

Combining Information from Related Regressions

FRANCESCA DOMINICI, GIOVANNI PARMIGIANI,
KENNETH H. RECKHOW AND ROBERT L. WOLPERT

May 1, 1997

Francesca Dominici is Visiting Assistant Professor, Department of Biostatistics, Johns Hopkins University. Giovanni Parmigiani is Assistant Professor, Institute of Statistics and Decision Sciences and Center for Health Policy Research and Education, Duke University. Robert Wolpert is Associate Professor, Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251. Kenneth Reckhow is Associate Professor, Nicholas School of the Environment, and Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251. This work was completed while Francesca Dominici was a visiting scholar at the Institute of Statistics and Decision Sciences. Research supported by NSF under grants DMS 9403818 and DMS-9305699.

The data used in this article are available on the web site <http://www.isds.duke.edu/~gp>.

Abstract

We propose and illustrate an approach for combining information from several regression studies, each considering only a subset of the variables of interest. Our approach uses a combination of Bayesian hierarchical modeling and data augmentation. Hierarchical models are a flexible tool for modeling study-to-study as well as within-study variability. Data augmentation methods address fully the uncertainty resulting from missing data and provide venues for combining information in a way that preserves the meaning of the regression coefficients across studies. We discuss in detail a normal-normal model, we suggest a simple and efficient numerical implementation based on a block Gibbs sampler, and we provide explicit full conditional distributions for an arbitrary pattern of variables missing by study.

We discuss an application of our model to investigating the level of chlorophyll-*a* in water quality management. Chlorophyll-*a* is one of the most important indicators of lake water quality. Scientists have developed a number and variety of forecasting models relating chlorophyll-*a* to nutrients such as phosphorus and nitrogen. These models often have to rely on sparse information from multiple sources – in this case lakes. We study the relationship among chlorophyll-*a* and phosphorus in twelve north temperate lakes by using data from the literature. An important covariate is nitrogen, which is reported only in some of the studies.

1 Introduction

1.1 Background and Motivation

If there are several studies that address the same research question, one might be interested in combining the information from the individual studies in order to draw overall conclusions about the research question of interest. The combining of the individual studies in order to learn about the whole is referred to in the literature as *meta-analysis*. In this paper we focus on meta-analysis of regression studies. In particular we discuss how to combine several multivariate regression data sets, each recording overlapping, but possibly different, sets of variables. This is a common situation: frequently, an initial study will identify a potentially interesting relationship between variables. New studies are then likely to follow, with more comprehensive designs and more variables, perhaps in an attempt to clarify potential confounding effects or biases in the initial study. Interest in similar questions from other agencies, technological progress in measuring potential explanatory variables, and emergence of new and interesting explanatory variables are all likely to lead to more studies, with yet different sets of variables. Often, studies have multiple endpoints or use different proxies for responses of interest. In practice, multi-study regression analyses carried out to support important policy decisions will very often require combining studies with different variables.

The goal of this paper is to provide a framework for handling some of the most urgent modeling problems arising in the situation just described. Examples are

i) combining several studies with a common response variable and overlapping, but different covariates;

ii) combining studies with the same covariates but different endpoints (responses), with the aid of one or more further studies investigating the dependence between the endpoints.

iii) combining multivariate analyses with differing sets of variables.

Our proposal is based on a combination of hierarchical modeling and data augmentation techniques. In meta-analysis, an important and well recognized concern is modeling vari-

ation from study to study, which can arise from differences in the studies subpopulations, data collection protocols and so forth. Hierarchical models are emerging as a flexible and practical modeling strategy, and are widely used in meta-analysis (DuMouchel and Harris, 1983, DuMouchel, 1990, Berry, 1990, Eddy, Hasselblad and Shachter, 1991, Gatsonis 1993) and other information synthesis problems (Morris and Normand, 1992, Wolpert and Warren-Hicks, 1992, Stangl, 1995). Studies are thought of as belonging to a population of studies addressing the same research question. Hierarchical models for regression problems are well understood when all studies report the same covariates (Lindley and Smith, 1972, DuMouchel, 1994). This suggests setting up a model based on the full set of predictors and responses recorded in at least one of the studies, and treating the variables that are not reported as missing. Information on missing variables in each study is provided by recorded variables in that study as well as the dependence structure inferred from other studies.

Advantages of this approach are both methodological and computational. Methodologically, the combination of regression models is carried out in a fashion that preserves the interpretation of the coefficient across different studies. Computationally, data augmentation can be handled conveniently using Markov chain Monte Carlo (MCMC) techniques (Tanner and Wong, 1987, Gelfand and Smith 1990, Tanner 1993, Gilks, Richardson and Spiegelhalter 1996). The main output of the analysis are posterior distributions of parameters and future outcomes of interest. These can be easily marginalized over the missing information, uncertainty about which is then addressed in full.

The outline of the paper is as follows. We begin with an elementary example motivating the need for careful treatment of the missing variables. In section 2 we introduce the model. In section 3 we discuss Markov chain Monte Carlo methods for deriving the marginal posterior and predictive distributions of interest, and we provide the full conditional distributions for implementing a block Gibbs sampler handling general missing variables patterns. Finally in section 4 we consider an application to investigating the relation between chlorophyll-*a*, phosphorus and nitrogen in twelve north temperate lakes by using data from several studies

in the literature.

1.2 A Tutorial Example

Imagine that studies of the effect of phosphorus on the concentration of chlorophyll-*a* in lakes are available. Studies may adjust for known covariates: the first study corrects for the effect of nitrogen, the second corrects for the effect of lake depth and the third corrects for both these effects. We use the notation:

Y = Chlorophyll-*a* concentration,

X_1 = Phosphorus concentration,

X_2 = Nitrogen concentration,

X_3 = Lake depth.

Our goal is to combine information from these three studies in order to find a pooled estimate of the regression coefficient β for the predictor variable X_1 .

Suppose that the four variables $Z = [Y, X_1, X_2, X_3]$ are, after suitable transformations, well approximated by a joint Gaussian distribution with $\mu = 0$ and covariance matrix:

$$\Sigma = \begin{bmatrix} 1 & -1/4 & 1/2 & -1/2 \\ \cdots & 1 & -1/2 & 1/2 \\ \cdots & \cdots & 1 & -1/4 \\ \cdots & \cdots & \cdots & 1 \end{bmatrix}$$

The conditional distribution of Y given any set of the X_i 's are easily computed linear functions of the X_i 's. Thus the first study that included phosphorus and nitrogen, should find approximately:

$$E[Y^1 \mid X_1^1, X_2^1] \simeq X_1^1 \times (0) + X_2^1 \times \left(\frac{1}{2}\right),$$

suggesting that nitrogen is affecting chlorophyll-*a*, but that phosphorus is not. The second study, that used only phosphorus and lake depth, should find approximately:

$$E[Y^2 \mid X_1^2, X_3^2] \simeq X_1^2 \times (0) + X_3^2 \times \left(-\frac{1}{2}\right),$$

suggesting that lake depth is an important predictor but that phosphorus is not. The third study, that included both concomitant variables, would find approximately:

$$E[Y^3 | X_1^3, X_2^3, X_3^3] \simeq X_1^3 \times \left(\frac{1}{4}\right) + X_2^3 \times \left(\frac{1}{2}\right) + X_3^3 \times \left(-\frac{1}{2}\right),$$

suggesting that phosphorus is affecting chlorophyll-*a*—though the first two studies agreed in suggesting that $\beta \simeq 0$.

When each study carries complete information, and the combination is based on a fixed effects model with equal variance, the combined maximum likelihood estimate MLE of the coefficients is a weighted average of the MLEs of the regression coefficients within each study, with the sample sizes as weights. But what happens in the presence of *incomplete* information and unequal variances? This simple combination method can be seriously misleading for the following reasons:

1. If each study includes a different set of covariates, then this makes almost it impossible to maintain that the regression coefficients have the same meaning and that they are constant over studies.
2. For each study, if the missing covariates are correlated with the observed covariates then the MLE obtained from the reduced data is no longer unbiased.
3. If all studies are considered, the individual studies' MLEs are not *sufficient statistics* for the coefficients in the full model.

