

Bayesian Synthesis

Qingzhao Yu, Steven N. MacEachern and Mario Peruggia
Department of Statistics, The Ohio State University,
Columbus, OH 43210-1247

July 2006

Abstract

Bayesian model averaging enables one to combine the disparate predictions of a number of models in a coherent fashion, leading to superior predictive performance. The improvement in performance arises from averaging models that make different predictions. In this work, we tap into perhaps the biggest driver of different predictions—different analysts—in order to gain the full benefits of model averaging. In a standard implementation of our method, several data analysts work independently on portions of a data set, eliciting separate models which are eventually updated and combined through Bayesian synthesis. The methodology helps to alleviate concerns about the sizeable gap between the foundational underpinnings of the Bayesian paradigm and the practice of Bayesian statistics.

This paper provides theoretical results that characterize general conditions under which data-splitting results in improved estimation which, in turn, carries over to improved prediction. These results suggest general principles of good modelling practice. In experimental work we show that the method has predictive performance superior to that of many automatic modelling techniques, including AIC, BIC, Smoothing Splines, CART, Bagged CART, Bayes CART, BMA, BART and LARS. Compared to com-

peting modelling methods, the data-splitting approach 1) exhibits superior predictive performance for real data sets and simulations; 2) makes more efficient use of human knowledge; 3) selects sparser models with better explanatory ability and 4) avoids multiple uses of the data in the Bayesian framework.

Keywords: Automatic Modelling, Data-Splitting, Human Intervention, Model Averaging.

Qingzhao Yu (email: yu@stat.ohio-state.edu) is Graduate Student, and Steve MacEachern (email: snm@stat.ohio-state.edu) and Mario Peruggia (email: peruggia@stat.ohio-state.edu) are Professors, Department of Statistics, The Ohio State University, Columbus, OH 43210. This material is based upon work supported by the NSF through Awards no. SES-0214574 and SES-0437251 and by the NSA through Award No. MSPF-04G-109.

1 Introduction

A coarse but conceptually useful taxonomy of modelling strategies distinguishes between two broad categories: automatic strategies and strategies which require human intervention. Automatic strategies typically rely on generic methods for model selection, perhaps allowing data-based choice of a couple of tuning parameters. They are appealing because, once the data are input, inferences are produced without requiring any further human interaction. By contrast, human modelling emphasizes exploratory data analysis and the accompanying notions of model development and refinement. The debate on the relative merits of these two *modi operandi* is vigorous and ongoing (see, for example, Breiman, 2001, or Hand, 2006, and the ensuing comments and rejoinders).

In our experience, much of data analysis is heavily based on subjective decisions which do not lend themselves to routine formulations. These range from what variables to include

in an analysis to what forms the variables should take, to insight about the parametric form of the response variable, to whether individual cases should be included in the analysis or trimmed as outliers. Many common instances of human interventions in the modelling cannot be easily carried out by automatic procedures.

Throughout, an adequate analysis must take into account what the variables are, whether they are well-measured or of lesser quality, whether individual influential cases drive the results, what the scientific background of the problem is, etc. (Weisberg, 1985). All of these elements are essential, both when modelling the data formally and when drawing conclusions from the analysis. In certain cases we might specify some aspects of a model and impose specific constraints based on scientific knowledge. For example, the popular Michaelis-Menten model relating the velocity of a chemical reaction to the concentration of a substrate specifies that the expected velocity must be non-negative, monotone and bounded as a function of concentration over the positive half-line. General purpose model selection methods may fail to recognize some or all of these facts and produce inferences based on assumptions that are at odds with the underlying scientific theories.

Because of these reasons, we strongly adhere to the belief that a good data analysis based on *human intervention* will often be far superior to a routinely implemented analysis. In this article we present a modelling strategy, called *Bayesian synthesis*, for combining analyses from several human modelers within the Bayesian framework. Bayesian synthesis, formalized in Section 3, relies on a number of different analysts each contributing a Bayesian model to a pool of models. Each model in the pool is given a weight, thus creating a “hyper-model”. The techniques of model averaging (e.g., Raftery et al., 1997) are used to synthesize the different analysts’ beliefs. Formal rules ensure that the analysts will contribute models that can be synthesized. Bayesian synthesis retains the benefits of subjective modelling while substantially enhancing the inferential and predictive strengths of each individual analy-

sis, producing combined inferences that vastly outperform inferences based on automatic methods.

The methodology we propose can be viewed as a means of constructing a useful space of models over which to perform a Bayesian analysis. In this regard, it is strongly connected to the literature on model selection (e.g., George and McCulloch, 1994, who describe a method of screening models for further development) and on accounting for model uncertainty (see Draper, 1994, and the following discussion for an extensive treatment). In contrast to earlier work, our approach emphasizes the role of subjective modelling and the need for multiple analysts.

The new methodology is developed both through theoretical investigation and through experimentation. The thrust of the theoretical work is a theory that describes an abstraction of subjective modelling, presented in Section 2. Our experimental work, presented in Section 4, compares subjective modelling combined with Bayesian synthesis to automated modelling methods. The results are uniform in demonstrating the success of our new method: in all cases that we considered, human modelling proved superior to a variety of automatic methods, including AIC (Akaike, 1974), BIC (Schwarz, 1978), Smoothing Splines (Craven and Wabba, 1979; Gu, 2002), CART (Breiman et al., 1994), a bagged version of CART (Breiman, 1996), LASSO (Tibshirani, 1996), Forward Statgewise (Hastie et al., 2001), LAR (Efron et al., 2004), Bayesian Model Averaging (Raftery et al., 1997), Bayesian CART (Chipman et al., 1998) and BART (Chipman et al., 2005) and the gains relative to these automated methods were large. In Section 5 we discuss related work and suggest directions for future research.

2 Theoretical Results for Data Splitting

In this section, we present theoretical results that suggest splitting the data can improve the overall analysis. To set up the theory, we ask the reader to envision an applied regression analysis. The analyst examines the data, exploring it and making choices that guide the ensuing analysis. Once these choices are made, they are frozen. As more data accrue, these choices remain frozen, with the additional data used only to update the posterior distribution in the model (or set of models).

In practice, freezing some aspect of the analysis before proceeding is universal. Common choices include selection of variables for the model (including selection of which variables to record), choice of the form for variables (original scale, transformation, discretization), choice of the dependence structure in the model (independence vs. AR models vs. ARIMA models vs. etc.) and choice of a particular parametric form for the response. With introspection, the readers can supply many more examples where they routinely freeze part of the model. Freezing a portion of the model is often justified on the basis that it makes computation simpler and quicker, with the understanding that one pays some price for restricting the scope of the model. In a similar vein, working with only a subset of the data often simplifies and speeds model exploration and model fitting.

2.1 Data-splitting and estimation

We examine the interplay between freezing aspects of the model and data splitting through a series of results. The results compare different analyses. One is the frozen analysis based on the entire data set. The second is the collection of frozen analyses, one for each of the k portions of the partitioned data set. Ultimately, we will cast the comparison of the analyses in terms of a prediction problem and, typically, improved performance should result from

better estimation of the frozen parts of the model. Thus, formal results establishing how data-splitting benefits estimation become relevant. From the full-data analysis, we obtain an estimate of some parameter to be frozen. From the partitioned analyses, we obtain k (different) estimates of the same parameter. The k different estimates are thought of as a grander model that includes k submodels. Focusing on an asymptotic, as more data arrive, a likelihood based analysis will result in either (classical) selection of the best model among the k submodels or (Bayesian) a posterior distribution that gives weight tending to 1 to the best of the submodels. A formal statement of this form requires substantial technical work and considerable notation. We refrain from a full statement here, and refer the reader to Berk (1966) who provides a formal result. We next turn to a formal comparison of full data and split-data estimators.

Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be a random sample drawn from a d -dimensional multivariate normal distribution with mean $\boldsymbol{\theta}$ and variance matrix $(n/k)\boldsymbol{\Sigma}$, where we know the variance and want to estimate $\boldsymbol{\theta}$. Suppose that we split the data set into k parts evenly, so that each part has n/k observations. We estimate $\boldsymbol{\theta}$ by its sufficient statistic, the mean of the sample. Denote the estimator from the i -th split data set by $\overline{\mathbf{X}}_i$ and that from the whole data set by $\overline{\mathbf{X}}$. Then the $\overline{\mathbf{X}}_i$'s are iid multivariate normal random vectors with mean $\boldsymbol{\theta}_d$ and variance matrix $\boldsymbol{\Sigma}$, and $\overline{\mathbf{X}}$ is the mean of the $\overline{\mathbf{X}}_i$'s. Let $1 \leq m \leq d$ and consider the problem of estimating $\boldsymbol{\theta}_m$ (these first m coordinates of $\boldsymbol{\theta}_d$ are the ones that will be frozen). In the following theorem we compare the best estimator from the split data set with that from the whole data set by evaluating their squared Euclidean norms. To simplify the notation, we assume, without loss of generality, that $\boldsymbol{\theta} = \mathbf{0}_d$, we remove the overlying bars from the estimators $\overline{\mathbf{X}}_i$, denoting them by \mathbf{X}_i , and we let $\|\mathbf{X}_i\|_m^2$ denote the squared norm of the first m coordinates of \mathbf{X}_i , i.e., we define

$$\|\mathbf{X}_i\|_m^2 = X_{i1}^2 + X_{i2}^2 + \dots + X_{im}^2.$$

Theorem 2.1 Let $\{\mathbf{X}_i = (X_{i1}, \dots, X_{id})^T\}$, $i = 1, \dots, k$, be k iid d -dimensional normal random vectors with mean $\mathbf{0}_d$ and variance matrix Σ , and let Σ_{ij} be the (i, j) -th element of Σ . Let $Y_1 = \min\{\|\mathbf{X}_1\|_m^2, \dots, \|\mathbf{X}_k\|_m^2\}$ and $Y_2 = \|\frac{1}{k} \sum_{i=1}^k \mathbf{X}_i\|_m^2$.

