IMS Collections
Borrowing Strength: Theory Powering Applications –
A Festschrift for Lawrence D. Brown
Vol. 6 (2010) 158–165

DOI: 10.1214/10-IMSCOLL611

Dilution priors: Compensating for model space redundancy

Edward I. George

University of Pennsylvania

Abstract: For the general Bayesian model uncertainty framework, the focus of this paper is on the development of model space priors which can compensate for redundancy between model classes, the so-called dilution priors proposed in George (1999). Several distinct approaches for dilution prior construction are suggested. One is based on tessellation determined neighborhoods, another on collinearity adjustments, and a third on pairwise distances between models.

1. Introduction

Suppose a space of models Γ is considered for modeling data Y. Under model $\gamma \in \Gamma$, Y is assumed to have density $f(Y|\theta_\gamma, \gamma)$ where θ_γ is a vector of unknown parameters. (Although we refer to γ as a model, it is more precisely, a set of models indexed by θ_γ). Bayesian formulations for this setup proceed by describing the uncertainty about γ with a prior $\pi(\gamma)$ on the model space Γ, and the conditional uncertainty about θ_γ given γ with a prior $\pi(\theta_\gamma \,|\, \gamma)$ on the parameter space of model γ. Posterior model probabilities are then obtained via Bayes rule as

$$\pi(\gamma \,|\, Y) = \frac{m(Y \,|\, \gamma)\pi(\gamma)}{\sum_{\gamma\in\Gamma} m(Y \,|\, \gamma)\pi(\gamma)}, \tag{1}$$

where

$$m(Y \,|\, \gamma) = \int f(Y \,|\, \theta_\gamma, \gamma)\pi(\theta_\gamma \,|\, \gamma)\, d\theta_\gamma \tag{2}$$

is the marginal distribution of Y given γ.

The posterior distribution $\pi(\gamma|Y)$ provides a comprehensive post-data representation of model uncertainty which can be used to solve a variety of problems. For example, a commonly used strategy for Bayesian model selection is to select the maximum posterior model [5, 6] or the median posterior model [1]. For the purpose of predicting a quantity of interest Δ, one might use the Bayesian model average

$$E(\Delta \,|\, Y) = \sum_{\gamma\in\Gamma} E(\Delta \,|\, Y, \gamma)\pi(\gamma \,|\, Y) \tag{3}$$

[9]. When the number of models is very large and the full posterior is intractable, attention is usually restricted to a manageable subset of models $S \subset \Gamma$, such as might be obtained by MCMC sampling. In this case, the maximum or the median

Keywords and phrases: model averaging, model selection, objective Bayes, prior distribution, variable selection.

AMS 2000 subject classifications: 62F15, 62J05.

posterior model in Γ would be approximated by the maximum or the median posterior model in S, and the unconditional probability $E(\Delta \mid Y)$ would be approximated by something of the form

$$\hat{E}(\Delta \mid Y) = \sum_{\gamma \in S} E(\Delta \mid Y, \gamma)\hat{\pi}(\gamma \mid Y, S). \tag{4}$$

The potential of the Bayesian approach lies in the formulation of the ingredients on which it is based - the model space, the model space prior and all the parameter priors. Of substantial interest for this problem are the so-called objective prior formulations intended for use as defaults in the absence of bonafide prior information. For this purpose, the development of objective parameter space priors $\pi(\theta_\gamma \mid \gamma)$ has received considerable attention in the literature. In contrast, the development of objective model space priors $\pi(\gamma)$ has received less attention in part because of the availability of some simple and convenient choices. For instance, a simple and commonly used model space prior is the discrete uniform prior, $\pi_U(\gamma) \equiv 1/K$, where K is the number of models in Γ. Although often treated as a natural representation of ignorance, this uniform prior does not account for similarities among the models. As a consequence and as will be seen in the next section, the uniform prior may assign excess probability to neighborhoods of redundant models.

