$	**0.828**	**0.869**	**0.925**	**0.922**	**0.919**	**0.924**	**0.913**	**0.913**	0.305	0.369	**0.717**	**0.689**	**0.677**	**0.699**	**0.609**	**0.622**
	(0.240)	(0.205)	(0.141)	(0.146)	(0.147)	(0.141)	(0.156)	(0.156)	(0.316)	(0.323)	(0.222)	(0.246)	(0.246)	(0.229)	(0.294)	(0.285)
β_{12}	0.065	0.109	0.301	0.288	0.289	0.315	0.259	0.260	0.054	0.093	**0.533**	0.480	0.469	**0.504**	0.393	0.404
	(0.036)	(0.053)	(0.105)	(0.109)	(0.107)	(0.111)	(0.102)	(0.102)	(0.067)	(0.109)	(0.171)	(0.196)	(0.192)	(0.182)	(0.217)	(0.211)
$\boldsymbol{\beta_{13}}$	0.437	**0.504**	**0.651**	**0.640**	**0.637**	**0.653**	**0.617**	**0.617**	0.083	0.129	**0.558**	**0.511**	0.500	**0.532**	0.426	0.437
	(0.332)	(0.322)	(0.261)	(0.268)	(0.267)	(0.258)	(0.277)	(0.277)	(0.119)	(0.151)	(0.189)	(0.216)	(0.212)	(0.200)	(0.239)	(0.233)
β_{14}	0.098	0.143	0.321	0.309	0.310	0.334	0.283	0.283	0.043	0.072	**0.504**	0.449	0.439	0.476	0.366	0.375
	(0.129)	(0.146)	(0.157)	(0.163)	(0.161)	(0.159)	(0.162)	(0.162)	(0.058)	(0.083)	(0.160)	(0.176)	(0.172)	(0.164)	(0.191)	(0.187)
β_{15}	0.165	0.212	0.386	0.372	0.373	0.396	0.347	0.347	0.074	0.109	**0.520**	0.470	0.461	0.496	0.387	0.398
	(0.147)	(0.156)	(0.148)	(0.160)	(0.158)	(0.156)	(0.161)	(0.161)	(0.114)	(0.131)	(0.155)	(0.175)	(0.171)	(0.162)	(0.195)	(0.191)
$E(k\|Y)$	5.027	5.629	7.711	7.572	7.568	7.844	7.260	7.264	2.359	2.965	8.830	8.167	8.011	8.467	6.928	7.103
$E(\frac{g}{1+g}\|Y)$	0.996	0.990	0.949	0.955	0.947	0.939	0.956	0.956	0.996	0.990	0.783	0.817	0.795	0.760	0.856	0.849

Table 3: Posterior Inclusion Probabilities for Setting 'A' with standard deviations in parentheses. Left Panel corresponds to a signal-to-noise ratio of $\sigma = 2.5$, right panel to a ratio of $\sigma = 5$ noise. Coefficients corresponding to variables of the data-generating model are highlighted in bold in the left column. PIP values exceeding 0.5 are highlighted in **bold**. Results are averaged over 50 Monte Carlo draws.