- 1) Case $m = 1$ Y_2 is stochastically larger than Y_1 .
- 2) Case $m = 2$ If $\Sigma_{11} = \Sigma_{22}$ and $\Sigma_{12} = 0$, then Y_1 and Y_2 have the same distribution. If the condition on Σ does not hold, then Y_2 is stochastically larger than Y_1 .
- 3) Case $2 < m \leq d$ Assume $\Sigma_{11} = \Sigma_{22} = \dots = \Sigma_{mm}$ and assume the first to the m -th elements of the \mathbf{X}_i 's are mutually independent. Then Y_1 is stochastically larger than Y_2 .

Proofs of this and other theorems in this section can be found in the Appendix. Theorem 2.1 shows that, if we decide to freeze a single coordinate, the best estimator from the split data sets is stochastically closer to the true value than the full-data estimator. Improved estimation of this single coordinate-to-be-frozen then results in superior APE performance for the prediction of any given coordinate of the vector.

Having established that data-splitting can be an effective strategy, we turn to the question of how best to split the data.

Theorem 2.2 Let $\{\mathbf{X}_i = (X_{i1}, \dots, X_{id})^T\}$, $i = 1, \dots, n$, be n iid d -dimensional normal random vectors with mean $\mathbf{0}_d$ and variance matrix Σ , and let Σ_{ij} be the (i, j) -th element of Σ . Let $Y_\alpha = \min\left\{\left\|\frac{1}{\lfloor n\alpha \rfloor} \sum_{i=1}^{\lfloor n\alpha \rfloor} \mathbf{X}_i\right\|_m^2, \left\|\frac{1}{n - \lfloor n\alpha \rfloor} \sum_{i=\lfloor n\alpha \rfloor + 1}^n \mathbf{X}_i\right\|_m^2\right\}$, where $\alpha \in [0, 1]$ and $\lfloor X \rfloor$ denotes the greatest integer less than or equal to X .

- 1) Case $m = 1$ As long as neither $\lfloor n\alpha \rfloor = \lfloor \frac{n}{2} \rfloor$ nor $\lfloor n(1 - \alpha) \rfloor = \lfloor \frac{n}{2} \rfloor$, Y_α is stochastically larger than $Y_{1/2}$.
- 2) Case $m = 2$ Assume that $\Sigma_{11} = \Sigma_{22}$ and X_{i1} and X_{i2} are independent. Then Y_α has the same distribution for all $\alpha \in [0, 1]$.

- 3) Case $2 < m \leq d$ Assume that $\Sigma_{11} = \Sigma_{22} = \dots = \Sigma_{mm}$ and the first to the m -th elements of the \mathbf{X}_i 's are mutually independent. Then Y_α is stochastically increasing in α , for $\alpha \in [0, 0.5]$.

Theorem 2.2 can be extended to the case when the data are split into more than two parts, yielding similar results. If we fix only one coordinate, it is optimal to split the data evenly.

The next theorem deals with the number of splits. For the case where data-splitting has been shown to be effective (cf. Theorem 2.1), we find that more splits are better.

Theorem 2.3 Let $\{\mathbf{X}_i = (X_{i1}, \dots, X_{id})^T\}$, $i = 1, \dots, n$, be n iid d -dimensional normal random vectors with mean $\mathbf{0}_d$ and variance matrix Σ , and let Σ_{ij} be the (i, j) -th element of Σ . Assume that n is divisible by a . Let $Y_a = \min_{i=1,2,\dots,a} \left\{ \left\| \frac{a}{n} \sum_{j=n(i-1)/a+1}^{ni/a} \mathbf{X}_j \right\|_m^2 \right\}$.

- 1) Case $m = 1$ Y_a is stochastically decreasing in a .
- 2) Case $m = 2$ Assume that $\Sigma_{11} = \Sigma_{22}$ and that X_{i1} and X_{i2} are independent. Then Y_a has the same distribution for all a .
- 3) Case $2 < m \leq d$ Let m be an integer with $2 < m \leq d$. Assume that $\Sigma_{11} = \Sigma_{22} = \dots = \Sigma_{mm}$ and that the first to the m -th elements of the \mathbf{X}_i 's are mutually independent. Then Y_a is stochastically increasing in a .

In practice, few estimators are normally distributed. However, many estimators are asymptotically normal. Defining asymptotic stochastic ordering to be stochastic ordering of appropriately centered and scaled limiting distributions, Theorems 2.1 through 2.3 extend to the case where estimators are asymptotically normal. These strengthened results capture the behavior of typical estimators when parametric models are *fit* to the data; the data generating mechanism need not be “parametric”. Adapting Bayesian versions of the central limit theorem (see, for example, Berger, 1985 and references therein), the results also apply to a large class of Bayes estimators. The strengthened results are to be found in Yu (2006).

2.2 The impact of data-splitting on prediction

Results 2.1-2.3 establish the benefits of data-splitting for parameter estimation, while our main concern is prediction. To this end, we consider a formal asymptotic evaluation of the predictive fit of a model. Here, we distinguish the variable to be predicted as Y with the remaining variables denoted by X .

Definition 2.4 For a sequence of random vectors $(X_i, Y_i), i = 1, 2, \dots$, associated predictors \hat{Y}_i , and loss function $L(Y_i, \hat{Y}_i)$, the asymptotic predictive evaluation (APE) is

$$APE_m = \lim_{n \rightarrow \infty} E \left[n^{-1} \sum_{i=m+1}^{m+n} L(Y_i, \hat{Y}_i) \right], \text{ and } APE = \lim_{m \rightarrow \infty} APE_m,$$

provided the limits exist and are finite.

Conditions are required on the random vectors and estimators to ensure that APE exists. For a relatively straightforward example, consider the simple linear regression problem.

Example 2.1 The data consist of an iid sequence of pairs (X_i, Y_i) . The model used for analysis is correct. It is $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, $E[\epsilon_i] = 0$, X_i and ϵ_i are independent, and the (X_i, ϵ_i) pairs are iid. The estimator \hat{Y}_i is the plug-in estimator $\hat{\beta}_0 + \hat{\beta}_1 X_i$, with $(\hat{\beta}_0, \hat{\beta}_1)$ the least squares estimator based on $(X_1, Y_1), \dots, (X_{i-1}, Y_{i-1})$. We take the loss function L to be quadratic. To ensure that the expected loss is finite, we impose the conditions that $\sigma^2 = \text{var}(\epsilon_i)$ and $\sigma_x^2 = \text{var}(X_i)$ be finite. Thus, $E[L(Y_i, \hat{Y}_i)] = E[\epsilon_i^2] + E[(\beta_0 - \hat{\beta}_0 + (\beta_1 - \hat{\beta}_1)X_i)^2]$ is finite, provided $\hat{\beta}_0$ and $\hat{\beta}_1$ have finite variances. To establish that these variances are finite, we must fully define our estimator. Set the estimator to $(\hat{\beta}_0, \hat{\beta}_1) = (0, 0)$ if there is no unique least squares solution. Add the condition that X_i is continuous with $\sup_x f_X(x) < \infty$, and we have $APE = \sigma^2$.

Our subsequent concern will be with analyses where a portion of the model is frozen.

Specifically, suppose that the intercept is frozen at $\beta_0 = 0$. In other words, suppose that we incorrectly fit a regression through the origin. Then, the least square estimator of β_1 based on $(X_1, Y_1), \dots, (X_{i-1}, Y_{i-1})$ is $\hat{\beta}_1 = (\sum_{j=1}^{i-1} x_j y_j) / \sum_{j=1}^{i-1} x_j^2$, which converges, as i goes to infinity, to $\beta_1^* = \beta_1 + \beta_0 \mu_x / (\sigma_x^2 + \mu_x^2)$, where $\mu_x = E(X_i)$. Under the model and conditions given above, $APE = \sigma^2 + E[(\beta_0 + (\beta_1 - \beta_1^*)X_i)^2] = \sigma^2 + \beta_0^2 [1 - \mu_x^2 / (\sigma_x^2 + \mu_x^2)]$. Thus, the asymptotic performance measured by APE is an increasing function of $|\beta_0|$ which attains its minimum when the restricted model is correct, i.e., when $\beta_0 = 0$. This is in agreement with the fact that the model misspecification implied by freezing the intercept at $\beta_0 = 0$ becomes more serious as $|\beta_0|$ gets larger.