The focus of this paper is on the development of objective model space priors which can compensate for model space redundancy, the so-called dilution priors proposed by George [7]. Instead of assigning prior probability uniformly across models, the goal of such priors is to assign probability more uniformly across neighborhoods of models. The label "dilution" stems from the observation that such priors work by diluting the neighborhood probabilities across the models within them. We begin in Section 2 with an extreme example to illustrate why and where dilution priors may be useful. In Sections 3, 4 and 5, we suggest three very distinct approaches for the general construction of dilution priors. The hope is that these approaches will help pave the way for the future development and implementation of dilution priors for model uncertainty problems.

2. The case for dilution

Suppose the class of linear regression subset models is considered for the relationship between n observations on Y and a set of p predictors $X_1, \ldots, X_p$. In this case, the γth model would be

$$Y = X_\gamma \beta_\gamma + \epsilon \text{ where } \epsilon \sim N_n(0, \sigma^2 I), \tag{5}$$

where the q_γ columns of X_γ correspond to a particular subset of the predictors. An early standard model space prior form for this setup was the independence prior

$$\pi_I(\gamma) = \prod_{i=1}^{p} w_i^{\delta_i}(1 - w_i)^{1-\delta_i}, \tag{6}$$

where $\delta_i = I(X_i \text{ in } \gamma)$ and $w_i = \pi(X_i \text{ in } \gamma)$, e.g., see Clyde, Desimone and Parmigiani (1996), [5, 6] and [10]. Note that the discrete uniform prior $\pi_U(\gamma) \equiv 1/2^p$ is a special case of (6).

To reveal the potential shortcomings of such an independence prior, consider the following extreme example (cf. [7] and [3]). Suppose that X_1 was uncorrelated with

$X_2, \ldots, X_p$, and that $X_2, \ldots, X_p$ were so highly multicollinear that they were all nearly identical proxies for each other. As a result, any subset of $X_2, \ldots, X_p$ would then have an equivalent effect in the model, (a conclusion which does not make use of Y). Effectively, adding $X_2, \ldots, X_p$ to the mix is tantamount to adding an equivalent single new potential predictor.

Thus, a reasonable model space prior for this situation, say π_*, would put $1/2$ probability on the inclusion of X_1 in the model, $1/2$ total probability on the inclusion of any subset of $X_2, \ldots, X_p$ in the model, and would treat these as two independent prior events. Indeed, this would make sense because the 2^p different models here collapse into four sets of equivalent models: the null model, the model containing only X_1, the models containing only at least one of $X_2, \ldots, X_p$, and the models containing X_1 and at least one of $X_2, \ldots, X_p$. Denoting these four sets by $\Gamma_0, \Gamma_1, \Gamma_2, \Gamma_{12}$ respectively, note that π_* would assign probability $1/4$ to each of these mutually exclusive sets.

π_* is a dilution prior in the sense that the probability assigned to each of Γ_2 and Γ_{12} is diluted across the models within these sets. Such dilution is desirable because it maintains the allocation of total prior probability to each of the four sets $\Gamma_0, \Gamma_1, \Gamma_2, \Gamma_{12}$ of equivalent models. For example, the prior probability allocated to the set of all subsets of $X_2, \ldots, X_p$ should remain the same regardless of the size of p.

In sharp contrast to π_*, the uniform prior here would assign $\pi_U(\Gamma_0) = \pi_U(\Gamma_1) = 1/2^p$ and $\pi_U(\Gamma_2) = \pi_U(\Gamma_{12}) = 1/2 - 1/2^p$. Indeed, this and all other cases of the independence priors π_I in (6) lack any dilution property. To see this, consider what would happen if a new X_{p+1}, high correlated with $X_2, \ldots, X_p$, were added to the mix in our example. In effect, all the model probabilities under π_I would be reduced by w_{p+1} for models in which X_{p+1} is included, and by $(1 - w_{p+1})$ for models in which X_{p+1} is excluded. In particular, the probability of the X_1 only model would be reduced by $(1 - w_{p+1})$. And further, if we continued to introduce more proxies for $X_2, \ldots, X_p$, the probability of the X_1 only model could be made arbitrarily small, a disturbing feature if Y was in fact related only to X_1. We note in passing that fully Bayes elaborations of (6) obtained by putting priors on the $w_i's$ would continue to lack any dilution property.