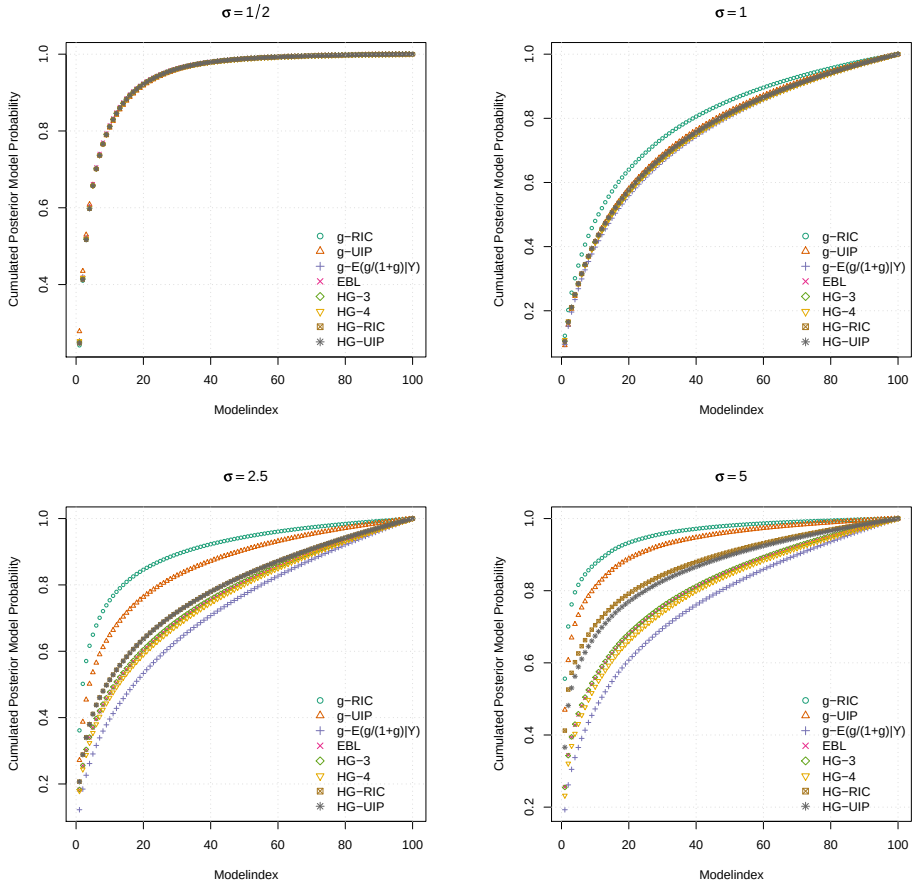


Figure 4: Cumulative posterior model probabilities for the best 100 models under Setting B. Top panel corresponds to a noise level of $\sigma = 1/2$ (left) and $\sigma = 1$ (right). Bottom panel corresponds to a ratio of $\sigma = 2.5$ (left) and $\sigma = 5$ (right).

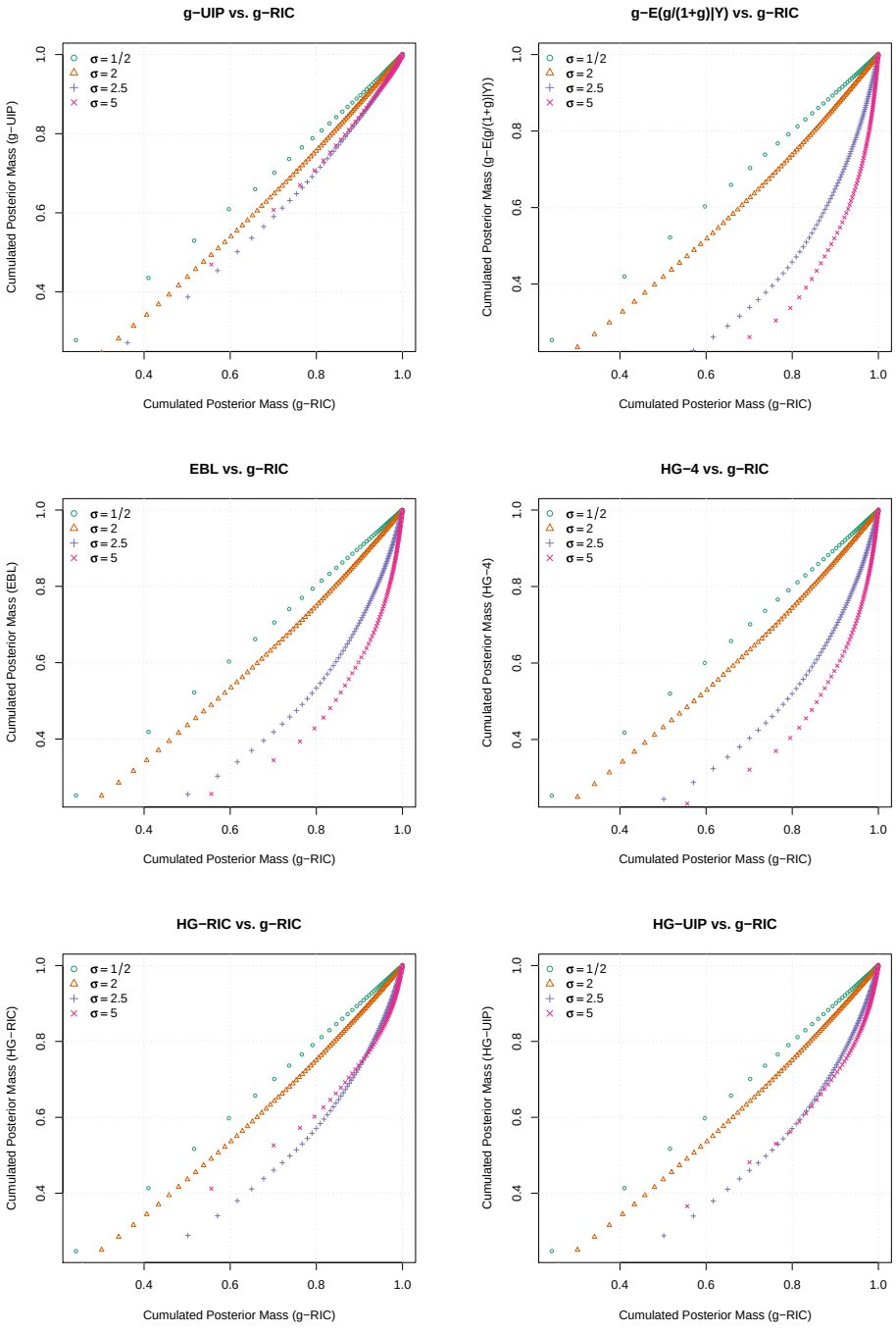


Figure 5: QQ-plot of cumulative posterior mass for different g-priors versus that of the g-RIC setting (Setting B, based on 50 Monte Carlo draws).