APE also applies to multiple regression problems, as the next example shows.

Example 2.2 The data consist of *iid* multivariate normal vectors (\mathbf{X}_i, Y_i) where \mathbf{X}_i is of dimension $d - 1 > 1$. We will freeze our estimate of the mean for the first coordinate. Thus, after fixing $\theta_1^* = \bar{X}_1$ from the initial sample, as $n \rightarrow \infty$, the estimator for $\boldsymbol{\theta}$ tends to $\hat{\boldsymbol{\theta}}^* = (\bar{X}_1, \theta_2 + \Sigma_{21}\Sigma_{11}^{-1}(\theta_1 - \bar{X}_1), \dots, \theta_d + \Sigma_{d1}\Sigma_{11}^{-1}(\theta_1 - \bar{X}_1))$ for the full sample analysis and to $\hat{\boldsymbol{\theta}}_i^* = (\bar{X}_{i1}, \theta_2 + \Sigma_{21}\Sigma_{11}^{-1}(\theta_1 - \bar{X}_{i1}), \dots, \theta_d + \Sigma_{d1}\Sigma_{11}^{-1}(\theta_1 - \bar{X}_{i1}))$ for each split sample analysis. We apply Definition 2.4 to prediction of Y_i and obtain that, for the full sample analysis, $APE = \sigma^2 + [\Sigma_{d1}\Sigma_{11}^{-1}(\theta_1 - \bar{X}_1)]^2$, while, for the split sample analysis, $APE = \sigma^2 + \min_i [\Sigma_{d1}\Sigma_{11}^{-1}(\theta_1 - \bar{X}_{i1})]^2$, where $\sigma^2 = \text{var}(Y_i | \mathbf{X}_i)$. From Theorem 2.1, we have that $(\theta_1 - \bar{X}_{i1})^2$ is stochastically smaller than $(\theta_1 - \bar{X}_1)^2$. Thus the split-sample analysis provides better asymptotic predictive performance than does the full sample analysis. Similarly, when two coordinates are frozen, data-splitting yields a stochastically better APE than does a full sample analysis. When more than two coordinates are frozen, the comparison of the APE s depends on how the unfrozen coordinates are related to the frozen coordinates and on how the frozen coordinates are related to each other.

There are many additional benefits to data-splitting. Some can be addressed with a formal treatment. MacEachern, Peruggia and Guha (2003) demonstrate the benefits of discarding data when data are analyzed with a computational algorithm that is more costly than $O(n)$. Fligner and MacEachern (2003) note that there is no loss of asymptotic relative efficiency due to data-splitting for estimators whose variance is $O(n)$. MacEachern and Peruggia (2000) show how a single, judicious data split can produce an estimator of smaller variance than one based on the entire data set. Some benefits cannot be addressed with a formal treatment, such as the belief (which we hold) that the large-scale features that an analyst extracts from data are relatively stable once an adequate amount of data is available.

3 A Bayesian Framework for Data-Splitting

Our primary focus is on Bayesian modelling, where a team of analysts builds models for a data set. The paradigm we envision is this. First, the data are split into several portions. Each analyst receives one portion of the data. Second, each analyst builds a Bayesian model for their portion of the data, reporting a “Bayesian summary” of their posterior distribution. Third, the Bayesian summaries are updated on portions of the data not used to build them, and they are combined to yield a single, overall posterior model.

Two features are essential for this procedure to work well. First, each analyst must produce a Bayesian summary that is amenable to updating with further data. Second, the various Bayesian summaries must be amenable to synthesis. Throughout, we must exercise care so that the data are not split into too many parts. The theorems of the preceding section provide some guidance on this issue. Throughout this section, we will assume that there are k analysts.

Splitting the Data. The data to be used for model development and synthesis are split

into k portions. Once split, the portions of the data are assigned to the k analysts at random. This produces an exchangeable partition and assignment of data to analysts. The theoretical results of the previous section suggest that (where data splitting is appropriate) the portions of the data should all contain approximately the same amount of information about the data-generating process. Following this theory, we seek to produce a set of splits that give conditionally iid data to the analysts. The following cases describe two of the splitting procedures that we have implemented.

The first case is that of a designed experiment where a structural balance is forced upon the data. For example, the two-sample, completely randomized design is often implemented in a balanced fashion, so that the same number of experimental units are assigned to each of the two treatment conditions. Additionally, covariates are recorded on the experimental units. For this type of experiment, we split at random, with the restriction that each analyst receive the same number of observations on each treatment. The additional covariates need not be balanced and need not be used by the analysts in constructing a model for the data.

The second case, matching the ozone example of Section 4, is one where there is a collection of experimental units, with a variety of information on each unit. In this case, we split the data at random, with each analyst receiving the same number of observations.

These methods of splitting the data have the advantage of not depending on the analysts' eventual models—an essential part of our paradigm. The methods are extremely easy to implement and do not require the help of an expert to split the data. The drawback to these methods is that the portions of the data will typically not convey the same amount of information to the different analysts. While “optimal” splits (see, for example, Theorem 2.2), might well differ, we would need to know the details of the analysts' models to formalize the notions of information in the splits and of optimality. For large samples, the splits of the data will contain approximately the same amount of information.

Building and Updating the Model. In order to carry out the analysis, each analyst is provided with a set of ground rules for model building. The rules include, most importantly, the task of modelling. Second, the analyst must know what kind of Bayesian summary to produce. Since synthesis of the analysts' summaries will be accomplished through Bayes factors, and since Bayes factors depend on the marginal likelihood of the data, the analyst must be informed of the quantity for which the likelihood will be calculated. Third, the analyst must know what conventions will be followed for computation of the likelihood. These conventions must guarantee that the analysts' models will be mutually absolutely continuous over the range of values that the data can assume.

Consider the prototypical experiments for which data splitting is described. In the first case, of a balanced two-sample experiment with case-specific covariates, interest may focus on the difference between treatment means. Implicitly, the analysts have been informed that the treatment means exist. The Bayesian summary for an analyst represents the analyst's posterior, given the portion of the data used for the analysis. The likelihood of responses to the two treatments will be computed; the mechanism assigning units to the treatments will not be part of the likelihood. The convention for the likelihood is that it be a density absolutely continuous with respect to Lebesgue measure with support on the real line. An alternate convention might be that the likelihood be discrete, rounded to a single decimal place, on the non-negative half-line.

An instance of the second case is described in some detail in the upcoming example, and so we leave off discussion for the moment. In any event, each analyst is left with the choice of constructing a model from the assigned portion of the data. The analyst may use any method whatsoever to build their model, ranging from automated methods, to subjectively elicited priors, to construction and refinement of models through diagnostics. The essence of the paradigm is to encourage the analysts to build creative models that can be combined

across analysts.

The Bayesian Summary. The Bayesian summary can take on a wide variety of forms, depending on the analyst’s modelling choices. Whatever the form, the summary must be amenable to updating and allow one to compute the marginal likelihood for the portions of the data not used to construct the model.

Several forms of summary work well in practice. Choice of a posterior distribution conjugate to the analyst’s chosen likelihood for the future data leads to a direct computation of the marginal likelihood. Choice of a collection of such distributions leads to a mixture of conjugate posteriors, and hence to quick computation of the marginal likelihood. For models that move beyond conjugacy, the posterior distribution can be represented in a discrete fashion, for example, by the output of a Monte Carlo simulation. Along with the representation, the summary must include a means of updating the summary, e.g., code to compute the marginal likelihoods and to produce summaries that enable one to address the inferential goals of the analysis.

Synthesizing the Analyses. When each analyst has produced a model, we can combine them to yield an overall model. We combine the models by computing pairwise Bayes factors for portions of the data and then reconciling them by passing to marginal likelihoods, eventually arriving at a single marginal likelihood for each analyst. These marginal likelihoods determine the weight that each analyst receives in predictions. Let Y_1, \dots, Y_k denote the k splits of the data; let f_1, \dots, f_k denote the likelihoods for the k models with possibly differing parameters $\theta_1, \dots, \theta_k$.