All these objections can be overcome by reparameterizing the problem and expanding the multivariate regression model to include the uncertain joint distribution of the covariates' X 's too, rather than only the conditional distribution of Y 's given X 's. One approach for implementing this strategy in a hierarchical setting is discussed next.

2 Model

Consider S studies each including n^s observations on at most l response variables and at most r explanatory variables. We begin by defining the general structure of the model under complete information. The analysis in the presence of missing covariates or missing response variables is then conducted by conditioning on all observed variables and treating the unobserved variables as unknown parameters.

2.1 Complete Information

We indicate the response variables in study s by $Y^s = (Y_1^s, \dots, Y_l^s)'$, the explanatory variables by $X^s = (X_1^s, \dots, X_r^s)'$ and the complete set of $l+r$ variables by $Z^s = (Y_1^s, \dots, Y_l^s, X_1^s, \dots, X_r^s)'$. In each study, we model Z^s by a joint multivariate normal distribution with mean μ^s and covariance matrix Σ^s . Complete sample mean and covariance matrices will be denoted by \bar{Z}^s and V^s . One implication is that each of the l -dimensional vectors of responses Y^s in study s is modeled as a conditional normal distribution given X^s : partitioning the study-specific parameters as

$$\mu^s = \begin{bmatrix} \mu_y^s \\ \mu_x^s \end{bmatrix}; \quad \Sigma^s = \begin{bmatrix} \Sigma_{yy}^s & \Sigma_{yx}^s \\ \Sigma_{xy}^s & \Sigma_{xx}^s \end{bmatrix},$$

we have that

$$Y^s | X^s \sim N_l \left(Y^s | \mu_y^s + B^s(X^s - \mu_x^s), \Sigma_{yy}^s - \Sigma_{yx}^s \Sigma_{xx}^{s-1} \Sigma_{xy}^s \right)$$

for $s = 1, \dots, S$. Here $B^s = \Sigma_{yx}^s \Sigma_{xx}^{s-1}$ is the matrix of regression coefficients for study s .

In general it is restrictive to assume that study parameters are homogeneous across studies. However it is reasonable to assume that they are sufficiently similar to be thought of as belonging to some common distribution. In this article we model this by assuming that μ^s are exchangeable draws from a normally distributed population with mean μ^* and covariance matrix Γ^* , and that Σ^s are exchangeable draws from an inverse Wishart distribution with w degrees of freedom and mean Σ^* . Some of the parameters of the distribution describing the

population of studies (in our case μ^* , Γ^* , and Σ^*) will be unknown, and will be assigned a prior distribution.

In summary, our assumptions define the following hierarchical model:

$$\begin{aligned}
\text{Stage I: } \quad \bar{Z}^s \mid \mu^s, \Sigma^s &\sim N_{l+r} \left(\bar{Z}^s \mid \mu^s, \frac{1}{n^s} \Sigma^s \right) \\
V^s \mid \Sigma^s &\sim W_{l+r} (V^s \mid (n^s - 1), \Sigma^s) \\
\text{Stage II: } \quad \mu^s \mid \mu^*, \Gamma^* &\sim N_{l+r} (\mu^s \mid \mu^*, \Gamma^*) \\
\Sigma^s \mid \Sigma^* &\sim IW_{l+r} (\Sigma^s \mid w, \Sigma^*); \\
\text{Stage III: } \quad \mu^* &\sim N_{l+r} (\mu^* \mid m, M) \\
\Gamma^* &\sim IW_{l+r} (\Gamma^* \mid g, G) \\
\Sigma^* &\sim W_{l+r} (\Sigma^* \mid a, D).
\end{aligned}$$

Here studies are assumed to be independent conditional on μ^* , Γ^* and Σ^* . The quantities w, m, M, g, G, a, D are known hyperparameters. We use the notation $IW_{l+r}(\Gamma \mid g, G)$ to denote the inverse Wishart density proportional to

$$\frac{|G|^{(g+l+r-1)/2}}{|\Gamma|^{(g+2(l+r))/2}} \exp \left\{ -\frac{1}{2} \text{tr} (G \Gamma^{-1}) \right\},$$

where g will be referred to as the degrees of freedom and the $(l+r) \times (l+r)$ positive definite matrix G as the scale matrix. Similarly, we use the notation $W_{l+r}(\Sigma \mid a, D)$ to denote the Wishart distribution with density proportional to

$$\frac{|\Sigma|^{(a-(l+r)-1)/2}}{|D|^{a/2}} \exp \left\{ -\frac{1}{2} \text{tr} D^{-1} \Sigma \right\}.$$

Again, a will be referred to as the degrees of freedom and the $(l+r) \times (l+r)$ positive definite matrix D as the scale matrix.

Interest is both in the study specific (stage II) parameters and in the population (stage III) parameters. The stage II parameters μ^s and Σ^s represent the location and the correlation structure of the $l+r$ variables for the s -th study. Inference on stage II parameters based on a model like ours is often preferable to separate analyses of individual studies, because of the well known ‘‘borrowing of strength’’ resulting from the hierarchical structure.

Inference on stage III parameters represents a synthesis of the S studies and may have an intrinsic scientific importance. For example, in Section 4, we consider the effect of phosphorus and nitrogen on the concentration of chlorophyll- a in twelve lakes; each study corresponds to a lake. In this case, from the distributional assumptions at the third stage, we have $E(\mu^s) = \mu^*$, $E(\Sigma^s) = \frac{1}{w-2}\Sigma^*$. Therefore μ^* and Σ^* determine the location and the correlation structure of the three variables on average over all the lakes. In addition, given that the observations in each study are recorded at the same location over time we can interpret Σ^* also as a temporal correlation structure common to all lakes. Finally Γ^* measures the variability and the correlation of the location parameter μ^s respect to the overall mean μ^* and can be viewed as a spatial correlation between lakes.

In the presence of multiple data sets it is often of interest to estimate parameters in regression models. In particular we are interested in the posterior distribution of the regression coefficients for the lake s , denoted by $B^s = \Sigma_{yx}^s \Sigma_{xx}^{s-1}$, and of the regression coefficients synthesis over all the lakes, denoted by $B^* = \Sigma_{yx}^* \Sigma_{xx}^{*-1}$.

2.2 Missing Variables

Consider now the situation where some of the variables are missing. We assume here that the presence or absence of a variable in a study is unrelated to the value of both the missing and observed variables. In other words we assume that the variables are missing completely at random (Rubin 1976). For each study we rearrange the vector Z^s , so that it can be written as (W^s, U^s) , where W^s denotes the vector of the p^s variables that are present in study s , and U^s denotes the vector of the remaining $q^s = (l + r) - p^s$ variables that are missing in study s . Both W^s and U^s can include responses as well as explanatory variables. The following decomposition of the study specific parameters is useful:

$$\mu^s = \begin{bmatrix} \mu_w^s \\ \mu_u^s \end{bmatrix}; \quad \Sigma^s = \begin{bmatrix} \Sigma_{ww}^s & \Sigma_{wu}^s \\ \Sigma_{uw}^s & \Sigma_{uu}^s \end{bmatrix}.$$

The dimensions of these subvectors and submatrices depends on the number of variables recorded in the study s ; in particular, μ_w^s and μ_u^s are vectors of dimensions p^s and q^s , and Σ_{ww}^s , Σ_{wu}^s , Σ_{uw}^s and Σ_{uu}^s are matrices of dimension $p^s \times p^s$, $p^s \times q^s$, $q^s \times p^s$ and $q^s \times q^s$.