The effect of a dilution prior on the posterior is manifested of course through Bayes rule $\pi(\gamma \mid Y) \propto m(Y \mid \gamma)\pi(\gamma)$. Because the marginal $m(Y \mid \gamma)$ is unaffected by changes to the model space, any dilution effect is therefore controlled completely by the model space prior $\pi(\gamma)$. Indeed, no dilution of posterior probabilities will occur under the uniform prior since it leads to $\pi(\gamma \mid Y) \propto m(Y \mid \gamma)$. Instead the posterior probability of every γ is reduced while all pairwise posterior odds are maintained.

When will dilution priors be useful? Dilution priors avoid placing too little probability on good, but unique, models as a consequence of massing excess probability on large sets of bad, but similar, models. Thus dilution priors may be useful for model averaging over the entire posterior to avoid biasing averages such as (3) away from good, but isolated, models. They also may be useful for MCMC sampling because such Markov chains gravitate toward regions of high probability. Failure to dilute the probability across clusters of many bad models would bias both model search and model averaging approximations (4) toward those bad models. That said, it should be noted that dilution priors would not be appropriate for pairwise model comparisons because the relative strengths of two models should not depend on whether another is considered. For this purpose, Bayes factors (corresponding to selection under uniform priors) seem preferable.

A prior construction that dilutes naturally is the tree generating process prior

proposed by Chipman, George and McCulloch [2] in the context of Bayesian CART model selection. There, each CART model corresponds to a binary tree model T_γ that partitions the range of $X_1, \ldots, X_p$ into regions determined by a sequence of splitting rules associated with the interior nodes of T_γ. The conditional distributions of Y given T_γ are typically simple models such as $Y \sim N(\mu_{\gamma i}, \sigma^2_{\gamma i})$ when $(X_1, \ldots, X_p)$ falls in the ith region determined by T_γ. The dilution prior on T_γ is determined by the following three step tree generating process: (1) 'grow' a tree by random successive splitting, (2) randomly (uniformly) assign available predictors to nodes, and then (3) randomly (uniformly) assign available split points to the assigned predictors. Note that in step (3), if a predictor with many available split points (i. e. realized values) has been assigned to a node, the probabilities associated with the split point assignments will be smaller, thereby downweighting such trees because there are more of them. This is precisely the dilution property at work balancing the prior probabilities across sets of similar modes. It should also be remarked that although Chipman, George and McCulloch [2] use this dilution prior to guide their MCMC search, for model selection they ignore the prior and instead use only the marginal likelihood $m(Y \mid \gamma)$.

Although dilution priors may occasionally arise naturally, such the CART tree generating process, in other settings such as the linear model selection problem above, the construction of dilution priors seems to require at least a bit more crafting. In the next three sections, we will suggest three distinct approaches for the construction of dilution priors.

3. Tessellation defined dilution priors

In this section, we begin with the following tessellation defined construction of a natural class of dilution priors for the linear model variable selection setup (5). Motivated by the idea that a dilution prior should assign uniform probabilities to neighborhoods of models, this class is obtained by identifying such neighborhoods with appropriate tessellations of the surface of a high dimensional sphere. This is facilitated by considering such a specification conditionally on the model dimension q,

$$\pi_V(\gamma) = \pi(\gamma \mid q_\gamma = q)\pi(q), \tag{7}$$

where $\pi(q)$ is any discrete prior with support on $\{0, 1, \ldots, p\}$. Now let

- $A_p \equiv$ surface of the unit radius p dimensional sphere in the space spanned by $X_1, \ldots, X_p$
- $V_\gamma \equiv$ subspace spanned by the columns of X_γ
- $S_\gamma \equiv \{a \in A_p : \| a - V_\gamma \| \leq \| a - V_{\gamma'} \| \text{ for } q_{\gamma'} = q_\gamma\}$

and define

$$\pi(\gamma \mid q_\gamma = q) \propto \mid S_\gamma \mid, \tag{8}$$

where $\mid S_\gamma \mid$ is the area of S_γ.