The pairwise Bayes factor is computed on the greatest set of data not used in constructing the two models, after the two models have been updated to include the same data. Thus,

the Bayes factor comparing analysts 1 and 2 is

$$B_{12} = \frac{\int f_1(Y_3, Y_4, \dots, Y_k | \boldsymbol{\theta}_1) \pi(\boldsymbol{\theta}_1 | Y_1, Y_2) d\boldsymbol{\theta}_1}{\int f_2(Y_3, Y_4, \dots, Y_k | \boldsymbol{\theta}_2) \pi(\boldsymbol{\theta}_2 | Y_1, Y_2) d\boldsymbol{\theta}_2} = \frac{m_{1(2)}}{m_{2(1)}}.$$

Note that the distribution on $\boldsymbol{\theta}_1$ used in the above calculation is the posterior, given both Y_1 and Y_2 . Similarly, the distribution on $\boldsymbol{\theta}_2$ is the posterior given both Y_1 and Y_2 .

If the Bayesian summaries yield models that are each well-represented by a set of N draws from the appropriate posterior distribution, the Bayes factor can be computed as

$$\hat{B}_{12} = \frac{\sum_{j=1}^N N^{-1} f_1(Y_3, Y_4, \dots, Y_k | \boldsymbol{\theta}_1^{(j)})}{\sum_{j=1}^N N^{-1} f_2(Y_3, Y_4, \dots, Y_k | \boldsymbol{\theta}_2^{(j)})}.$$

Weighted distributions, such as those produced by importance sampling, can be used to obtain the Bayes factor. For more complex models, sophisticated methods of estimating the marginal likelihoods produce these Bayes factors. See Chen et al. (2000) for a recent book that describes methods for estimating Bayes factors/marginal likelihoods.

We recognize $\log(BF_{ij}) = \log(m_{i(j)}) - \log(m_{j(i)})$ and average the log marginal likelihoods across all $j \neq i$ to obtain $m_i = \exp((k-1)^{-1} \sum_{j=1, j \neq i}^k \log(m_{i(j)}))$. These m_i are then used as weights to yield the synthesized posterior: $f(\boldsymbol{\theta} | Y) = \sum_{i=1}^k m_i f(\boldsymbol{\theta}_i | Y_1, \dots, Y_k) / \sum_{j=1}^k m_j$. In this expression, $\boldsymbol{\theta}$ runs over the parameter spaces for all of the analysts' models.

4 Applications

In this section, we describe an experiment which demonstrates the benefits of Bayesian synthesis. To conduct the experiment, we selected a data set which has been used by other authors to illustrate the benefits of automated modelling methods. None of us was familiar with the data set and we each received one third of the data. This allowed us to create three

pairs of analysts, with one third of the data reserved for evaluation of the pair’s synthesis. The Bayesian syntheses were compared to a variety of automated procedures. We find that Bayesian synthesis outperforms all of these methods.

4.1 Ozone Data

The ozone data set consists of daily measurements of ozone concentration and eight meteorological quantities in the Los Angeles basin for 330 days in the year 1976. Breiman (2001) describes the origin of the data set. The data set is contained and documented in the software package R. The data frame contains 330 observations on the following variables: *upo3* – upland ozone concentration, in ppm; *vdht* – Vandenberg 500 millibar height, in meters; *wdsp* – wind speed, in miles per hour; *hmdt* – humidity; *sbtp* – Sandburg air base temperature, in degrees Celsius; *ibht* – inversion base height, in feet; *dpgg* – Daggett pressure gradient, in mmhg; *ibtp* - inversion base temperature, in degrees Fahrenheit; *vsty* - visibility, in miles; *day* – calendar day, an integer number between 1 and 366.

Each analyst was charged with the task of constructing a Bayesian model that can be used to predict ozone concentration. Each model should produce a distribution for ozone concentration supported on the non-negative integers.

4.1.1 The Split-Data Analysis

We split the data into three sets of 110 observations each with a complete randomization. Each of us (Analysts 1-3) received one part of the data (data 1-3). All three analysts decided to model log ozone level as a continuous variable and to produce the agreed-upon distribution for ozone (over the positive integers) by integrating the continuous density of the modeled variable.

Model 1. Analyst 1 used data set 1 to build a model, pursuing a strategy of first discovering

which variables appeared to be important in predicting ozone level and then determining the forms in which the variables should enter the model.

Matrices of scatter plots of the response variable and explanatory variables were examined. Serial dependence was investigated by including lagged responses as explanatory variables. Several variables (*sbtp*, *ibht*, *vsty* and *day*) appeared to be quite important, and so were chosen to appear in the models. There was no apparent serial dependence in the data, after adjusting for other variables.

Having identified important variables, the analyst searched for appropriate forms. The term *ibht* was modeled as four variables, a linear term, two further variables developed to capture non-linearity, and an indicator for $ibht = 5000$, an apparent truncation point for the variable. The indicator allows for the jump that we expect at the truncation point and provides a way to incorporate additional variability at this point. The analyst forced the effect of variable *day* to be periodic with period 1 year.

After basic models were created, the analyst reexamined variables previously judged to be of lesser import with added variable plots and best subsets regressions. The variable *hmdt* was included as a predictor, in a piecewise linear fashion. The variables *dgpq* (with linear and quadratic terms) and *vdht* were considered to be potential predictors. Plots of *vsty* showed a wiggly pattern of nonlinearity. Two forms for this effect were considered—a linear effect and a Gaussian process centered at a linear effect. The prior on the Gaussian process version was chosen to force the realized effect curve to be close to linear.

Finally, eight models (all including the initial variables and *hmdt*; then the 2^3 combinations including or excluding *dgpq* and *vdht* and with two forms of prior for *vsty*) were selected to receive positive probability. The prior distribution on each model was improper, uniform for some coefficients and vague for most other coefficients. Weights were formed for the eight models through estimated likelihoods. Each model was updated with 99 cases and a

Table 1: Weights for Analyst 2’s component models, given data set 2.

	CAR 1	CAR 2
main effects	0.4	0.3
main effects plus interactions	0.2	0.1

predictive likelihood computed for the remaining 11 cases. This process was repeated 10 times, yielding ten predictive likelihoods. The weight given to each model was proportional to the geometric mean of its predictive likelihoods.

Model 2. Based on data set 2, Analyst 2 plotted log ozone concentration and all other covariates against “day” to detect evident trends. The response and the covariates were each detrended through local fitting (by means of the `loess()` function in R) using the variable “day” as a predictor. All subsequent modelling was conducted on the residuals from these fits.

Analyst 2 believed that time proximity might constitute an important factor and decided to specify conditional autoregressive (CAR) models for the detrended data. Denoting the response variable by Y , a CAR model takes the form $Y_t \sim Normal(\mu_t, \sigma^2)$, where $\mu_t = \mathbf{X}'_t\boldsymbol{\beta} + \boldsymbol{\theta}_t$, with \mathbf{X}_t denoting a vector of covariate values at time t and $\boldsymbol{\beta}$ a vector of model parameters. The models specified random walk priors of order either one or two for the vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{366})'$, as explained in Thomas et al. (2004).

Analyst 2 built two models for the regression $\mathbf{X}'_t\boldsymbol{\beta}$. The first has an intercept and four main effects selected by means of graphical and exploratory data analysis techniques. The second has many more predictors selected through a stepwise procedure, starting from the model with all main effects and two-way interactions. The two regression models and the two CAR structures were combined to produce four models that were averaged according to weights given in Table 1. The weights were chosen subjectively to reflect the analyst’s higher degree of confidence in simpler rather than more complicated models. Non-informative priors were specified for the model parameters and Winbugs was used to draw separate samples

from the posterior distributions for the four models.

Model 3. Analyst 3 used data set 3 and applied a modification of Least Angle Regression (LARS; Efron et al., 2004) to fit the model: first modified LARS was used to choose the variables to be included in the model, and then Bayesian linear regression was implemented to quantify the relationship between log ozone concentration and the selected variables.

Two modifications are applied to LARS. The first is the restriction that an interaction term can be selected only after the corresponding main effects have entered the model. As soon as the main effects enter, the interaction term becomes a candidate variable. The second modification to LARS is that some variables (in this analysis, one main effect) are forced to enter the model at the beginning of the procedure.

Assume there are p candidate main effects. Order these variables by the strength of their correlation with the response variable, from strongest to weakest. Label the ordered variables $1, \dots, p$. Suppose that variable 2 will be forced into the model. We start with only variables 2 through p as candidate variables, and so LARS selects variable 2. We continue with the solution path until another variable is added. At this point, the list of candidate variables is expanded to include variable 1 and the second order term for variable 2. A second variable is chosen from the list of candidate variables according to the LARS criterion. Then the second order term for this variable and its interaction with variable 2 are included as candidate variables. The above process is repeated until the solution path is completed.

Analyst 3 used modified LARS to decide, with different forced in variables, the order in which variables entered the models. This produced several sequences of models. Each sequence was examined by C_p and by differences in AIC and BIC to subjectively determine which models were viable. A Bayesian linear regression was computed for each viable model, against an improper prior distribution. Finally, BIC was used to obtain a weight for each of the four models. With new data, both the weight for each model and the distributions of

parameters within the model were updated.