	signal-to-noise ratio of $\sigma = 1/2$								signal-to-noise ratio of $\sigma = 1$							
	g-RIC	g-UIP	g-E($\frac{g}{1+g}\|y$)	EBL	HG-3	HG-4	HG-RIC	HG-UIP	g-RIC	g-UIP	g-E($\frac{g}{1+g}\|y$)	EBL	HG-3	HG-4	HG-RIC	HG-UIP
β_1	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
β_2	**0.801**	**0.835**	**0.818**	**0.810**	**0.808**	**0.813**	**0.804**	**0.804**	0.374	0.487	**0.567**	**0.562**	**0.560**	**0.575**	**0.544**	**0.544**
	(0.223)	(0.182)	(0.205)	(0.218)	(0.217)	(0.213)	(0.223)	(0.222)	(0.266)	(0.262)	(0.242)	(0.247)	(0.245)	(0.240)	(0.250)	(0.250)
β_3	**0.762**	**0.802**	**0.781**	**0.773**	**0.772**	**0.777**	**0.767**	**0.767**	0.393	0.480	**0.550**	**0.548**	**0.548**	**0.560**	**0.533**	**0.534**
	(0.239)	(0.195)	(0.220)	(0.228)	(0.227)	(0.222)	(0.232)	(0.232)	(0.295)	(0.270)	(0.243)	(0.248)	(0.247)	(0.241)	(0.253)	(0.253)
β_4	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**	**0.747**	**0.806**	**0.840**	**0.840**	**0.838**	**0.845**	**0.831**	**0.831**
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.282)	(0.242)	(0.207)	(0.206)	(0.206)	(0.199)	(0.215)	(0.215)
β_5	**0.885**	**0.897**	**0.891**	**0.889**	**0.888**	**0.889**	**0.886**	**0.886**	0.437	**0.516**	**0.577**	**0.573**	**0.572**	**0.584**	**0.559**	**0.560**
	(0.147)	(0.119)	(0.135)	(0.140)	(0.140)	(0.136)	(0.143)	(0.143)	(0.293)	(0.271)	(0.245)	(0.249)	(0.247)	(0.241)	(0.254)	(0.254)
β_6	**0.943**	**0.952**	**0.948**	**0.947**	**0.946**	**0.947**	**0.944**	**0.944**	0.470	**0.560**	**0.625**	**0.621**	**0.620**	**0.632**	**0.606**	**0.606**
	(0.116)	(0.093)	(0.106)	(0.107)	(0.108)	(0.105)	(0.111)	(0.111)	(0.332)	(0.303)	(0.272)	(0.277)	(0.276)	(0.269)	(0.283)	(0.283)
β_7	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.980**	**0.988**	**0.991**	**0.990**	**0.990**	**0.990**	**0.989**	**0.989**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.039)	(0.023)	(0.018)	(0.020)	(0.021)	(0.019)	(0.022)	(0.022)
β_8	**0.696**	**0.749**	**0.720**	**0.712**	**0.711**	**0.718**	**0.704**	**0.704**	0.325	0.430	**0.509**	**0.506**	**0.506**	**0.520**	0.489	0.490
	(0.234)	(0.190)	(0.216)	(0.221)	(0.220)	(0.215)	(0.225)	(0.225)	(0.237)	(0.242)	(0.227)	(0.233)	(0.231)	(0.227)	(0.236)	(0.236)
β_9	**0.679**	**0.734**	**0.704**	**0.694**	**0.693**	**0.700**	**0.686**	**0.686**	0.255	0.364	0.452	0.449	0.449	0.465	0.431	0.431
	(0.256)	(0.211)	(0.238)	(0.247)	(0.246)	(0.241)	(0.252)	(0.252)	(0.189)	(0.196)	(0.188)	(0.197)	(0.195)	(0.192)	(0.198)	(0.198)
β_{10}	**0.998**	**0.998**	**0.999**	**0.999**	**0.998**	**0.998**	**0.998**	**0.998**	**0.737**	**0.804**	**0.841**	**0.838**	**0.836**	**0.843**	**0.828**	**0.828**
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.271)	(0.225)	(0.190)	(0.192)	(0.193)	(0.186)	(0.200)	(0.200)
β_{11}	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.997**	**0.998**	**0.999**	**0.999**	**0.998**	**0.999**	**0.998**	**0.998**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.022)	(0.013)	(0.010)	(0.010)	(0.010)	(0.010)	(0.011)	(0.011)
β_{12}	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**	**0.671**	**0.754**	**0.804**	**0.795**	**0.793**	**0.802**	**0.782**	**0.782**
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.300)	(0.248)	(0.207)	(0.217)	(0.218)	(0.210)	(0.227)	(0.227)
β_{13}	**0.989**	**0.990**	**0.989**	**0.989**	**0.989**	**0.989**	**0.989**	**0.989**	**0.728**	**0.773**	**0.804**	**0.801**	**0.800**	**0.806**	**0.793**	**0.794**
	(0.053)	(0.043)	(0.048)	(0.050)	(0.050)	(0.049)	(0.052)	(0.052)	(0.303)	(0.265)	(0.233)	(0.236)	(0.235)	(0.229)	(0.243)	(0.243)
β_{14}	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.993**	**0.996**	**0.997**	**0.996**	**0.996**	**0.996**	**0.996**	**0.996**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.030)	(0.021)	(0.016)	(0.018)	(0.018)	(0.017)	(0.020)	(0.020)
β_{15}	0.486	**0.575**	**0.524**	**0.508**	**0.509**	**0.519**	0.499	0.499	0.285	0.380	0.460	0.457	0.457	0.472	0.441	0.441
	(0.191)	(0.155)	(0.177)	(0.185)	(0.183)	(0.179)	(0.187)	(0.187)	(0.201)	(0.194)	(0.180)	(0.185)	(0.184)	(0.180)	(0.188)	(0.188)
$E(k\|Y)$	13.238	13.530	13.373	13.318	13.313	13.348	13.275	13.275	9.391	10.335	11.015	10.976	10.964	11.089	10.822	10.824
$E(\frac{g}{1+g}\|Y)$	0.996	0.990	0.994	0.995	0.994	0.993	0.994	0.994	0.996	0.990	0.982	0.983	0.980	0.978	0.982	0.982