Essentially, each S_γ is the subset of points on A_p that are closer to V_γ than to any other $V_{\gamma'}$. The set of regions S_γ such that $q_\gamma = q$ forms a Voronoi tessellation of A_p, hence the label π_V. Thus π_V assigns uniform probability to neighborhoods rather than to models, thereby diluting the probability of clusters of similar models. Such uniformity may be a more reasonable representation of "ignorance" than the "ignorance" often associated with π_U.

A useful representation of π_V is obtained by the following "Spinner Process", which provides an intuitive method for sampling from $\pi_V(\gamma)$:

1. First, sample the model dimension q from $\pi(q)$
2. Simulate $Y^* \sim N_n(0, I)$
3. Select the X_γ with $q_\gamma = q$ that is 'closest' to Y^* by minimizing $Y^*[I - X_\gamma(X'_\gamma X_\gamma)^{-1}X_\gamma]Y^*$.

Step 3 is equivalent to selecting the γ that maximizes R^2 for the regression of Y^* on X_γ, and effectively minimizes the smallest angle between Y^* and V_γ. Selection by the random direction of Y^*, which can be thought of as imaginary data, is a natural way of describing prior ignorance about the actual data Y. Intuitively, Y^* is a high dimensional random spinner. This process makes it easy to see that the probability of selecting X_γ is diminished by the presence of other $X_{\gamma'}$ which span a nearby subspace - the dilution property.

The Spinner Process provides a method of constructing $\pi_V(\gamma)$ by repeated simulation, a useful starting point for the study of $\pi_V(\gamma)$. Interesting directions for future research would include the study of $\pi_V(\gamma)$ with different covariance structures, the predictive advantages of using these $\pi_V(\gamma)$ with some of the standard parameter prior formulations, the extent to which $\pi_V(\gamma)$ can be considered to be a least favorable distribution for particular decision theoretic frameworks, and how $\pi_V(\gamma)$ may be extended for different model classes. For example, for a class of generalized linear models, the distribution on Y^* might be changed and the selection distance in step 3 of the Spinner Process might be replaced by the deviance.

To avoid the need for a prior specification of $\pi(q)$, an alternative to the above is the following modified One-Step Spinner Process, which avoids initial sampling of q from $\pi(q)$:

1. Simulate $Y^* \sim N_n(0, I)$
2. Select the X_γ 'closest' to Y^* by minimizing $Y^*[I - X_\gamma(X'_\gamma X_\gamma)^{-1}X_\gamma]Y^*/g(q_\gamma)$

for some decreasing function $g(\cdot)$. Such a $g(q_\gamma)$ serves to adjust the distance between Y^* and X_γ for the dimensionality q_γ. A natural choice is the familiar degrees of freedom correction, $g(q_\gamma) = (n - q_\gamma)$ for which the selection corresponds to maximizing adjusted R^2 for the regression of Y^* on X_γ. More stringent choices such as $g(q_\gamma) = (n - q_\gamma)^2$ would also merit consideration.

From a practical point of view, the calculation of π_V is computationally expensive. Although the simulation of the Spinner Processes can be streamlined by reducing the problem to p dimensions from n dimensions, the real difficulty is that π_V is globally defined. Thus, all 2^p models must be considered for its construction, an impediment to obtaining easily implementable transition kernels $\pi(\gamma \to \gamma')$ for MCMC exploration of the posterior. However, an approximate alternative for MCMC that might be considered is the Local Spinner Process:

1. Simulate $Y^* \sim N_n(0, I)$
2. For $X_{\gamma'}$ in a 'neighborhood' of X_γ, select the one that is 'closest' to Y^* by minimizing $Y^*[I - X_{\gamma'}(X'_{\gamma'}X_{\gamma'})^{-1}X_{\gamma'}]Y^*/g(q_\gamma)$

The choice of neighborhood here might be something like the set $X_{\gamma'}$ obtained by added or deleting a variable from X_γ as is used in the random walk Metropolis algorithm, see Raftery, Madigan and Hoeting (1996). This would at least accomplish dilution within such neighborhoods.