4.1.2 Human Modelling versus Automated Modelling

Many authors have advocated the use of automated modelling strategies, arguing that such methods provide better predictive performance than corresponding subjectively built models. Breiman et al. (1984) and Gu (2002) analyze the ozone data with the goal of predicting ozone concentration. Using the methods described in their work, we reanalyzed the data, comparing their predictive performance to that of the single and combined models of Analysts 1-3.

Table 2 contains three comparisons. For each comparison, one split of the ozone data is reserved as test data, with the other two splits used to fit the models. The automated methods include CART, a bagged version of CART, smoothing splines (SS), and AIC and BIC applied to linear regressions with the original variables as candidate predictors. Previous analyses of these data have focused on prediction of log ozone, leading to our choice of log ozone as the response variable. The accuracy of predictions is measured in terms of sum of squared prediction errors.

The table presents results on two versions of the prediction problem. The first is a static prediction problem; the latter a sequential prediction problem. For the static problem, the training data were used to develop the model. A prediction was made for each case in the test data, and the measure of fit was computed. We refer to this as making a prediction “once and for all.” For the sequential problem, we randomly partitioned the test data into 11 sets of 10 cases each. The model was fit to the training data, and a prediction made for the first set of cases in the test data. The model was updated (getting the posterior distributions both within and across models) based on the first set of cases in the test data, and predictions made for the second set of cases. This procedure was continued, updating the model on successively larger sets of data and making predictions for the next set of cases,

Table 2: Comparison of Automatically Fitted Models with Human Models by Sum of Squared Errors for Log Ozone. Mean Human and Bayesian Synthesis Outperform All of the Automatic Methods.

Test Data	data split 1		data split 2		data split 3	
Updating Method	Once	10 by 10	Once	10 by 10	Once	10 by 10
ANALYST 1	-	-	12.76	12.72	14.96	14.25
ANALYST 2	18.00	17.48	-	-	12.13	12.03
ANALYST 3	15.96	16.07	14.21	14.32	-	-
MEAN HUMAN	16.98	16.78	13.49	13.52	13.55	13.14
BAYESIAN SYNTHESIS	16.00	16.31	12.50	13.11	12.10	11.89
AIC	21.96	21.98	16.63	16.96	16.75	16.63
BIC	21.51	21.54	17.44	17.69	16.75	16.63
SMOOTHING SPLINE	26.85	26.39	17.60	16.73	18.01	15.96
CART	27.51	28.43	17.87	17.01	19.37	19.51
BAGGED CART	19.63	19.16	14.94	14.19	16.28	15.51
BAYES TREE	24.42	21.43	21.71	20.18	20.07	20.15
BART	24.70	23.00	23.51	23.09	20.49	20.81
BMA	21.96	21.89	17.61	17.67	16.90	16.40
LAR	21.34	21.41	17.36	17.57	17.30	17.20
LASSO	21.32	21.74	17.48	17.98	19.40	17.45
FORWARD STAGEWISE	21.10	21.28	17.37	17.88	17.11	17.45

until the test data were exhausted. We used the same partition of the test data (in the same order) to evaluate each of the methods. We refer to this as “ten by ten” evaluation.

Table 2 contains rows for the “Mean Human” and for “Bayesian Synthesis.” The “Mean Human” is defined by selecting an analyst to make predictions at random. The measure of fit is the mean of the two analysts’ measures. “Bayesian Synthesis” implements the method of Section 3, combining the two analysts eligible to make predictions for the test data. The initial weights given to each analyst are equal. When updating “ten by ten,” the weights adjust, based on the relative performance of the analysts’ models. The predictions were taken to be the posterior predictive means.

Table 2 shows the success of data splitting and of human modelling. We first note that the “Mean Human” provides a better predictive fit than do any of the automatic methods. It corresponds to randomly selecting an analyst to develop a model. This comparison establishes the benefit of subjective modelling.

Second, we turn to the main purpose of the experiment—to see whether Bayesian Synthe-

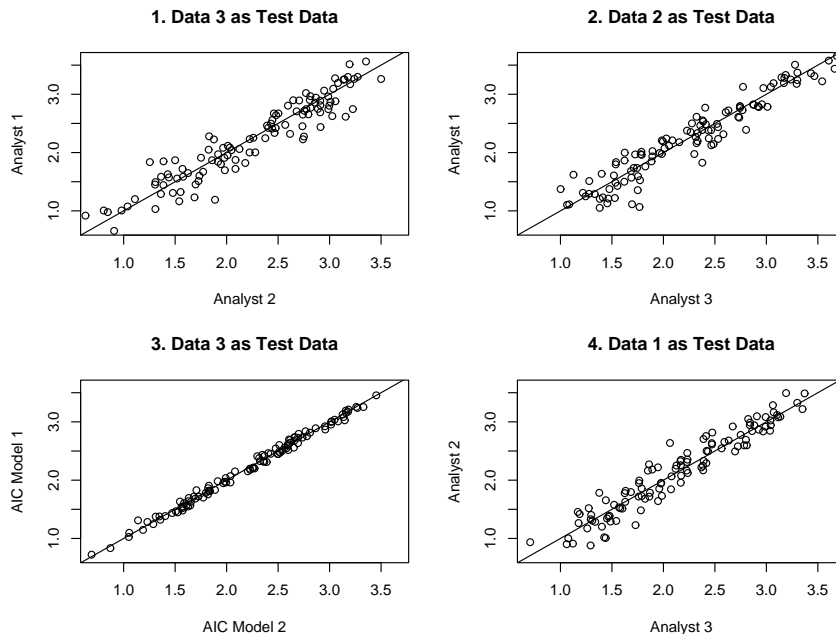


Figure 1: Plots showing the differences in model predictions.

sis outperforms rival methods. In every instance we find that the method does outperform competing procedures. Bayesian Synthesis yields much smaller predictive mean square errors than any of the automatic methods. The predictive mean square error is also smaller than the “Mean Human”. The comparisons also show the magnitude of the benefit to human modelling. The differences between amongst the best of the automated techniques are considerably smaller than the difference between these automated techniques and Bayesian Synthesis.

Third, the comparison between the static and sequential problems shows, on the whole, a modest benefit to continually updating the model. It also makes clear the dominant role that modelling plays in effective prediction—building a better model (more precisely, a better collection of models) is far more important than having a bit more data with which to update the model.

We next turn to an explanation of the benefits of model synthesis. Bayesian synthesis, indeed all Bayesian model averaging, provides the greatest benefits when the models to be

synthesized provide different predictions. It is here that averaging allows one to make a different prediction than either model, and it is here that further information collected in data allows the posterior weights given to different models to select the better model. The benefits of bagging/averaging models arising from relatively stable procedures such as AIC, BIC and SS are minimal (results not presented in table), because the bulk of the bagged models provide the same or similar predictions. Figure 1 shows that differences in predictions from different analysts show more variation than do differences from different AIC models.

The results outlined in Table 2 show clearly that there are large benefits stemming from human modelling with additional improvements attributable to Bayesian synthesis. Interestingly, large benefits can also ensue from Bayesian synthesis of a human and an automatically fitted model, as evidenced by the summaries presented in Table 3. This is in part due to the fact that the predictions produced by human and automatically fitted models are typically different. Also, the gains appear to be more sizeable when the human models are synthesized with methods based on the creation of new variables (e.g., Smoothing Spline, CART, Bagged CART, BART) than when they are synthesized with methods based on regressions with the original variables (e.g., AIC, BIC, BMA, LAR, LASSO, Forward Stagewise). Overall, the empirical results confirm the indication provided by the stylized theoretical results of Section 2: The predictions produced by a Bayesian synthesis usually outperform the predictions of the single constituent elements and inherit many of the performance properties of the best of the constituent elements.

4.2 Simulations

To further explore the performance of data-splitting, we designed a simulation study so that we could compare the models arising from data splitting to known, true models. We followed the simulations of Raftery et al. (1997) for models (linear regression), patterns of covariates

Table 3: Comparisons of Model Combination by Sum of Squared Error. The entries in the table are the sum of squared prediction errors for the given “Test Data” based on “once and for all” updating using Bayesian synthesis. The columns headed “none” report the results for individual methods without synthesis.