Because of the normality of assumption at the first stage, the conditional distribution of the unobserved variables given the observed variables and the study-specific parameters, within each study, is:

$$U^s \mid W^s, \mu^s, \Sigma^s \sim N_{q^s} \left(U^s \mid \mu_u^s + \Sigma_{uw}^s \Sigma_{ww}^{s-1} (W^s - \mu_w^s), \Sigma_{uu}^s - \Sigma_{uw}^s \Sigma_{ww}^{s-1} \Sigma_{wu}^s \right).$$

Using this framework, we can draw inferences on any of the unknowns using the posterior distribution

$$p(\mu^1, \dots, \mu^S, \Sigma^1, \dots, \Sigma^S, U^1, \dots, U^S, \mu^*, \Sigma^*, \Gamma^* \mid W^1, \dots, W^S) \quad (1)$$

Inference on the parameters of interest μ^* and Σ^* can be based on computing marginal distributions. Neither the posterior distribution (1) nor its marginal distributions are available in closed form in this case. However, practical algorithms for simulating from (1) can be based on Markov chain Monte Carlo schemes, discussed next.

2.3 Sampling

A practical choice for simulating from the joint posterior distribution (1) is to use a block Gibbs sampler (Gelfand and Smith 1990). This is based of partitioning the unknowns into groups and sampling each group in turn given all others. This requires the so-called full conditional distributions which are given explicitly next.

Define $\Psi^s = (\Sigma^s)^{-1}$, $\mu = \frac{1}{S} (\mu^1 + \dots + \mu^S)$, $\Psi = \Psi^1 + \dots + \Psi^S$, then:

$$\begin{aligned} \mu^* | \Gamma^*, \mu^1, \dots, \mu^S &\sim N_{l+r} \left(\mu^* | \left[S\Gamma^{*-1} + M^{-1} \right]^{-1} \left[\Gamma^{*-1} S\mu + M^{-1}m \right], \left[S\Gamma^{*-1} + M^{-1} \right]^{-1} \right) \\ \Gamma^* | \mu^1, \dots, \mu^S &\sim IW_{l+r} \left(\Gamma^* | g + S, \sum_{s=1}^S (\mu^s - \mu^*) (\mu^s - \mu^*)' + G \right) \\ \Sigma^* | \Psi &\sim W_{l+r} \left(\Sigma^* | a + Sw, [\Psi + D^{-1}]^{-1} \right) \\ \mu^s | \bar{Z}^s, \Sigma^s, \mu^*, \Gamma^* &\sim N_{l+r} \left(\mu^s | \left[n^s \Psi^s + \Gamma^{*-1} \right]^{-1} \left[n^s \Psi^s \bar{Z}^s + \Gamma^{*-1} \mu^* \right], \left[n^s \Psi^s + \Gamma^{*-1} \right]^{-1} \right) \\ \Sigma^s | V^s, \Sigma^* &\sim IW_{l+r} (\Sigma^s | w + n^s - 1, V^s + \Sigma^*) \\ U^s | W^s, \mu^s, \Sigma^s &\sim N_{q^s} (U^s | \mu_u^s + \Sigma_{uw}^s (\Sigma_{ww}^s)^{-1} (\mu_w^s - W^s), \Sigma_{uu}^s - \Sigma_{uw}^s (\Sigma_{ww}^s)^{-1} \Sigma_{wu}^s). \end{aligned}$$

The full distribution of U^s reflects all the uncertainty arising from the missing variables. In a practical implementation it is used to simulate each of the n^s missing observations. The derivation of these distributions is routine (see, e.g. Bernardo and Smith 1994) and is not discussed here.

3 Chlorophyll-Phosphorus relations in Lakes

In this section we apply the hierarchical model of Section 2 to investigating the relation between chlorophyll-*a*, phosphorus, and nitrogen in lakes. Chlorophyll-*a* is one of the most widely measured and predicted indicators of lake water quality. It serves as a measure of the density of algal cells and also reflects the “greenness” or clarity of the water in a lake. Higher concentrations of chlorophyll-*a* are associated with higher algal densities and poorer water quality (a condition called “eutrophication”).

The nutrients phosphorus and nitrogen stimulate the growth of algae, and hence are indicators, or predictors, of the potential for algal growth. As a consequence, scientists have developed a number and variety of simulation and forecasting models relating phosphorus and nitrogen to chlorophyll-*a* (see Reckhow and Chapra 1983). These models are extensively used by scientists and engineers to guide lake management.

Lake	1	2	3	4	5	6	7	8	9	10	11	12
C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TP	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TN	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓
n	16	8	4	7	4	4	4	10	6	6	3	2

Table 1: Summary of the variables recorded in each of the 12 lakes under study, and corresponding sample sizes. Each column represents a lake. A check mark indicates that the variable was recorded.

Our investigation is based on data on twelve of the north temperate lakes, reported by Smith and Shapiro (1981), who also give background and motivation.

Their analysis suggests that reductions in total phosphorus (TP) concentration in the lakes are generally accompanied by consistent declines in algal biomass, as measured by chlorophyll-*a* (C). However, the amount of such decline tends to vary from lake to lake. The data also suggest that chlorophyll-*a*'s response can be expected whether algal growth in a lake is phosphorus-limited or nitrogen-limited, although the magnitude of the response may differ. Also, not all studies report total nitrogen concentration (TN), and sample sizes within each lake are small.

The present hierarchical methodology provides an effective strategy to build a global model to assess the effect of phosphorus in a heterogeneous population of lakes, and to accomodate the fact that TN is not always reported. In particular, our assessment of the chlorophyll-phosphorus relations in individual lakes takes into account that:

- It is necessary to include in the analysis the effect of the nitrogen, even though some studies do not report nitrogen levels;
- It is of interest to investigate both the geographical and temporal dependencies between the variables and to model those separately, as temporal variation is more strongly related to human intervention;
- It can be important to provide a predictive distribution for the effect of phosphorus concentration reduction in a north temperate lakes not included in the sample.

We consider the following twelve related regression models:

$$\log(\text{C}) = \beta_0^s + \beta_1^s \log(\text{TP}) + \beta_2^s \log\left(\frac{\text{TN}}{\text{TP}}\right), \quad s = 1, \dots, 12. \quad (2)$$

The choice of the logarithmic scale is common in this application (Smith and Shapiro, 1981). Using $\log\left(\frac{\text{TN}}{\text{TP}}\right)$ to model the effect of nitrogen is intended to reduce correlation between the two explanatory variables, and to improve the stability of estimated coefficients. Mathematically, the model is equivalent to one including $\log(\text{TP})$ and $\log(\text{TN})$ as explanatory variables.

We used dispersed prior distributions to let the observed data drive the conclusions of the data analysis. Our prior distributions are all proper. The hyperparameters were chosen to reflect basic scientific knowledge of the allowable ranges of the observations. In practice, we specified a plausible range for each of the coefficients and chose prior densities that would vary at most by a factor of 2 within the range. The values used are given in the appendix. The implied prior distributions on the level III regression coefficients B^* are indicated by the dashed lines in Figure 4.

Using the Gibbs sampler of section 3 we obtained a sample from the joint posterior distribution of all unknown quantities

$$p(\mu^1, \dots, \mu^S, \Sigma^1, \dots, \Sigma^S, U^1, \dots, U^S, \mu^*, \Sigma^*, \Gamma^* | W^1, \dots, W^S) \quad (3)$$

Samples of the vectors B^s of regression coefficients in each of the twelve lakes and of the vector B^* of overall regression coefficients can be obtained simply by variable transformation on the sampled parameters. The transformations are defined in section 2.