Table 4: Posterior Inclusion Probabilities for Setting B with standard deviations in parentheses. Left Panel corresponds to a signal-to-noise ratio of $\sigma = 1/2$, right panel to a ratio of $\sigma = 1$. PIP values exceeding 0.5 are highlighted in **bold**. Results are averaged over 50 Monte Carlo draws.

	signal-to-noise ratio of $\sigma = 2.5$								signal-to-noise ratio of $\sigma = 5$							
	g-RIC	g-UIP	g-E($\frac{g}{1+g}\|y$)	EBL	HG-3	HG-4	HG-RIC	HG-UIP	g-RIC	g-UIP	g-E($\frac{g}{1+g}\|y$)	EBL	HG-3	HG-4	HG-RIC	HG-UIP
β_1	**0.999**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.819**	**0.851**	**0.944**	**0.940**	**0.930**	**0.927**	**0.804**	**0.857**
	(0.004)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.274)	(0.244)	(0.113)	(0.109)	(0.115)	(0.107)	(0.239)	(0.192)
β_2	0.060	0.105	0.355	0.377	0.373	0.409	0.327	0.328	0.030	0.048	0.412	0.392	0.376	0.423	0.261	0.282
	(0.059)	(0.096)	(0.179)	(0.196)	(0.193)	(0.188)	(0.197)	(0.197)	(0.050)	(0.068)	(0.163)	(0.183)	(0.174)	(0.168)	(0.171)	(0.172)
β_3	0.112	0.165	0.414	0.432	0.426	0.460	0.382	0.382	0.040	0.063	0.422	0.399	0.384	0.428	0.264	0.288
	(0.150)	(0.170)	(0.200)	(0.211)	(0.209)	(0.203)	(0.213)	(0.213)	(0.069)	(0.094)	(0.184)	(0.206)	(0.197)	(0.187)	(0.177)	(0.184)
β_4	0.055	0.105	0.376	0.394	0.389	0.426	0.342	0.342	0.033	0.053	0.436	0.412	0.395	0.442	0.276	0.298
	(0.060)	(0.114)	(0.208)	(0.225)	(0.221)	(0.216)	(0.224)	(0.224)	(0.049)	(0.066)	(0.168)	(0.186)	(0.177)	(0.170)	(0.178)	(0.178)
β_5	0.056	0.093	0.332	0.355	0.351	0.387	0.306	0.307	0.028	0.047	0.420	0.397	0.382	0.428	0.263	0.286
	(0.066)	(0.083)	(0.159)	(0.185)	(0.183)	(0.179)	(0.186)	(0.186)	(0.025)	(0.039)	(0.168)	(0.187)	(0.178)	(0.173)	(0.169)	(0.171)
β_6	0.083	0.128	0.374	0.396	0.391	0.428	0.344	0.345	0.024	0.040	0.395	0.372	0.357	0.404	0.239	0.262
	(0.154)	(0.174)	(0.210)	(0.216)	(0.213)	(0.207)	(0.218)	(0.218)	(0.034)	(0.051)	(0.149)	(0.180)	(0.170)	(0.163)	(0.148)	(0.157)
β_7	0.230	0.304	**0.547**	**0.555**	**0.549**	**0.579**	**0.508**	**0.508**	0.062	0.086	0.446	0.423	0.407	0.451	0.288	0.312
	(0.262)	(0.298)	(0.292)	(0.299)	(0.296)	(0.283)	(0.311)	(0.311)	(0.144)	(0.165)	(0.193)	(0.216)	(0.209)	(0.199)	(0.209)	(0.210)
β_8	0.057	0.095	0.335	0.359	0.355	0.392	0.308	0.309	0.024	0.040	0.396	0.375	0.360	0.407	0.245	0.267
	(0.072)	(0.103)	(0.172)	(0.188)	(0.185)	(0.183)	(0.185)	(0.185)	(0.027)	(0.042)	(0.157)	(0.183)	(0.174)	(0.168)	(0.166)	(0.168)
β_9	0.056	0.097	0.332	0.356	0.353	0.388	0.309	0.309	0.025	0.042	0.397	0.379	0.364	0.411	0.250	0.271
	(0.063)	(0.098)	(0.179)	(0.193)	(0.190)	(0.186)	(0.193)	(0.193)	(0.029)	(0.046)	(0.158)	(0.175)	(0.166)	(0.162)	(0.160)	(0.161)
β_{10}	0.107	0.154	0.402	0.424	0.419	0.453	0.376	0.376	0.046	0.071	0.435	0.417	0.401	0.446	0.286	0.308
	(0.193)	(0.210)	(0.235)	(0.252)	(0.249)	(0.241)	(0.255)	(0.255)	(0.072)	(0.102)	(0.197)	(0.209)	(0.202)	(0.193)	(0.206)	(0.205)
β_{11}	**0.503**	**0.580**	**0.764**	**0.768**	**0.761**	**0.781**	**0.727**	**0.728**	0.138	0.170	**0.524**	**0.501**	0.484	**0.525**	0.360	0.387
	(0.352)	(0.344)	(0.269)	(0.260)	(0.261)	(0.245)	(0.283)	(0.282)	(0.243)	(0.257)	(0.236)	(0.245)	(0.241)	(0.226)	(0.249)	(0.250)
β_{12}	0.115	0.173	0.433	0.451	0.445	0.480	0.399	0.399	0.040	0.065	0.433	0.410	0.394	0.440	0.274	0.297
	(0.164)	(0.197)	(0.241)	(0.243)	(0.239)	(0.234)	(0.243)	(0.243)	(0.067)	(0.098)	(0.173)	(0.196)	(0.188)	(0.179)	(0.187)	(0.187)
β_{13}	0.126	0.178	0.411	0.430	0.426	0.458	0.384	0.384	0.073	0.094	0.436	0.415	0.400	0.444	0.284	0.307
	(0.175)	(0.201)	(0.227)	(0.243)	(0.240)	(0.232)	(0.246)	(0.246)	(0.178)	(0.188)	(0.187)	(0.211)	(0.205)	(0.193)	(0.212)	(0.212)
β_{14}	0.234	0.318	**0.600**	**0.600**	**0.592**	**0.625**	**0.546**	**0.547**	0.051	0.080	0.475	0.448	0.431	0.475	0.308	0.332
	(0.273)	(0.291)	(0.276)	(0.286)	(0.285)	(0.269)	(0.304)	(0.304)	(0.072)	(0.097)	(0.198)	(0.216)	(0.208)	(0.198)	(0.208)	(0.209)
β_{15}	0.045	0.083	0.327	0.350	0.346	0.383	0.300	0.300	0.033	0.051	0.410	0.391	0.376	0.422	0.260	0.282
	(0.039)	(0.066)	(0.151)	(0.169)	(0.167)	(0.163)	(0.170)	(0.170)	(0.073)	(0.092)	(0.159)	(0.181)	(0.173)	(0.168)	(0.173)	(0.173)
$E(k\|Y)$	2.837	3.577	7.001	7.244	7.173	7.649	6.556	6.564	1.465	1.803	6.981	6.670	6.444	7.073	4.662	5.036
$E(\frac{g}{1+g}\|Y)$	0.996	0.990	0.933	0.926	0.913	0.897	0.931	0.931	0.996	0.990	0.782	0.796	0.767	0.716	0.866	0.850