Test Data	Data Set 1			Data Set 2			Data Set 3		
Combined With Analyst	2	3	none	1	3	none	1	2	none
Analyst 1	-	-	-	-	12.50	12.76	-	12.10	14.96
Analyst 2	-	16.00	18.00	-	-	-	12.10	-	12.13
Analyst 3	16.00	-	15.96	12.50	-	14.21	-	-	-
AIC	18.22	17.16	21.96	14.21	14.46	16.63	14.43	13.05	16.75
BIC	18.45	17.50	21.51	14.45	14.77	17.44	14.43	13.05	16.75
SMOOTHING SPLINE	17.11	17.54	26.85	14.06	14.02	17.60	14.12	12.61	18.01
CART	18.84	18.57	27.51	12.67	12.76	17.87	13.26	12.59	19.37
BAGGED CART	16.71	16.33	19.63	12.75	13.08	14.94	13.50	12.33	16.28
BAYES TREE	19.09	18.71	24.42	14.91	15.62	21.71	14.75	12.99	20.07
BART	16.58	16.02	24.70	15.51	15.90	23.51	14.36	12.93	20.49
BMA	18.60	17.67	21.96	14.53	14.82	17.61	14.50	13.09	16.90
LAR	18.45	17.47	21.34	14.29	14.71	17.36	15.62	14.05	17.30
LASSO	18.44	17.46	21.32	14.09	14.32	17.48	14.63	13.16	19.40
FORWARD STAGEWISE	18.37	17.34	21.10	14.27	14.67	17.37	15.62	14.05	17.11

and sizes of data sets. In the simulations, we split the data set into 2 parts. To automate the process of model building, we chose models via AIC and via BIC for each portion of the data, and also for the full data set. Posterior distributions for the selected models were computed against a vague prior.

The results of the simulation were examined in two fashions. The first was whether data splitting captured the correct model. When AIC was used for the full data set, it tended to over fit, selecting too many variables for inclusion in the model. Performance was enhanced with data splitting, where, with a smaller number of cases, over fitting was not as great a problem. BIC led to fine performance for the full data set. For the split data sets, BIC tended toward modest under fit. Combining the results across criteria, the pattern emerges that data splitting is most effective when used with criteria that tend to over fit models.

The second analysis of the simulation focused on predictive performance. Under this criterion, AIC applied to the split data sets led to better predictive performance than did AIC applied to the full data set. BIC applied to the split data sets performed on a virtual

par with BIC applied to the full data set. The strength of data splitting arose when the models based on the split data sets were combined. When the models were combined, with both AIC and BIC used to select models, data splitting combined with model synthesis led to greater predictive accuracy than did an analysis of the full data set.

The simulations suggest strategies for subjective modelling that work well with data splitting. The strategies are to include a model that cuts across models, and to move modelling in the direction of “over fitting”, realizing that the additional data moderate the effects of an over fit and realizing that other analysts can rescue one from over fitting. We note that this advice parallels “staking out the corners of model space” (Draper et al., 1987).

5 Discussion and Further Research

In this paper, we propose Bayesian synthesis, a new paradigm for Bayesian data analysis. The paradigm is motivated by the concern that using a set of data both to develop a model and to subsequently fit the model with the same data violates the spirit of Bayes theorem. The paradigm has been developed with an eye to which parts of a modelling effort appear to be stable—model development by a single analyst—and which appear to yield highly variable results—model development by different analysts. Tapping into the variable parts of an analysis while retaining enough information to preserve stability of the other parts of the analysis allows us to obtain the greatest benefits of Bayesian model averaging. This also provides us with a more appropriate accounting of model uncertainty.

We have explored the new paradigm experimentally and theoretically. Experimentally, the ozone data analysis shows the remarkable benefits that accrue to subjective modelling and the further benefits that follow from synthesizing subjective models across analysts. On the theoretical side, the series of results in Section 2 provide a basic theory that justifies split-data analyses.

This work raises several issues. One issue is how to most effectively split the data. In this work, we have focused on partitioning the data set with randomization playing a dominant role. An alternative route is to allow overlapping splits of the data, so that each analyst receives a more than $1/k$ fraction of the data. We expect overlapping splits to be of most use when data sets are small or when they contain large numbers of potential predictors. Overlapping splits also allow us to benefit from the modelling efforts of a larger set of analysts.

A second issue is the development of prototypical problems so that a precise methodology can be specified depending on the goal(s) of the analysis and the type of data collected. Investigation of these problems will give us more guidance on how to split the data and on what restrictions to place on the Bayesian summaries.

A third issue is application of the methodology with non-Bayesian components. The benefits of averaging non-stable or different models applies more broadly than in the Bayesian setting. Noting differences between the models built by CART and by the information criteria, one could average them as well. However, without a Bayesian summary and with incomplete likelihoods, model synthesis becomes somewhat more ad-hoc. Natural routes to pursue include the prequential approach (e.g., Dawid and Vovk, 1999) and predictive model selection (e.g., Laud and Ibrahim, 1995).

Appendix A: Proofs of the Theorems

Proof. Proof of Theorem 2.1.

1) We want to show that $P(\min(X_{11}^2, X_{21}^2, \dots, X_{k1}^2) \leq t) \geq P((\frac{1}{k} \sum_{i=1}^k X_{i1})^2 \leq t)$, for any $t \geq 0$, with strict inequality for some t . Without loss of generality, assume $\Sigma_{11} = 1$. Thus

$$P(\min(X_{11}^2, X_{21}^2, \dots, X_{k1}^2) \leq t) = 1 - 2^k(1 - \Phi(\sqrt{t}))^k, P((\frac{1}{k} \sum_{i=1}^k X_{i1})^2 \leq t) = 2\Phi(\sqrt{kt}) - 1.$$

We want to show that $G(x) = 2 - 2^k(1 - \Phi(x))^k - 2\Phi(\sqrt{k}x) \geq 0$ for all $x > 0$. Let $T(x) = \frac{1}{2} \frac{\partial G(x)}{\partial x}$, then

$$\begin{aligned} T(x) &= 2^{k-1}k(1 - \Phi(x))^{k-1}\phi(x) - \sqrt{k}\phi(\sqrt{k}x) \\ &= \phi(x)e^{-\frac{k-1}{2}x^2} \{2^{k-1}k(2\pi)^{-\frac{k-1}{2}} [R(x)]^{k-1} - \sqrt{k}\}, \end{aligned}$$

where $R(x) = \frac{1-\Phi(x)}{\phi(x)}$. From Gordon (1941), we have that $R(x)$ is non-negative and, for $x > 0$, that $R(x)$ is monotone decreasing and tends to 0. Thus $\{2^{k-1}k(2\pi)^{-\frac{k-1}{2}} [R(x)]^{k-1} - \sqrt{k}\}$ is monotone decreasing in x for $x > 0$. Since $\phi(x)e^{-\frac{k-1}{2}x^2} \geq 0$ for any x , the sign of $T(x)$ is decided by $\{2^{k-1}k(2\pi)^{-\frac{k-1}{2}} [R(x)]^{k-1} - \sqrt{k}\}$, which is positive until some $x = x_0$ and then becomes negative. Thus $G(x)$ is increasing for $0 \leq x < x_0$ and decreasing for $x > x_0$. Thus $G(x)$ is minimized at $x = 0$ or at $x = +\infty$. But $G(0) = G(+\infty) = 0$. Hence $G(x) \geq 0$ for any $x \geq 0$, with strict inequality when $x \neq 0$ or ∞ .

2) We first consider the case where $\Sigma_{11} = \Sigma_{22}$ and $\Sigma_{12} = 0$, i.e. the first and second elements of X are independent and have equal variance. Without loss of generality, assume $\Sigma_{11} = \Sigma_{22} = 1$. Then $Y_2 \sim \text{Exp}(k/2)$. $X_{i1}^2 + X_{i2}^2 \sim \text{Exp}(1/2)$, and so $Y_1 = \min(X_{11}^2 + X_{12}^2, \dots, X_{k1}^2 + X_{k2}^2) \sim \text{Exp}(\sum_{i=1}^k \frac{1}{2}) = \text{Exp}(k/2)$. Hence Y_1 and Y_2 have the same distribution.

Suppose that the minimum defining Y_1 is realized at $X_{j1}^2 + X_{j2}^2$ and let $\mathbf{Z}_1 = (X_{j1}, X_{j2})^T$. Also, let $\mathbf{Z}_2 = (1/k) \sum_{l=1}^k (X_{l1}, X_{l2})^T$. The distributions of \mathbf{Z}_1 and \mathbf{Z}_2 are spherically symmetric about the origin. Thus \mathbf{Z}_1 and \mathbf{Z}_2 have the same joint distribution. Next, consider the case of $\Sigma_{11} > \Sigma_{22} > 0$ and $\Sigma_{12} = 0$. This distribution arises as a transformation of the equal variance normal distribution. Define $X_{ij}^* = \sqrt{\Sigma_{jj}} X_{ij}$ for $i = 1, \dots, k$ and $j = 1, \dots, d$. Applying this transformation to \mathbf{Z}_1 and \mathbf{Z}_2 from the equal variance case, we obtain \mathbf{Z}_1^* and \mathbf{Z}_2^* . By definition, $Y_2^* = \|\mathbf{Z}_2^*\|^2$. However, $Y_1^* \leq \|\mathbf{Z}_1^*\|^2$, with equality holding if \mathbf{Z}_1^* is still the closest of the \mathbf{X}_j^* (the transformed \mathbf{X}_j) to the origin after transformation. This occurs with probability less than 1, and so Y_1^* is stochastically smaller than Y_2^* . The normal

distribution with non-zero covariance can be obtained by rotating a zero-covariance normal.