Figures 1 and 2 summarize point inference about the regression coefficients. The thicker lines are the regression lines implied by the posterior mean of the stage III coefficients B^* , for a fixed value of the concomitant variable. There are also two shorter lines for each lake ranging from the smallest to the largest phosphorus level in the study. Each line extends from the smallest to the largest phosphorus level in a particular study. Solid lines are based on posterior means of lake-specific regression coefficients, and dashed lines are based on the

ordinary least squares estimator. Dashed lines are missing for the studies that did not report TN. The line thickness is proportional to the square root of the number of observations in the lake. From Figure 1, the coefficients of the chlorophyll-*a*-phosphorus relations (ie, the slope of the short lines) appear to be relatively stable across lakes, even though the ranges of the observations vary substantially. In Figure 2, coefficients appear more variable, indicating that the TN/TP ratio has a more variable effect across lakes, possibly a slight indication of a nonlinear relation. This is made more pronounced by the presence of one extreme observation with $\log\left(\frac{\text{TN}}{\text{TP}}\right) = 4.27$ in lake 6. The second largest value in lake 6 is 3.58, and there are only 4 observations, indicating that the evidence in favor of nonlinearity from lake 6 is less strong than the display suggests. The change in slope may also be indicative of a limiting nutrient mechanism (Kaiser et al. 1994). A nonlinear transformation of one of the covariates, capturing such a mechanism, could be accommodated in our framework. A further strategy for extending our framework is to model the coefficients as a function of lake-specific covariates as in Wong and Mason (1985).

Figure 3 displays the posterior distributions of the coefficients B^s for the twelve lakes. Lakes 2 and 3 have a more dispersed distribution as a result of the missing nitrogen measurement. In addition lake 3 has a limited number of observations. The distribution of lake 6 is also more dispersed, because of the discrepancy between the observations for that lake, suggesting a negative slope, and the overall tendency. For comparison, the marginal posterior distributions of the corresponding element of B^* is displayed at the far right. B^* represents the mean of the population of lake specific coefficients B^s . Figure 4 focuses on the marginal distributions of the elements of B^* and shows both the prior and posterior distributions. It emphasizes that the effect of the data is strong even on stage III parameters. Figure 5 shows the joint distribution of β_1^* and β_2^* , highlighting the correlation between the two estimated coefficients. Accounting for such correlation is especially important in computing the predictive distribution for measurements in lakes not included in the sample.

The spatial variability of measurements from lake-to-lake is captured by the matrix Γ^* .

Marginal distributions of the elements of Γ^* are shown in Figure 6. A large diagonal element signifies large lake to lake variability in the corresponding measurement, while a large off-diagonal element signifies a large spatial correlation between the two variables. Elements of Γ^* can themselves be correlated in our model. All pairwise joint posterior distributions of elements of Γ^* are shown in Figure 7.

Similarly, the matrix Σ^* represents the variability of measurements over time. Marginal distributions of the elements of Σ^* are shown in Figure 8. A large diagonal element signifies large variability over time in the corresponding measurement, on average over lakes. A large off-diagonal element signifies a large average temporal correlation between the two variables. All pairwise joint posterior distributions of elements of Σ^* are shown in Figure 9.

Finally, Figure 10 shows the predictive distribution of C given three levels of TP for fixed TN for a hypothetical 13-th lake, randomly selected from north temperate lakes. This distribution accounts fully for the uncertainty in parameter estimation and missing covariates, as well as for lake-specific noise in the 13-th lake.

4 Discussion

In this paper we considered the problem of combining information from several regression studies, each considering only a subset of the variables of interest. We approached the problem using Bayesian hierarchical models. These combine flexibility in modelling study-to-study as well as within-study variability with reliable computational algorithms for variance components and imputation of missing values. We provided full conditional distributions for the implementation of a Gibbs sampler, useful for arbitrary patterns of variables missing by study.

The situation of interest in this study, missing covariates in the context of multiple studies addressing a common issue, is encountered frequently in water quality research and assessment. With respect to chlorophyll-*a* prediction, it is not uncommon to have data on a set of lakes with data missing for one or more of the predictor variables for one or more of

the studies (the situation encountered here). Furthermore, scientists recognize that every lake has unique features but also has common features that can be exploited with the hierarchical analysis presented here. These conditions hold for other predictions of interest in lake studies, such as: prediction of nutrient concentrations as a function of watershed and hydrologic variables (Reckhow and Chapra 1983), prediction of fish population response to lake acidification (Wolpert and Warren-Hicks 1993), and assessment of lake trophic state (Reckhow and Chapra 1983).

We conclude by some potential extensions of this work. First, the general strategy of combining hierarchical models and data augmentation can be applied beyond normal distributions and linear models. While Gibbs samplers may not be always be available, other MCMC simulation algorithms can be implemented (Tierney, 1994). Similar considerations apply to more complex modeling of the relation between nutrients and chlorophyll-*a*, such as limiting nutrient models (Kaiser 1994). Next, our treatment of missing variables is based on the assumption that both the covariates and the response variables are random. Missing covariates are also common in designed experiments, such as clinical trials or toxicity studies for risk assessment. If the variables that are fixed by design are present in all studies our approach can be applied with small modifications. Otherwise, predicting a missing design variable based on the observed ones is in general inappropriate. Also, our results and algorithms are based on the assumption that no information is carried by the absence of a certain covariate in a study. This is an important assumption and could be problematic if predictors had been initially recorded and then eliminated from a study based on a preliminary exploratory analysis or variable selection. It is possible to extend our modeling strategies to handle such censored information by incorporating the variable selection mechanism in the imputation of the missing variables. Finally, we assumed throughout that the study's means and covariance matrices are available for analysis. Meta-analysis of regression studies requires different approaches when more limited information, such as significance test results, is reported.

Appendix

The values of the hyperparameters used in our model are as follows: $w = 3, g = 3, a = 2.5$,

$$G = \begin{bmatrix} 6 & 3.6 & .6 \\ 3.6 & 6 & 0 \\ .6 & 0 & 6 \end{bmatrix}.$$

This implies that the diagonal elements are $\text{Gamma}(3, 18)$.

$$m = (4.6, 5.5, -1.94).$$

$$M = \begin{bmatrix} 5.28 & 0 & 0 \\ 0 & 7.28 & 0 \\ 0 & 0 & 6.25 \end{bmatrix}.$$

This choice of m and M lead a range of $(-5, 5)$ for the mean of $\log(\frac{TN}{TP})$ with 75% of coverage.

$$D = \begin{bmatrix} 1.6 & .96 & .16 \\ .96 & 1.6 & 0 \\ .16 & 0 & 1.6 \end{bmatrix}.$$

This implies that the diagonal elements are $\text{IGamma}(2.5, 4)$.

References

- Bernardo, J.M. and Smith, A.F.M. (1994), *Bayesian Theory*, New York: Wiley.
- Berry, D.A. (1990), "A Bayesian approach to multicenter trials and meta-analysis," *ASA Proceedings of Biopharmaceutical Section*, 1–10.
- DuMouchel, W. and Harris J. E. (1983), "Bayes Methods for Combining the Results of Cancer Studies in Humans and Other Species," *Journal of American Statistical Association*, **78**, 293–308.
- DuMouchel, W. (1990), "Bayesian meta-analysis," in *Statistical Methodology in the Pharmaceutical Sciences*, ed. Berry, D.A., 509–529, New York.
- DuMouchel, W. (1994), "Hierarchical Bayes Linear Models for Meta-analysis," Technical Report, n. 27, NISS.
- Eddy, D.M., Hasselblad, V., and Shachter, R. (1991), *Meta-Analysis by the Confidence Profile Method*, New York: Academic Press.
- Gatsonis, C., Normand, S.L., Liu C., Morris C. (1993), "Geographic variation of procedure utilization: A Hierarchical Model Approach," *Medical Care*. 31 :YS54-YS59.
- Gelman, A. Carlin, J., Stern, H. and Rubin, D. (1995), *Bayesian Data Analysis*, London: Chapman and Hall.
- Gilks, W.R., Richardson, S., and Spiegelhalter, D. (1996), *Markov chain Monte Carlo in practice*, London: Chapman and Hall.
- Kaiser, M.S., Speckman, L. and Jones J.R. (1994), "Statistical Models for Limiting Nutrient Relations in Inland Waters " *Journal of the American Statistical Association*, **89**, 410–423.
- Lindley, D. V. and Smith, A. F. M. (1972), "Bayes Estimates for the Linear Model (with Discussion)," *Journal of the Royal Statistical Society, Series B, Methodological*, **34**, 1–41.
- Morris, C. and Normand, S.L. (1992), "Hierarchical models for combining information and for meta-analyseys (with discussion)." In *Bayesian Statistics 4*, ed. J.M. Bernardo, J.O.