Table 5: Posterior Inclusion Probabilities for Setting B with standard deviations in parenthesis. Left Panel corresponds to a signal-to-noise ratio of $\sigma = 2.5$, right panel to a ratio of $\sigma = 5$ noise. PIP values exceeding 0.5 are highlighted in **bold**. Results are averaged over 50 Monte Carlo draws.

		g-RIC	g-UIP	g-E($\frac{g}{1+g}\|y$)	EBL	HG-3	HG-4	HG-RIC	HG-UIP
$\sigma = \frac{1}{2}$	Min.	0.1349	0.1279	0.1573	0.1888	0.1934	0.1809	0.2096	0.2094
	Mean	0.4618	0.3725	0.5306	0.5951	0.5973	0.5758	0.6220	0.6217
	Max.	0.6297	0.5106	0.7037	0.7669	0.7644	0.7461	0.7845	0.7843
	St.Dev.	0.1342	0.1019	0.1482	0.1551	0.1490	0.1487	0.1484	0.1484
$\sigma = 1$	Min.	0.0539	0.0317	0.0426	0.0308	0.0320	0.0289	0.0362	0.0362
	Mean	0.4433	0.3290	0.3922	0.3944	0.3932	0.3690	0.4219	0.4215
	Max.	0.6115	0.4849	0.5578	0.5954	0.5922	0.5658	0.6220	0.6216
	St.Dev.	0.1373	0.1138	0.1283	0.1358	0.1342	0.1293	0.1392	0.1392
$\sigma = 2.5$	Min.	0.0021	0.0022	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Mean	0.1201	0.1048	0.0556	0.0665	0.0660	0.0606	0.0721	0.0720
	Max.	0.4609	0.3392	0.1493	0.1978	0.1968	0.1768	0.2216	0.2213
	St.Dev.	0.1133	0.0834	0.0387	0.0487	0.0478	0.0441	0.0524	0.0524
$\sigma = 5$	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Mean	0.0012	0.0023	0.0010	0.0025	0.0024	0.0021	0.0027	0.0028
	Max.	0.0215	0.0324	0.0114	0.0323	0.0308	0.0254	0.0347	0.0347
	St.Dev.	0.0035	0.0057	0.0026	0.0062	0.0060	0.0051	0.0067	0.0067

Table 6: Summary statistics of posterior model probabilities for true model based on setting 'A' and 50 Monte Carlo Steps. Top panel corresponds to $\sigma = 1/2$, second panel to $\sigma = 1$, third panel to $\sigma = 2.5$, fourth panel to $\sigma = 5$.