3) Without loss of generality, assume $\Sigma_{11} = 1$. An absolutely continuous random variable X , with cdf $F(t)$ and pdf $f(t)$, has hazard rate $r_x(t) = \frac{f(t)}{1-F(t)}$. For two absolutely continuous variables X and Y , $r_x(t) \geq r_y(t)$ implies that X is stochastically smaller than Y .

Let $F_m(t)$, $f_m(t)$ and $r_m(t)$ denote the cdf, pdf and hazard rate for the chi-square distribution with m degrees of freedom. We have the cdf and pdf for Y_1 and Y_2 as follows:

$$F_{Y_1}(t) = 1 - (1 - F_m(t))^k \text{ and } f_{Y_1}(t) = F'_{Y_1}(t) = k(1 - F_m(t))^{(k-1)} f_m(t)$$

$$F_{Y_2}(t) = F_m(kt) \text{ and } f_{Y_2}(t) = F'_{Y_2}(t) = k f_m(kt).$$

To show that Y_2 is stochastically smaller than Y_1 , it suffices to show that $r_{Y_2}(t) \geq r_{Y_1}(t)$ or $\frac{1}{r_{Y_2}(t)} \leq \frac{1}{r_{Y_1}(t)}$. Since $k > 1$, for any $t > 0$, $kt > t$. In addition,

$$\begin{aligned} \frac{1}{r_{Y_1}(t)} &= \frac{(1 - F_m(t))^k}{k(1 - F_m(t))^{k-1} f_m(t)} = \frac{1}{k r_m(t)}, \\ \frac{1}{r_{Y_2}(t)} &= \frac{1 - F_m(kt)}{k f_m(kt)} = \frac{1}{k r_m(kt)}. \end{aligned}$$

Thus it is sufficient to show that for $x > 0$, $g(x) = \frac{1}{r_m(x)} = e^{\frac{x}{2}} x^{1-\frac{m}{2}} \int_x^\infty t^{\frac{m}{2}-1} e^{-\frac{t}{2}} dt$ is decreasing. Barlow and Proschan (1981) prove that this property holds when $m > 2$.

Proof. Proof of Theorem 2.2.

1) Without loss of generality, assume $\Sigma_{11} = 1$. We want to show that $P(Y_{1/2} \leq x) \geq P(Y_\alpha \leq x)$, for any $x \geq 0$ with strict inequality for some x . First, we consider only $n\alpha$ and $n/2$, not restricting them to be integers. Since

$$P(Y_\alpha \leq x) = 1 - 4(1 - \Phi(\sqrt{n\alpha x}))(1 - \Phi(\sqrt{n(1-\alpha)x})).$$

We want to show that

$$(1 - \Phi(\sqrt{\frac{nx}{2}}))^2 \leq (1 - \Phi(\sqrt{n\alpha x}))(1 - \Phi(\sqrt{n(1-\alpha)x})),$$

or that $T(\alpha) = (1 - \Phi(\sqrt{n\alpha x}))(1 - \Phi(\sqrt{n(1-\alpha)x}))$ is minimized at $\alpha = \frac{1}{2}$ for any $x \geq 0$ and $n > 0$. With $R(x)$ defined as in the proof of Theorem 2.1,

$$\begin{aligned} \frac{dT(\alpha)}{d\alpha} &= [1 - \Phi(\sqrt{n\alpha x})]\phi(\sqrt{n(1-\alpha)x})\frac{nx}{2\sqrt{n(1-\alpha)x}} - [1 - \Phi(\sqrt{n(1-\alpha)x})]\phi(\sqrt{n\alpha x})\frac{nx}{2\sqrt{n\alpha x}} \\ &= \frac{nx\phi(\sqrt{n(1-\alpha)x})\phi(\sqrt{n\alpha x})}{2\sqrt{n(1-\alpha)x}\sqrt{n\alpha x}} \left[\frac{\sqrt{n\alpha x}(1 - \Phi(\sqrt{n\alpha x}))}{\phi(\sqrt{n\alpha x})} - \frac{\sqrt{n(1-\alpha)x}(1 - \Phi(\sqrt{n(1-\alpha)x}))}{\phi(\sqrt{n(1-\alpha)x})} \right] \\ &= \frac{\phi(\sqrt{n(1-\alpha)x})\phi(\sqrt{n\alpha x})}{2\sqrt{\alpha(1-\alpha)}} [\sqrt{n\alpha x}R(\sqrt{n\alpha x}) - \sqrt{n(1-\alpha)x}R(\sqrt{n(1-\alpha)x})]. \end{aligned}$$

$$\frac{\phi(\sqrt{n(1-\alpha)x})\phi(\sqrt{n\alpha x})}{2\sqrt{\alpha(1-\alpha)}} \geq 0 \text{ for any } n, \alpha, x \geq 0. \text{ Let } B = \sqrt{n\alpha x}R(\sqrt{n\alpha x}) - \sqrt{n(1-\alpha)x}R(\sqrt{n(1-\alpha)x}).$$

From Gordon (1941), we have $\frac{dR(x)}{dx} = xR(x) - 1 < 0$ and $\frac{d^2R(x)}{dx^2} > 0$ for any $x \geq 0$.

Thus we have $\frac{dxR(x)}{dx} > 0$ for any $x \geq 0$. That is $xR(x)$ is increasing in x . When $\alpha \leq \frac{1}{2}$, $\sqrt{n\alpha x} \leq \sqrt{n(1-\alpha)x}$, so $B \leq 0$, and $\frac{dT(\alpha)}{d\alpha} \leq 0$. When $\alpha > \frac{1}{2}$, $\sqrt{n\alpha x} > \sqrt{n(1-\alpha)x}$, so $B > 0$, and $\frac{dT(\alpha)}{d\alpha} > 0$. So when $\alpha \leq \frac{1}{2}$, $T(\alpha)$ is increasing and when $\alpha > \frac{1}{2}$, $T(\alpha)$ is decreasing. Thus $T(\alpha)$ attains its minimum at $\alpha = \frac{1}{2}$.

If $n/2$ is an integer, Theorem 2.2 is proved. Otherwise, for any α for which $\lfloor n\alpha \rfloor \neq \lfloor \frac{n}{2} \rfloor$, and $\lfloor n(1-\alpha) \rfloor \neq \lfloor \frac{n}{2} \rfloor$ either $n/2 > \lfloor n/2 \rfloor > \lfloor n\alpha \rfloor$ or $n/2 < \lfloor n/2 \rfloor + 1 < \lfloor n\alpha \rfloor$.

$$\begin{aligned} P(Y_{1/2} \leq x) &= 1 - 4(1 - \Phi(\sqrt{\lfloor \frac{n}{2} \rfloor x}))(1 - \Phi(\sqrt{(n - \lfloor \frac{n}{2} \rfloor)x})) \\ &= 1 - 4(1 - \Phi(\sqrt{\lfloor \frac{n}{2} \rfloor x}))(1 - \Phi(\sqrt{(1 + \lfloor \frac{n}{2} \rfloor)x})). \end{aligned}$$

It is easy to see that $T(\alpha)$ is symmetric around $\frac{1}{2}$. Thus if $n/2 > \lfloor n/2 \rfloor > \lfloor n\alpha \rfloor$, $\alpha < 1/2$, $T(\alpha)$ is decreasing, and so $(1 - \Phi(\sqrt{\lfloor \frac{n}{2} \rfloor x}))(1 - \Phi(\sqrt{(1 + \lfloor \frac{n}{2} \rfloor)x})) \leq (1 - \Phi(\sqrt{\lfloor n\alpha \rfloor x}))(1 - \Phi(\sqrt{(n - \lfloor n\alpha \rfloor)x}))$, with strict inequality when $x \neq 0$ or ∞ . If $n/2 < \lfloor n/2 \rfloor + 1 < \lfloor n\alpha \rfloor$, then $\alpha > 1/2$, so $T(\alpha)$ is increasing, and we have $(1 - \Phi(\sqrt{\lfloor \frac{n}{2} \rfloor x}))(1 - \Phi(\sqrt{(1 + \lfloor \frac{n}{2} \rfloor)x})) \leq$

$(1 - \Phi(\sqrt{\lfloor n\alpha \rfloor x}))(1 - \Phi(\sqrt{(n - \lfloor n\alpha \rfloor)x}))$, with strict inequality when $x \neq 0$ or ∞ .

2) Without loss of generality, assume $\Sigma_{11} = 1$. Then $Y_\alpha \sim \text{Exp}(n/2)$ for all α .