- Berger, A.P. Dawid and Smith, A.F.M., 321–335.
- Reckhow, K.H., and Chapra, S.C. (1983), “Confirmation of Water Quality Models,” *Ecological Modelling*, **20**, 113–133.
- Rubin, D.B. (1976), “Inference and Missing data,” *Biometrika*, **63**, 581–592.
- Stangl, D., (1995) ”Prediction and Decision Making Using Bayesian Hierarchical Models,” *Statistics in Medicine*, **14(20)**, 2173–2190.
- Smith, V.H. and Shapiro J. (1981), “Chlorophyll-Phosphorus relations in Individual Lakes. Their Importance to Lake Restoration Strategies.” *Environmental Science & Technology*, **15**, 444-451.
- Tanner, M.A. (1993), *Tools for Statistical Inference*,” New York: Springer-Verlag.
- Tierney, L. (1994), “Markov Chains for exploring posterior distributions (with discussion),” *Annals of Statistics*, **22**, 1701–1762.
- Wolpert R.L. and Warren-Hicks W.J. (1992), ”Bayesian hierarchical logistic models for combining field and laboratory survival data (with discussion),” *In Bayesian Statistics 4*, ed. J.M. Bernardo, J.O. Berger, A.P. Dawid and Smith, A.F.M., 525–546.
- Wong G. Y. and Mason W.M (1985), ”The Hierarchical Logistic Regression Model for Multilevel Analysis,” *JASA*, **80**,513–524.

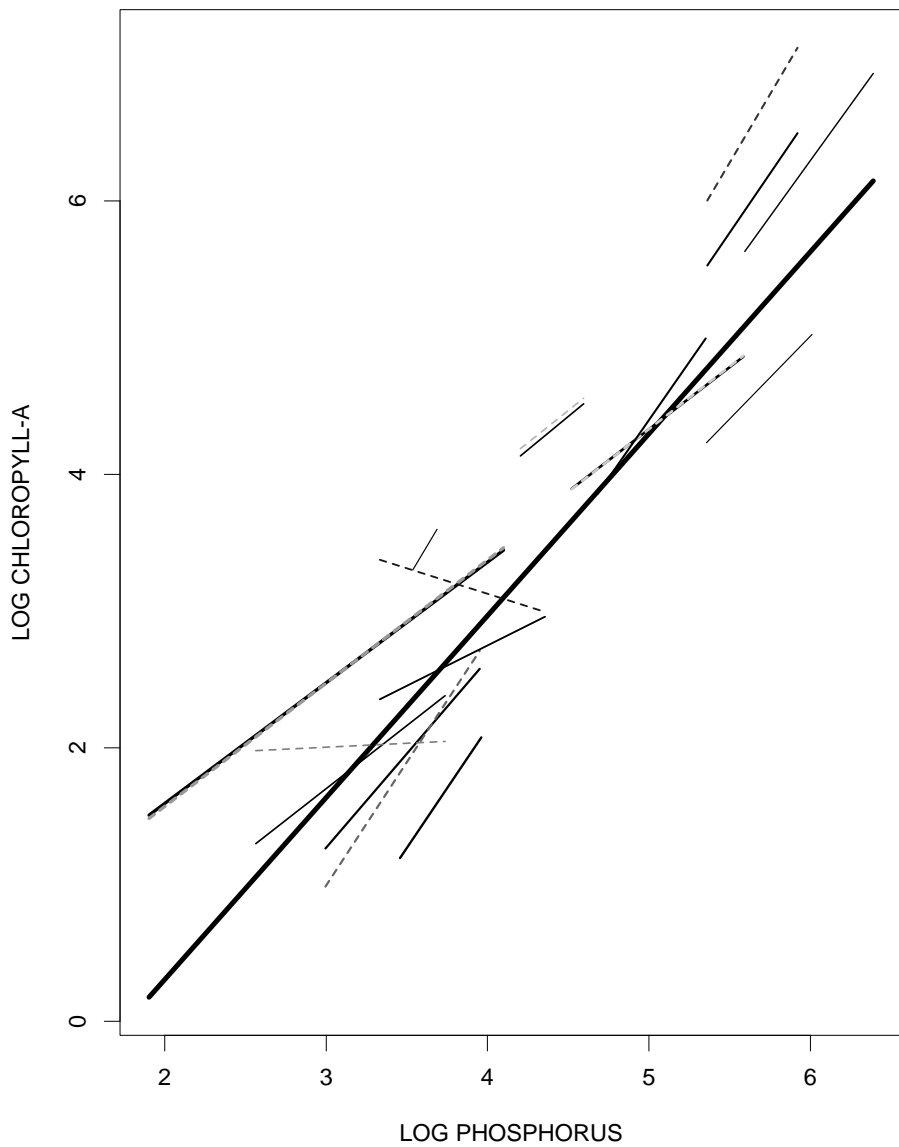


Figure 1: Summary of point inference about the regression coefficient of $\log(C)$ and $\log(TP)$. The thicker line is the regression line given by the posterior mean of the stage III coefficients B^* . The value of $\log TN/TP$ is fixed at its mean value 2.71. Smaller lines refer to the lake-specific regression coefficients. Smaller solid lines are based on posterior means and dashed lines are based on the ordinary least squares estimators. Line thickness is proportional to the square root of the number of observations in the lake.