		g-RIC	g-UIP	g-E($\frac{g}{1+g}\|y$)	EBL	HG-3	HG-4	HG-UIP	HG-RIC
$\sigma = \frac{1}{2}$	Min.	0.4752	0.6133	0.4704	0.5192	0.5386	0.5175	0.5695	0.5691
	Mean	0.9806	0.9919	0.9807	0.9872	0.9908	0.9902	0.9914	0.9914
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	St.Dev.	0.1342	0.1019	0.1482	0.1551	0.1490	0.1487	0.1484	0.1484
$\sigma = 1$	Min.	0.1363	0.1115	0.1226	0.1131	0.1188	0.1158	0.1239	0.1238
	Mean	0.9650	0.9552	0.9604	0.9612	0.9626	0.9604	0.9657	0.9656
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	St.Dev.	0.1373	0.1138	0.1283	0.1358	0.1342	0.1293	0.1392	0.1392
$\sigma = 2.5$	Min.	0.0071	0.0076	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Mean	0.4683	0.5202	0.5516	0.5325	0.5331	0.5274	0.5382	0.5383
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	St.Dev.	0.1133	0.0834	0.0387	0.0487	0.0478	0.0441	0.0524	0.0524
$\sigma = 5$	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Mean	0.0067	0.0173	0.0070	0.0203	0.0200	0.0147	0.0284	0.0285
	Max.	0.1643	0.3744	0.0842	0.2835	0.2737	0.1685	0.4719	0.4688
	St.Dev.	0.0035	0.0057	0.0026	0.0062	0.0060	0.0051	0.0067	0.0067

Table 7: Ratio of posterior model probability for true model divided by the PMP of the best model (summary statistics for setting 'A', based on 50 Monte Carlo draws). Top panel corresponds to $\sigma = 1/2$, second panel to $\sigma = 1$, third panel to $\sigma = 2.5$, fourth panel to $\sigma = 5$.

	g-RIC	g-UIP	g-E($\frac{g}{1+g}\|y$)	EBL	HG-3	HG-4	HG-RIC	HG-UIP
$\sigma = 1/2$	-	1.00877	0.99754	0.99798	0.99793	0.99817	0.99771	0.99771
$\sigma = 1$	-	1.00347	1.00200	1.00128	1.00219	1.00315	1.00126	1.00127
$\sigma = 2.5$	-	0.99501	1.00079	1.00556	1.00320	1.00699	1.00039	1.00042
$\sigma = 5$	-	0.99034	1.00697	1.00594	1.00720	1.01958	1.01256	1.00692
$\sigma = 1/2$	-	0.99754	0.99926	0.99794	0.99948	0.99910	0.99998	0.99998
$\sigma = 1$	-	0.98166	0.97396	0.97316	0.97501	0.97398	0.97648	0.97647
$\sigma = 2.5$	-	0.98875	0.97284	0.96760	0.97580	0.97847	0.97631	0.97627
$\sigma = 5$	-	0.99578	1.00427	0.99968	1.00747	1.02216	1.01853	1.00976

Table 8: Root mean squared errors relative to g-RIC, averaged over 50 Monte Carlo draws of data (y, X) and based on 30 out of forecasts over random sample splits of data under each draw. Values below 1 indicate predictive performance that is superior to the g-RIC setting. Top panel corresponds to setting 'A' and bottom panel to setting 'B'.