3) Without loss of generality, assume $\Sigma_{11} = 1$. Let W denote the chi-square random variable with m df. It is easy to show that $P(Y_\alpha \leq \epsilon) = 1 - P(W \geq n\alpha\epsilon)P(W \geq n(1 - \alpha)\epsilon)$. We want to show that $p(\alpha) = P(W \geq n\alpha\epsilon)P(W \geq n(1 - \alpha)\epsilon)$ is increasing in α for $\alpha \in [0, 0.5)$. We rely on $f_m(x)$ and $g(x)$ as defined in the proof of Theorem 2.1. The derivative $\frac{dp(\alpha)}{d\alpha} = n\epsilon f_m(n\alpha\epsilon)f_m(n(1 - \alpha)\epsilon)[g(n\alpha\epsilon) - g(n(1 - \alpha)\epsilon)]$. From the proof of Theorem 2.1, we have that $g(x)$ is decreasing in x , and hence that $g(n\alpha\epsilon) - g(n(1 - \alpha)\epsilon) > 0$. Thus $\frac{dp(\alpha)}{d\alpha} > 0$ and $p(\alpha)$ is increasing in α for $\alpha \in [0, 0.5)$.

Proof. Proof of Theorem 2.3.

1) Take $a > b$. We want to show that $P(Y_a \leq x) \geq P(Y_b \leq x)$ for any $x \geq 0$, with strict inequality for some x . Without loss of generality, assume that $\Sigma_{11} = 1$. Then $P(Y_a \leq x) = 1 - 2^a(1 - \Phi(\sqrt{\frac{nx}{a}}))^a$. We want to show that $G(a) = \log[2^a(1 - \Phi(\sqrt{\frac{nx}{a}}))^a]$ is a non-increasing function in $a > 0$ for any fixed $x \geq 0$. Noting that $\frac{dG(a)}{da} = \log 2 + \log(1 - \Phi(\sqrt{\frac{nx}{a}})) + \frac{\sqrt{\frac{nx}{a}}\phi(\sqrt{\frac{nx}{a}})}{2(1 - \Phi(\sqrt{\frac{nx}{a}}))}$, let $y = \sqrt{\frac{nx}{a}}$. We want to show that $H(y) = \log 2 + \log(1 - \Phi(y)) + \frac{y\phi(y)}{2(1 - \Phi(y))} \leq 0$ for any $y \geq 0$.

$$\begin{aligned} \frac{d \log H(y)}{dy} &= -\frac{\phi(y)}{1 - \phi(y)} + \frac{d \frac{y}{2R(y)}}{dy} \\ &= -\frac{1}{R(y)} + \frac{2R(y) - 2y(yR(y) - 1)}{4R^2(y)} \\ &= -\frac{1 + y^2}{2R(y)} + \frac{y}{2R(y)} \cdot \frac{1}{R(y)} \\ &\leq -\frac{1 + y^2}{2R(y)} + \frac{y}{2R(y)} \cdot \frac{y^2 + 1}{y} = 0. \end{aligned}$$

The last inequality follows from Gordon (1941), that $\frac{x^2+1}{x} \geq \frac{1}{R(x)} \geq x$ for any $x > 0$, with strict inequality for some x . So $H(y)$ is non-increasing when $y \geq 0$ and the maximum $H(y)$

is attained at $y = 0$, where $H(0) = 0$. Thus $H(y) \leq 0$ for all $y \geq 0$, and so $G(a)$ is nonincreasing in a for $a > 0$.

2) Without loss of generality, assume $\Sigma_{11} = 1$. Then $Y_\alpha \sim \text{Exp}(n/2)$ for all α .

3) Without loss of generality, assume $\Sigma_{11} = 1$. Then $P(Y_a \leq \epsilon) = 1 - P^a(W \geq n\epsilon/a)$, where W follows the chi-square distribution with m df. We want to show that $p(a) = \log(P^a(W \geq n\epsilon/a))$ is non-decreasing in a . Note that $\frac{dp(a)}{da} = \log(1 - F_m(\frac{n\epsilon}{a})) + \frac{f_m(\frac{n\epsilon}{a})n\epsilon/a}{1 - F_m(\frac{n\epsilon}{a})}$. Let $x = n\epsilon/a$. We want to show that $\log(1 - F_m(x)) + \frac{xf_m(x)}{1 - F_m(x)} \geq 0$ for all $x > 0$.

Note that $\frac{d[-\log(1 - F_m(x))]}{dx} = \frac{f_m(x)}{1 - F_m(x)}$, which is increasing in x by the proof of Theorem 2.1.

By Lagrange's theorem, for any $x > 0$, there exists $x_0 \in [0, x]$ such that

$$\frac{-\log(1 - F_m(x))}{x} = \frac{-\log(1 - F_m(x)) - [-\log(1 - F_m(0))]}{x - 0} = \frac{f_m(x_0)}{1 - F_m(x_0)} \leq \frac{f_m(x)}{1 - F_m(x)}$$

which implies that $\log(1 - F_m(x)) + \frac{xf_m(x)}{1 - F_m(x)} \geq 0$.

References

- Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716-723.
- Barlow R.E. and Proschan F. (1981), *Statistical Theory of Reliability and Life Testing*, TO BEGIN WITH, Silver Spring, Maryland.
- Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., Springer-Verlag, New York.
- Berk, R.H. (1966). "Limiting Behavior of Posterior Distributions when the Model is Incorrect," *The Annals of Mathematical Statistics*, 37, 51-58.
- Breiman L. (1996), "Bagging Predictors," *Machine Learning*, 26, 123-140.
- Breiman L. (2001), "Statistical Modeling: The Two Cultures," *Statistical Science*, 16, 199-215, (Disc: p216-231).

- Breiman L., Friedman J.H., Olshen R.A., and Stone. C.J. (1984), *Classification and Regression Trees*, Wadsworth, Belmont, Ca.
- Chen, M., Shao Q. and Ibrahim J.G. (2000), *Monte Carlo Methods in Bayesian Computation*, Springer-Verlag, New York.
- Chipman, H.A., George, E.I., and McCulloch, R.E. (1998), “Bayesian CART model search,” *Journal of the American Statistical Association*, 93, 935-948, (C/R: p948-960).
- Chipman H. A., George E. I., and McCulloch R. E. (2005), “BART: Bayesian Additive Regression Tree”, <http://gsbwww.uchicago.edu/fac/robert.mcculloch/research/code/BART-7-05.pdf>.
- Dawid, A.P. and Vovk V.G. (1999), “Prequential probability : principles and properties,” *Bernoulli*, 5, 125-162.
- Draper, D. (1995), “Assessment and propagation of model uncertainty,” *Journal of the Royal Statistical Society, Series B*, 57, 45–70, (Disc: p71–97).
- Draper, D.C., Hodges, J.S., Leamer, E.E., Morris, C.N., and Rubin, D.B. (1987), “A Research Agenda for Assessment and Propagation of Model Uncertainty,” *Report N-2683-RC*, RAND Corporation.
- Efron B., Hastie T., Johnstone I. and Tibshirani R. (2004), “Least Angle Regression (with discussion),” *Annals of Statistics*, 32, 2, 407-499.
- Fligner M.A. and MacEachern S.N. (2003), “Ranked Set Sampling: Models and Distribution Free Two Sample Methods under Imperfect Ranking,” *to appear in the Journal of the American Statistical Association*.
- George, E.I. and McCulloch, R.E. (1993), “Variable Selection Via Gibbs Sampling,” *Journal of the American Statistical Association*, 88, 881-889.
- Gordon R.D. (1941), “Values of Mills’ Ratio of Area to Bounding Ordinate and of the Normal Probability Integral for Large Values of the Augment,” *The Annals of Mathematical*

- Statistics*, 12, 3, 109-122.
- Gu C. (2002), *Smoothing Spline ANOVA Models*, Springer-Verlag.
- Hand, D.J. (2006), "Classifier Technology and the Illusion of Progress," *Statistical Science*, 21, 1-14, (Disc: p15-34).
- Hastie T., Tibshirani R., and Friedman J. (2001), *The Elements of Statistical Learning*, Springer-Verlag, New York.
- Kass R.E. and Raftery, A.E. (1995), "Bayes factors," *Journal of the American Statistical Association*, 90, 773-795.
- Laud, P.W. and Ibrahim, J.G. (1995), "Predictive model selection," *Journal of the Royal Statistical Society, Series B*, 57, 247-262.
- MacEachern S. N., S. Guha and M. Peruggia (2003), Discussion of "A theory of statistical models for Monte Carlo integration" by Kong, McCullagh, Nicolae, Tan and Meng, *Journal of the Royal Statistical Society, Ser. B*, 65, 612.
- MacEachern, S. N., and Peruggia, M. (2000), "Subsampling the Random Scan Gibbs Sampler: Variance Reduction," *Statistics and Probability Letters*, 47, 91-98.
- Raftery A.E., Madigan D. and Hoeting J.A. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179-191.
- Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461-4.
- Thomas A., Best N., Lunn D., Arnold R. and Spiegelhalter D. (2004), *GeoBUGS User Manual*.
- Tibshirani R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, 58(1), 267-288.
- Weisberg, S. (1985), *Applied Linear Regression*, John Wiley & Sons, New York.
- Yu, Q. (2006), "Bayesian Synthesis." Unpublished Ph.D. Dissertation, The Ohio State

University.