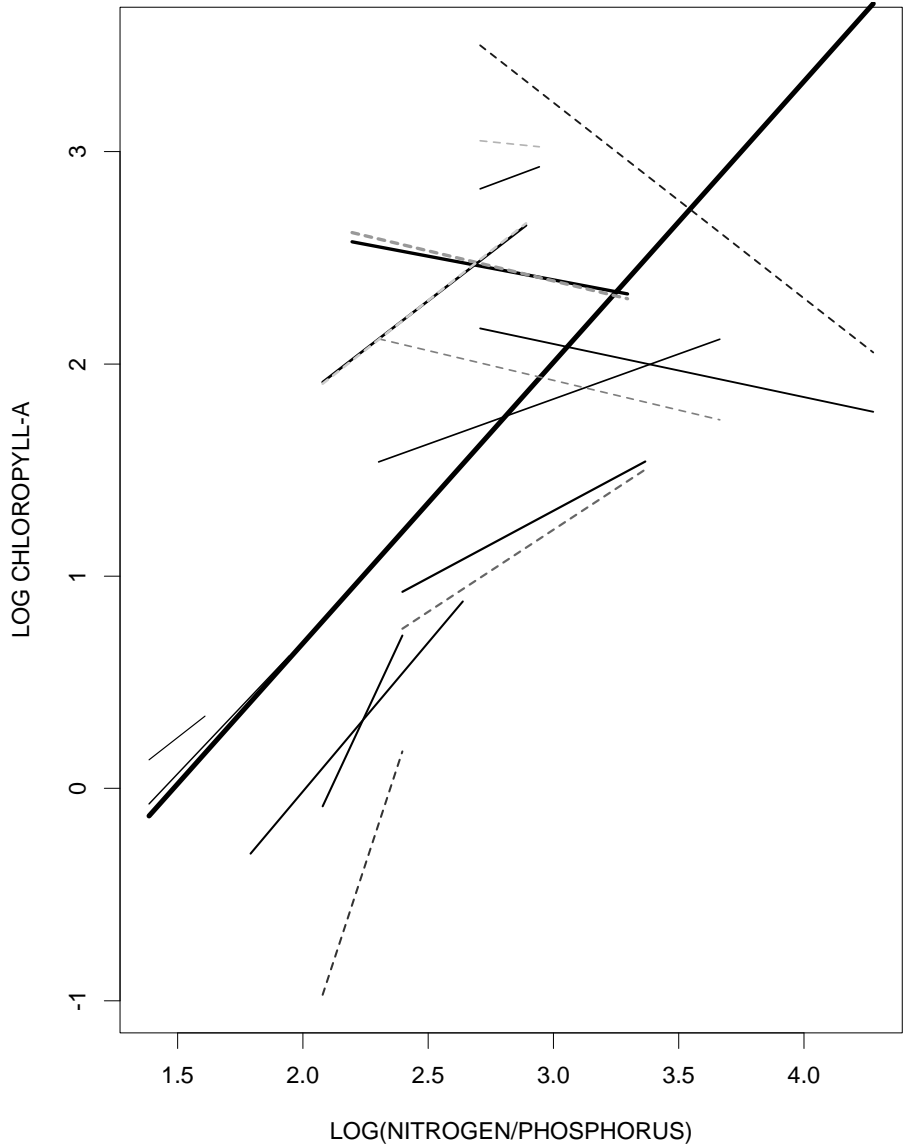


Figure 2: Point inference about the regression coefficient of $\log(C)$ and $\log TN/TP$. The thicker line is the regression line implied by the posterior mean of the stage III coefficients B^* . The value of $\log TP$ is fixed at its mean value 3. Smaller lines refer to the lake-specific regression coefficients. Smaller solid lines are based on posterior means and dashed lines are based on the ordinary least squares estimators. Line thickness is proportional to the square root of the number of observations in the lake.

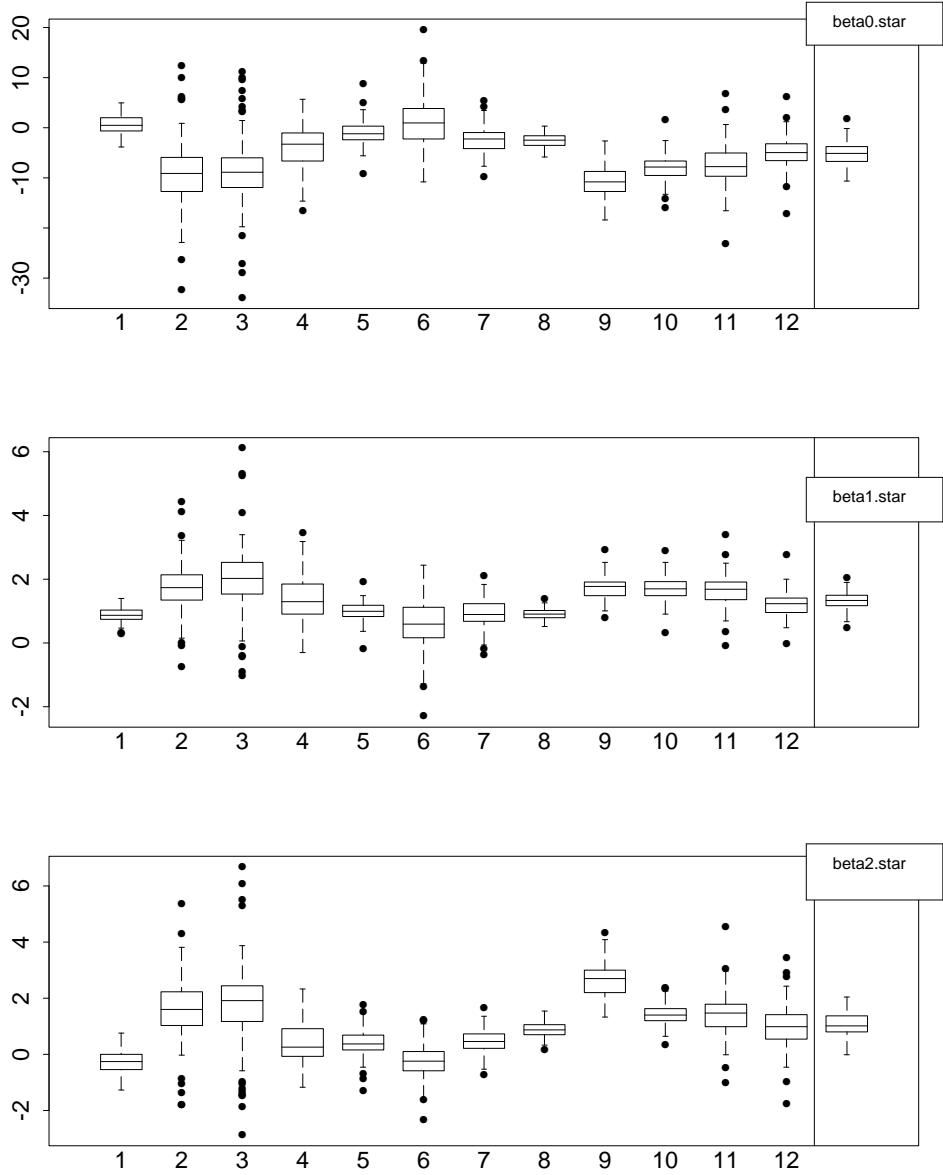


Figure 3: Boxplots of samples from the posterior distributions of lake-specific regressions coefficients B^* . For comparison, sample from the marginal posterior distributions of the corresponding element of B^* are displayed at the far right.