4. Collinearity adjusted dilution priors

Another route to the construction of dilution priors for the linear model is to consider downweighting the probability of γ for the collinearity in X_γ. For each γ, let R_γ be the correlation matrix such that $R_\gamma \propto X_\gamma' X_\gamma$. Note that $|R_\gamma|$ is an overall measure of collinearity. Indeed, $|R_\gamma| = 1$ when the columns of X_γ are orthogonal, and $|R_\gamma|$ decreases to 0 as the columns of X_γ become more redundant. This suggests modifications of the independence prior (6) of the form

$$\pi_R(\gamma) \propto h(|R_\gamma|) \prod_{i=1}^{p} w_i^{\delta_i} (1 - w_i)^{1-\delta_i} \tag{9}$$

for some monotone function h satisfying $h(1) = 1$ and $h(0) = 0$. (Recall $\delta_i = I(X_i \text{ in } \gamma)$). Simple natural choices for h, which controls the size of the downweighting, would be $h(r) = r$ and $h(r) = r^{1/2}$.

Compared to π_V and the related priors above, π_R offers great computational advantages. Except for the norming constant, π_D is easily computable. Thus it is ideal for Metropolis-Hastings algorithms where the norming constant is not needed. Furthermore, fast methods for sequential updating of $|R_\gamma|$ are available, and can be used with proposal distributions that transition by adding or deleting variables . The previous Local Spinner Process may in fact be an effective Metropolis-Hastings proposal with this prior.

Although π_R is a dilution prior in the sense of downweighting redundant models, it is not a dilution prior in the sense of assigning uniform probability to neighborhoods as does π_V. This is because π_R does not account for nearby similar models that use very different sets of variables. Thus, it is more like the Local Spinner Process in that its dilution occurs over some, but not all, nearby models. It should also be mentioned that the penalization of redundancy by π_R may be of interest in and of itself.

5. Model distance based dilution priors

A very different and more general route than the previous dilution prior constructions is obtained by basing it on a distance function between models. Because such a distance can be obtained for any class of models, such constructions are not limited to the set of linear models.

Let $D(\gamma, \gamma')$ be a "distance" between models γ and γ'. Such distances arise naturally. For example, one might consider distances between marginal distributions

$$m(Y \mid \gamma) = \int f(Y \mid \theta_\gamma, \gamma)\, \pi(\theta_\gamma \mid \gamma)\, d\,\theta$$

for particular prior choices of $\pi(\theta_\gamma \mid \gamma)$ such as the Hellinger distance

$$D^H(\gamma, \gamma') = \int [m^{1/2}(Y \mid \gamma) - m^{1/2}(Y \mid \gamma')]^2\, dy.$$

For the linear model class (5) coupled with conjugate priors $\beta_\gamma \sim N(0, \sigma^2 c (X_\gamma^T X_\gamma)^{-1})$, the Hellinger distance yields

$$D^H(\gamma, \gamma') \propto 1 - \frac{|I + cP_\gamma|^{1/4} |I + cP_{\gamma'}|^{1/4}}{|I + (c/2)(P_\gamma + P_{\gamma'})|^{1/2}},$$

where $P_\gamma \equiv X_\gamma (X_\gamma^T X_\gamma)^{-1} X_\gamma^T$.