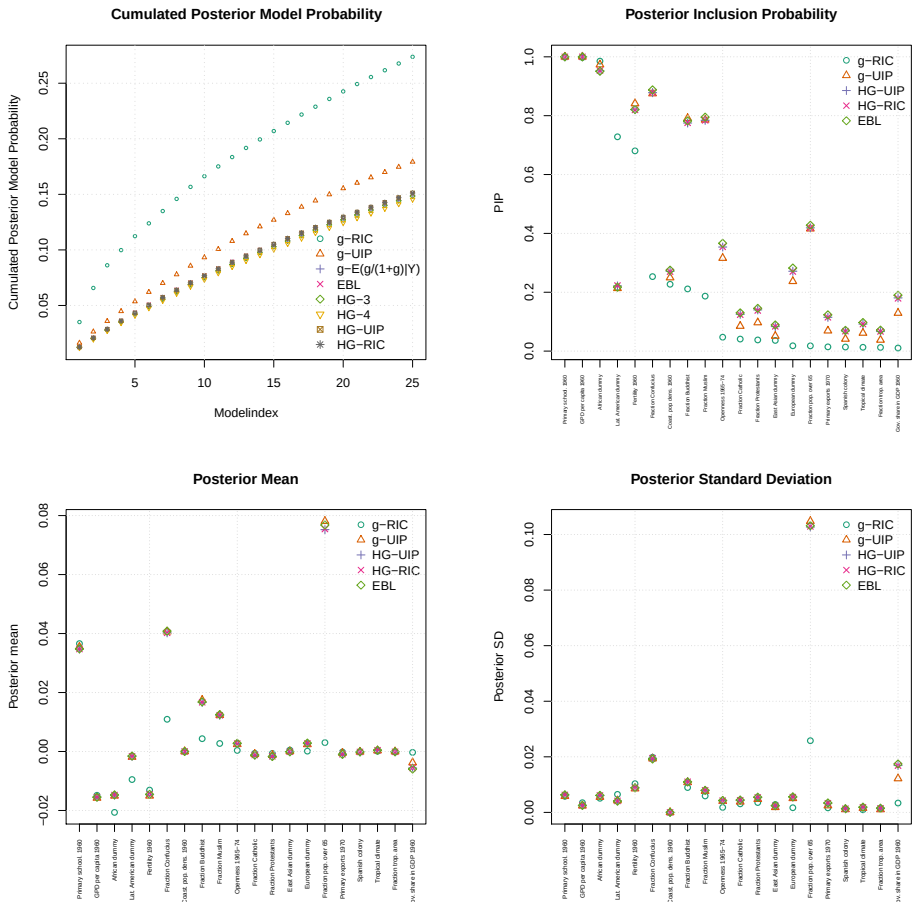


Figure 6: Estimation results for the PWT 6.1 data set from section 6: Top left panel shows the cumulative posterior mass, left panel the posterior inclusion probabilities for the most important 20 variables, bottom left panel the (expected) standardized coefficients and bottom right panel these standardized coefficients' posterior standard deviation.

	PWT 6.0	PWT 6.1	PWT 6.2	PWT 6.3
g-RIC	0.01384	0.01357	0.01123	0.01476
g-UIP	0.01359	0.01227	0.01070	0.01502
HG-UIP	0.01336	0.01199	0.01053	0.01473
HG-RIC	0.01336	0.01198	0.01054	0.01471
EBL	0.01344	0.01202	0.01061	0.01475

Table 9: Root mean squared errors (RMSE) based on 30 random data partitions, as described in section 6. Each random sample split assigns 75% as training sample retaining 25% of the data for the forecast evaluation.

	g=K^2	g=N	hyperUIP	hyperBRIC	EBL
PWT 6.0 vs. 6.1	5.4655	1.7971	1.5700	1.5714	1.5580
PWT 6.0 vs. 6.2	4.0735	1.9079	1.6364	1.6336	1.6116
PWT 6.0 vs. 6.3	5.5748	2.6756	1.9654	1.9715	1.9536
PWT 6.1 vs. 6.2	1.6936	1.5184	1.3208	1.3198	1.3037
PWT 6.1 vs. 6.3	2.0679	1.9518	1.6568	1.6573	1.6484
PWT 6.2 vs. 6.3	1.7748	2.2089	1.7516	1.7468	1.7334
PWT Overall Max / Min Ratio	7.3494	3.3100	2.3961	2.3938	2.3565

Table 10: Average PIP Max/Min ratios: for each revision pair, these are the mean of the ratio maximum vs. minimum PIP per variable (over all 67 variables).