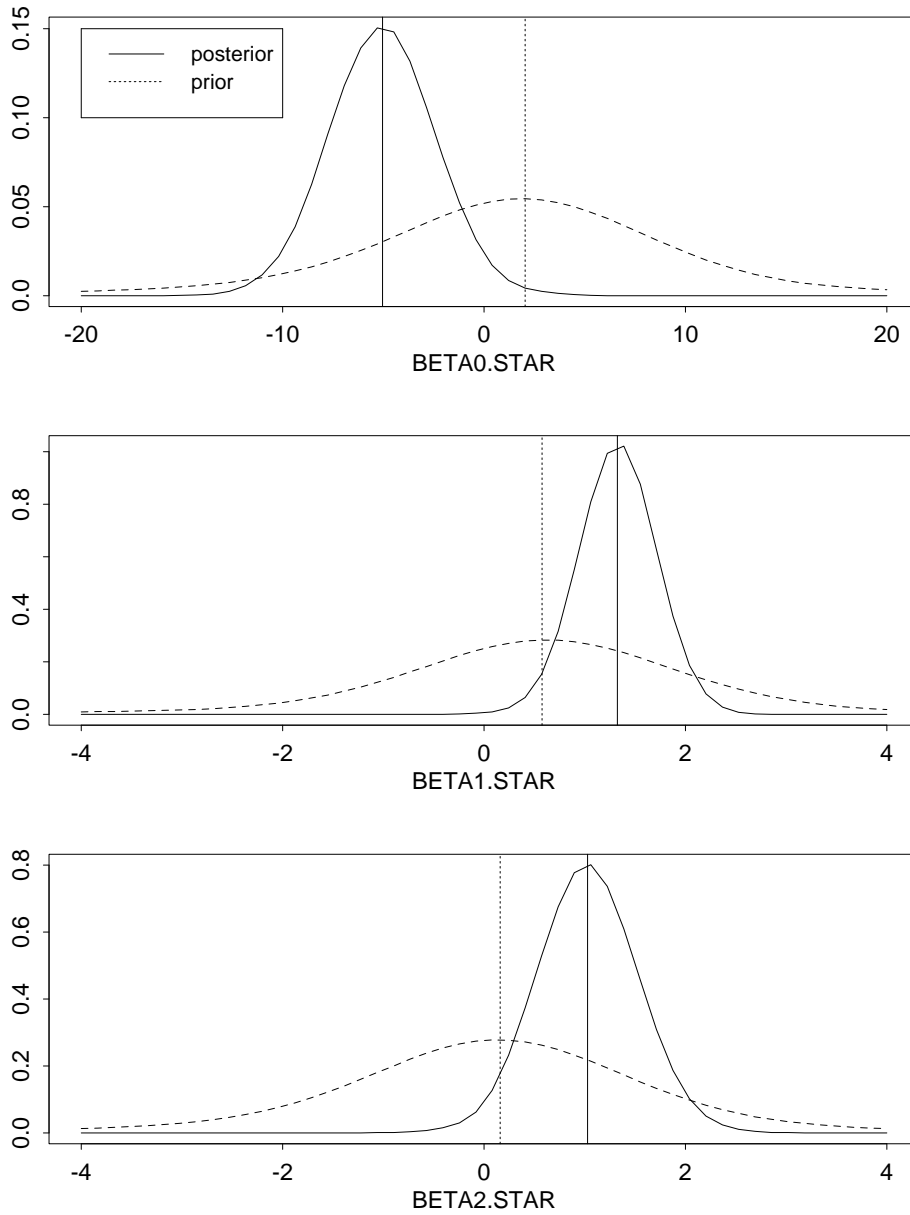


Figure 4: Comparison of prior and posterior distribution on individual components of B^* .

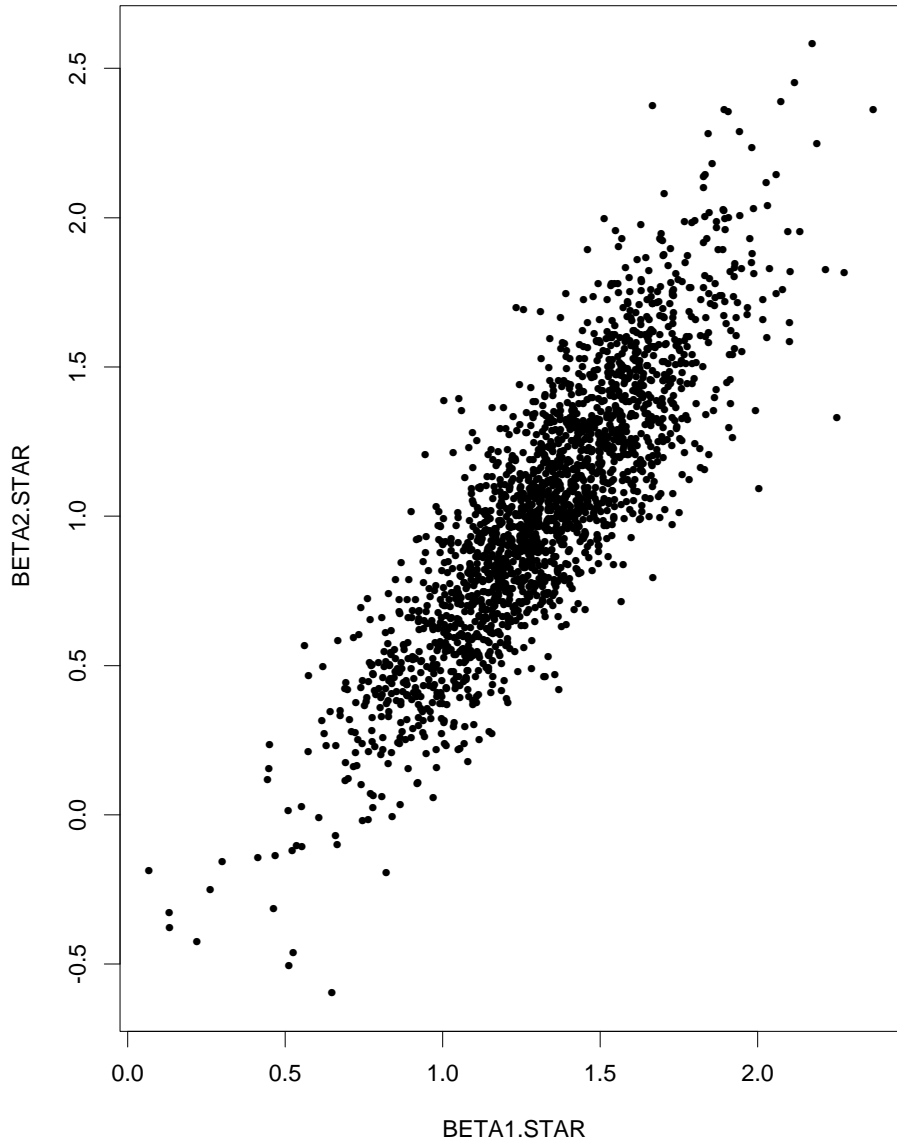


Figure 5: Sample from the joint posterior distribution of β_1^* and β_2^* .

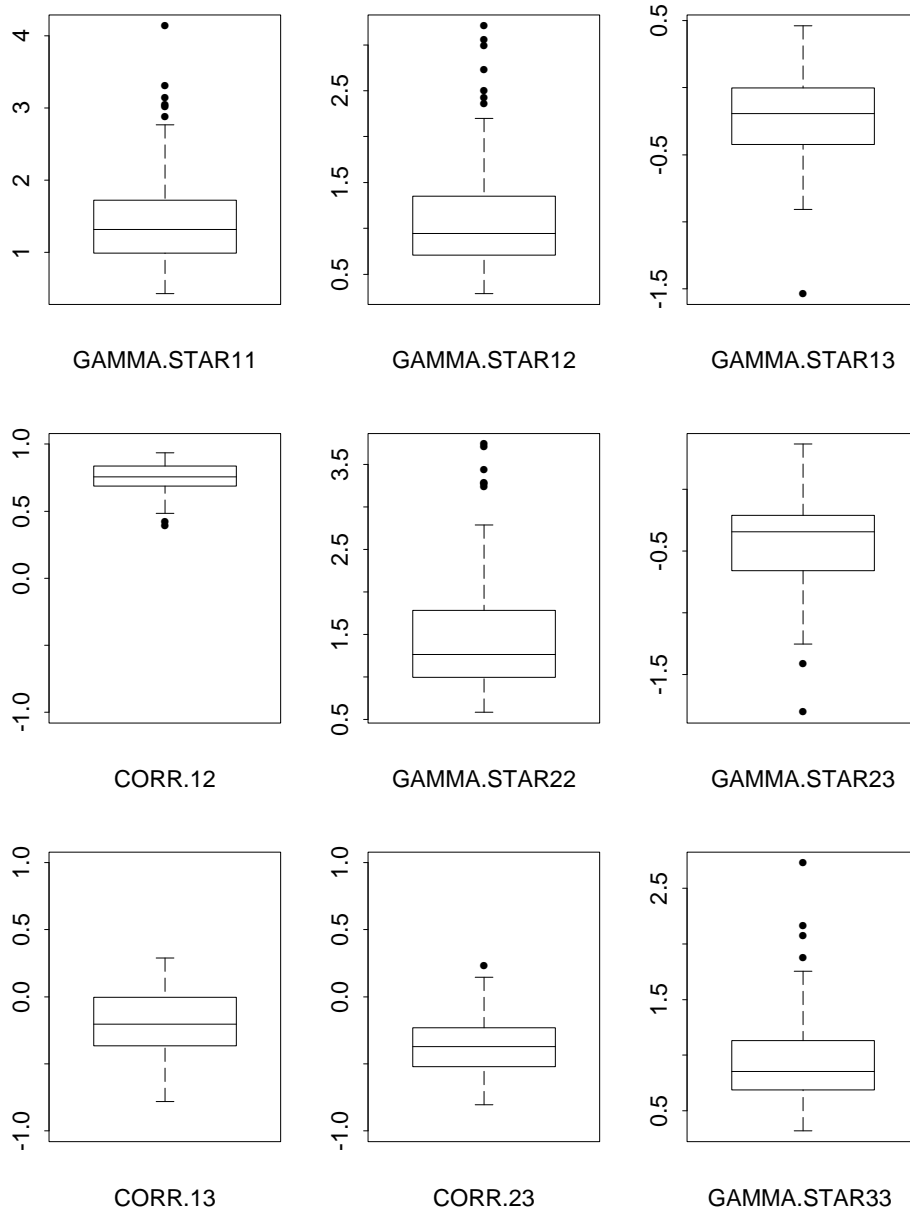


Figure 6: Boxplots of samples from the marginal posterior distributions of the elements of Γ^* . The boxplots are arranged as the corresponding elements in the matrix. Boxplots above the diagonal refer to covariances and below the diagonal refer to correlations. A large diagonal element signifies a large variability in the corresponding measurement from lake to lake. A large off-diagonal element signifies a large spatial correlation between the two variables.