Of course, many other distances are also available. Again for the linear model class, one might consider distances between subspaces V_γ spanned by the columns of X_γ. When V_γ and $V_{\gamma'}$ are both q dimensional, natural choices might be the geodesic distance

$$D^G(\gamma, \gamma') = (\theta_1^2 + \cdots + \theta_q^2)^{1/2}$$

or the chordal distance

$$D^C(\gamma, \gamma') = (\sin^2 \theta_1 + \cdots + \sin^2 \theta_q)^{1/2},$$

where $\theta_1, \ldots, \theta_q$ are principal angles between V_γ and $V_{\gamma'}$

The challenge is to construct $\pi(\gamma)$ based on any such set of distances $\{D(\gamma, \gamma')\}$. The following potentially promising idea was suggested to me by J. M. Steele [personal communication]. Define

$$\pi_D(\gamma) \propto a_\gamma,$$

where $a_\gamma = \sum_{\gamma'} D(\gamma, \gamma')$, the sum of the distances from γ to every other model. Such a prior will give more weight to distant, isolated models and less weight to models that are closely surrounded by other models. For a simple example, suppose we had three equidistant models. Then

$$\pi_D(\gamma_1) = \pi_D(\gamma_2) = \pi_D(\gamma_3) = 1/3.$$

Suppose now $\gamma_3 \to \gamma_2$ so that $D(\gamma_2, \gamma_3) \to 0$. Then $\pi_D(\gamma_1) \to 1/2$ and $\pi_D(\gamma_2)$, $\pi_D(\gamma_3) \to 1/4$, which is exactly the kind of dilution we want to happen.

Furthermore, the computational features of π_D are appealing. When $|\Gamma|$, the number of models in Γ, is small, $\pi_D(\gamma)$ along with its norming constant can be easily computed. When $|\Gamma|$ is moderate so that at least a_γ can be computed, $\pi_D(\gamma)$ can be easily incorporated into M-H algorithms since

$$\frac{\pi_D(\gamma)}{\pi_D(\gamma')} = \frac{a_\gamma}{a_{\gamma'}}.$$

When $|\Gamma|$ is large so that even computing a single a_γ would be expensive, any a_γ can be easily estimated by

$$\bar{a}_\gamma = \frac{1}{m} \sum a_{\gamma_i}$$

where $\gamma_1, \ldots, \gamma_m$ is an iid sample from $\Gamma - \gamma$. This yields an inexpensive approximation of $\pi_D(\gamma)/\pi_D(\gamma')$ by $\bar{a}_\gamma / \bar{a}_{\gamma'}$ that can again be incorporated into an M-H scheme.

6. Discussion

The contribution of this paper has been to suggest three distinct approaches for the construction of dilution priors, a tessellation defined approach, a collinearity adjustment approach and a pairwise model distance approach. Hopefully these approaches will ultimately lead to the development and implementation of dilution priors. But there is clearly much more work to be done. The main ideas of these approaches have only been outlined here and there are many variations and possibilities that need to be developed and studied. Indeed, it is probably the case that better and richer approaches are needed. With that in mind, I should like to mention a very interesting recent paper by Garthwaite and Mubwandarikwa [4] that further advances the case for dilution priors and proposes some other promising approaches to dilution prior construction based on predictive and empirical Bayes ideas.

References

[1] Barbieri, M. and Berger, J. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897.

[2] Chipman, H., George, E. I. and McCulloch, R. E. (1998). Bayesian CART Model Search (with discussion). *J. Amer. Statist. Assoc.* **93** 935–960.

[3] Chipman, H., George, E. I. and McCulloch, R. E. (2001). The practical implementation of Bayesian model selection. In *Model Selection (with discussion). IMS Lecture Notes – Monograph Series* (P. Lahiri, ed.) **38** 65–134. IMS.

[4] Garthwaite, P. H. and Mubwandarikwa, E. (2010). Selection of prior weights for weighted model averaging. *Austr. N. Z. J. Stat.* To appear.

[5] George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Soc.* **88** 881–889.

[6] George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–373.

[7] George, E. I. (1999). Sampling considerations for model averaging and model search. Invited discussion of "Model Averaging and Model Search, by M. Clyde." In *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 175–177. Oxford Univ. Press.

[8] Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Sci.* **14** 382–417.

[9] Raftery, A. E., Madigan, D. M. and Hoeting, J. (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.* 179–191.

[10] Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econom.* **75** 317–344.