	g-RIC				hyper-g (UIP)			
PWT Revision	6.0	6.1	6.2	6.3	6.0	6.1	6.2	6.3
East Asian Dummy	0.85	0.04	0.22	0.18	0.12	0.08	0.11	0.18
Primary Schooling in 1960	0.76	1.00	0.96	1.00	1.00	1.00	1.00	1.00
Fraction of Tropical Area	0.60	0.01	0.04	0.01	0.13	0.07	0.07	0.08
African Dummy	0.23	0.99	0.94	0.96	0.76	0.95	0.91	0.82
GDP in 1960 (log)	0.16	1.00	0.96	1.00	0.95	1.00	1.00	1.00
Latin American Dummy	0.13	0.73	0.36	0.18	0.29	0.22	0.29	0.19
Malaria Prevalence in 1960s	0.13	0.00	0.01	0.01	0.08	0.09	0.10	0.25
Population Density Coastal in 1960s	0.10	0.23	0.03	0.04	0.35	0.27	0.10	0.14
Fraction Buddhist	0.06	0.21	0.17	0.11	0.63	0.78	0.60	0.36
Spanish Colony	0.05	0.01	0.03	0.03	0.29	0.07	0.09	0.11
Higher Education 1960	0.04	0.00	0.00	0.00	0.18	0.06	0.06	0.09
Fraction Confucius	0.04	0.25	0.24	0.20	0.72	0.88	0.80	0.82
Fertility in 1960s	0.02	0.68	0.91	0.91	0.38	0.82	0.88	0.77
Nominal Government GDP Share 1960s	0.02	0.01	0.01	0.00	0.72	0.28	0.38	0.25
Primary Exports 1970	0.02	0.01	0.01	0.09	0.09	0.11	0.11	0.64
Real Exchange Rate Distortions	0.02	0.00	0.00	0.00	0.53	0.07	0.05	0.08
Life Expectancy in 1960	0.02	0.00	0.01	0.00	0.06	0.05	0.05	0.07
Fraction Population In Tropics	0.01	0.00	0.01	0.01	0.07	0.06	0.06	0.08
Openness measure 1965-74	0.01	0.05	0.03	0.09	0.29	0.35	0.24	0.60
Colony Dummy	0.01	0.00	0.00	0.00	0.09	0.07	0.05	0.08
Civil Liberties	0.01	0.00	0.00	0.00	0.14	0.06	0.05	0.06
Fraction Protestants	0.01	0.04	0.04	0.01	0.28	0.14	0.17	0.13
Absolute Latitude	0.01	0.01	0.01	0.01	0.12	0.10	0.10	0.14
Fraction Catholic	0.01	0.04	0.05	0.02	0.29	0.12	0.14	0.13
Years Open 1950-94	0.01	0.01	0.01	0.02	0.13	0.07	0.08	0.17
European Dummy	0.01	0.02	0.01	0.01	0.31	0.27	0.14	0.15
Fraction Muslim	0.01	0.19	0.09	0.05	0.43	0.78	0.58	0.56
Fraction Population Over 65	0.01	0.02	0.01	0.01	0.31	0.42	0.26	0.15
Population Growth Rate 1960-90	0.00	0.01	0.01	0.01	0.09	0.08	0.08	0.14
Fraction Hindus	0.00	0.00	0.00	0.00	0.13	0.09	0.07	0.16
Government Share of GDP in 1960s	0.00	0.01	0.00	0.00	0.12	0.18	0.08	0.14
Air Distance to Big Cities	0.00	0.01	0.01	0.03	0.11	0.09	0.11	0.10
Fraction Population Less than 15	0.00	0.01	0.01	0.01	0.11	0.09	0.08	0.10
Gov. Consumption Share 1960s	0.00	0.01	0.00	0.00	0.10	0.10	0.06	0.11
Fraction GDP in Mining	0.00	0.00	0.00	0.00	0.07	0.05	0.05	0.07
Investment Price	0.00	0.00	0.00	0.00	0.06	0.08	0.05	0.07
Timing of Independence	0.00	0.01	0.01	0.01	0.08	0.08	0.13	0.40
Fraction Speaking Foreign Language	0.00	0.00	0.00	0.00	0.07	0.05	0.06	0.07
Ethnolinguistic Fractionalization	0.00	0.00	0.00	0.00	0.06	0.11	0.07	0.07
Population Density 1960	0.00	0.00	0.00	0.00	0.07	0.08	0.06	0.09
Defence Spending Share	0.00	0.00	0.01	0.00	0.09	0.10	0.11	0.28
Political Rights	0.00	0.00	0.00	0.00	0.10	0.06	0.05	0.08
Population in 1960	0.00	0.00	0.00	0.00	0.07	0.06	0.05	0.10
War Participation 1960-90	0.00	0.00	0.00	0.00	0.06	0.06	0.07	0.06
Tropical Climate Zone	0.00	0.01	0.01	0.01	0.07	0.09	0.12	0.12
Fraction Orthodox	0.00	0.00	0.00	0.00	0.07	0.09	0.08	0.07
Square of Inflation 1960-90	0.00	0.00	0.01	0.01	0.06	0.06	0.08	0.09
Average Inflation 1960-90	0.00	0.00	0.01	0.02	0.06	0.06	0.10	0.09
English Speaking Population	0.00	0.00	0.00	0.00	0.06	0.05	0.05	0.08
Land Area	0.00	0.00	0.00	0.00	0.15	0.18	0.09	0.11
Terms of Trade Ranking	0.00	0.00	0.00	0.00	0.06	0.05	0.05	0.07
Public Education Spending Share in GDP in 1960s	0.00	0.00	0.00	0.00	0.08	0.05	0.05	0.07
Religion Measure	0.00	0.00	0.00	0.00	0.08	0.09	0.06	0.09
Revolutions and Coups	0.00	0.00	0.00	0.00	0.07	0.07	0.07	0.25
Landlocked Country Dummy	0.00	0.01	0.00	0.02	0.10	0.09	0.06	0.24
Fraction of Land Area Near Navigable Water	0.00	0.00	0.00	0.00	0.10	0.09	0.06	0.08
Size of Economy	0.00	0.00	0.00	0.01	0.07	0.10	0.07	0.11
Public Investment Share	0.00	0.00	0.00	0.00	0.07	0.05	0.05	0.07
Socialist Dummy	0.00	0.01	0.00	0.01	0.08	0.19	0.07	0.34
Oil Producing Country Dummy	0.00	0.00	0.00	0.00	0.06	0.05	0.06	0.07
Outward Orientation	0.00	0.00	0.00	0.01	0.07	0.05	0.05	0.07
Hydrocarbon Deposits in 1993	0.00	0.01	0.01	0.01	0.09	0.24	0.17	0.61
British Colony Dummy	0.00	0.00	0.00	0.00	0.07	0.05	0.05	0.06
Capitalism	0.00	0.00	0.00	0.00	0.06	0.10	0.06	0.36
Terms of Trade Growth in 1960s	0.00	0.00	0.00	0.00	0.07	0.05	0.05	0.07
Interior Density	0.00	0.00	0.00	0.00	0.05	0.05	0.04	0.11
Fraction Spent in War 1960-90	0.00	0.00	0.00	0.00	0.06	0.05	0.05	0.07
# Regressors	3.42	5.70	5.30	5.17	12.81	12.60	11.26	14.40
$E(g/(1+g)\|Y)$	0.9998	0.9998	0.9998	0.9998	0.9594	0.9665	0.9625	0.9630

Table 11: Posterior inclusion probabilities over PWT revisions. Left panel corresponds to fixed g=K^2, right panel to hyper-g (UIP).