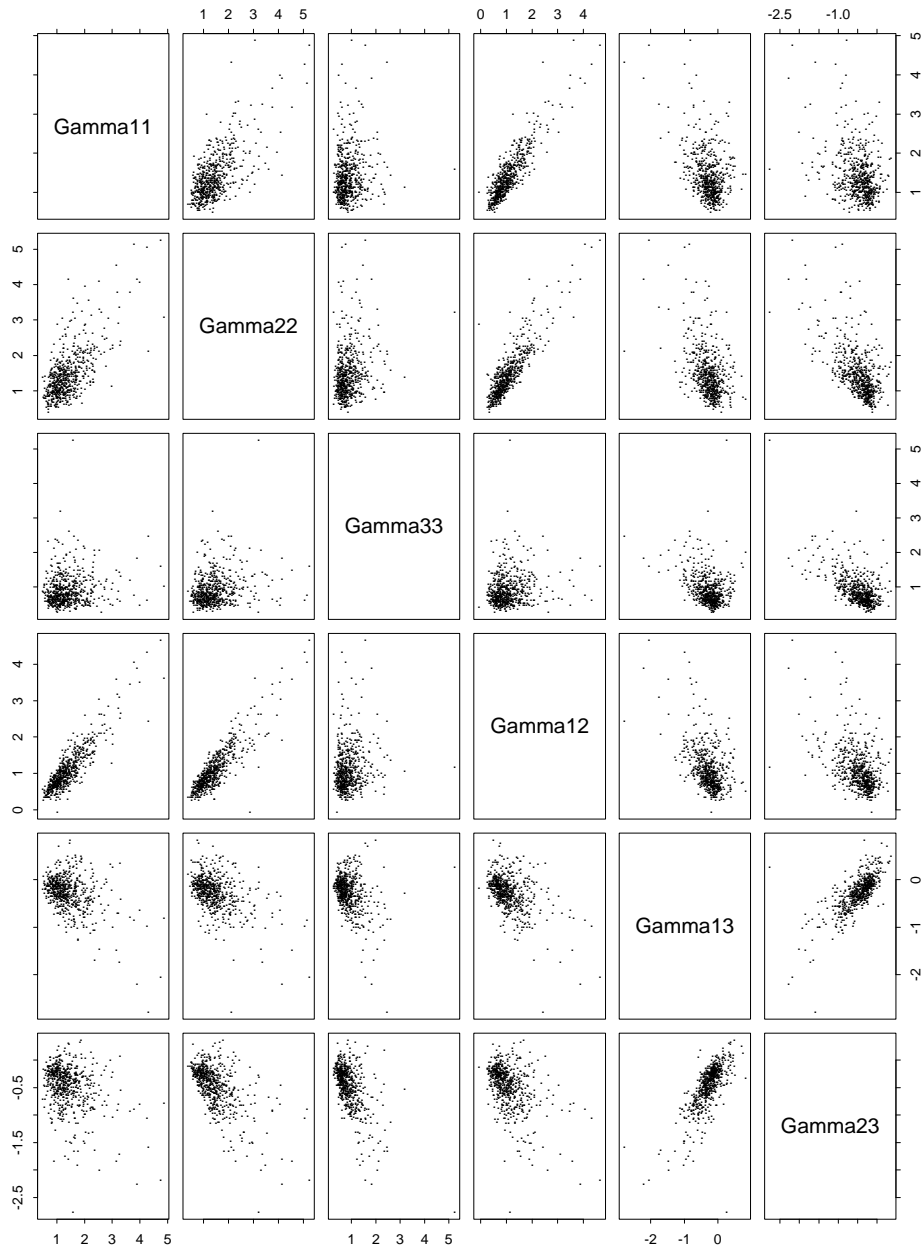


Figure 7: Scatterplots of sample points from the elements of Γ^* .

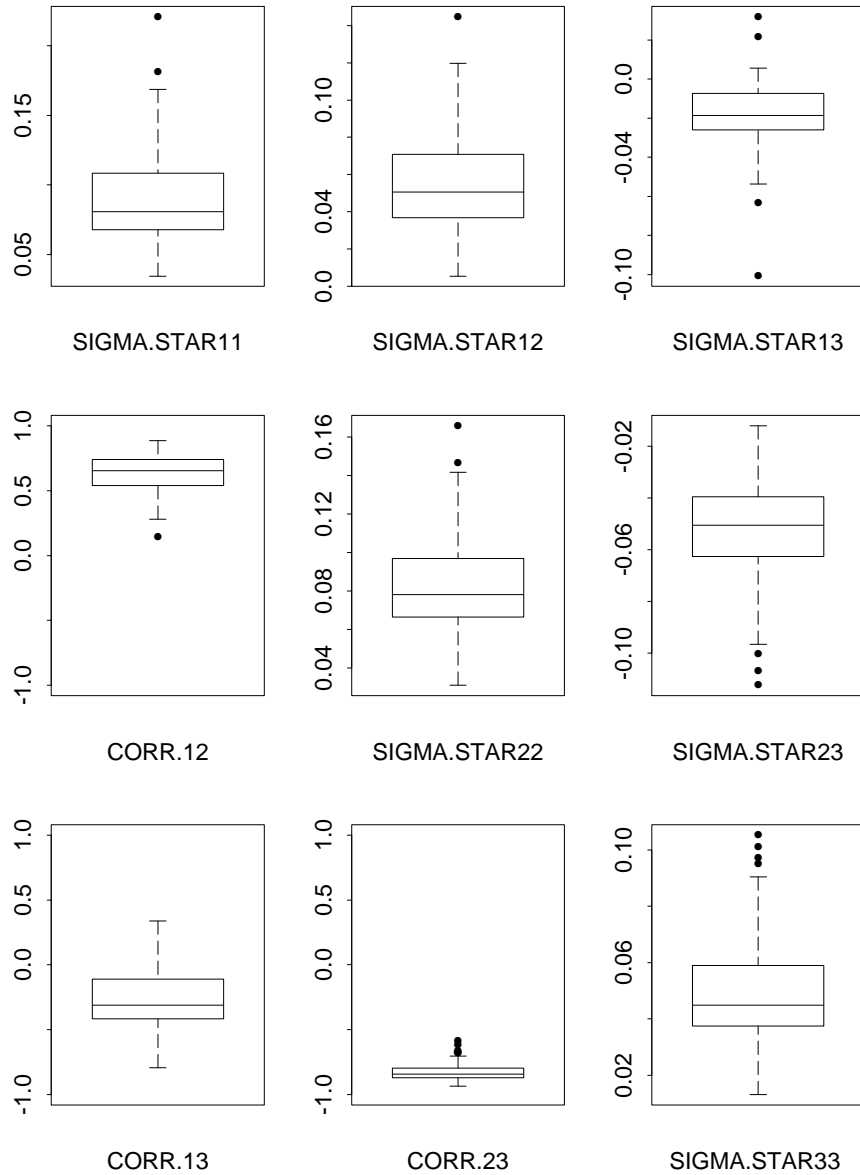


Figure 8: Boxplots of samples from the marginal posterior distributions of the elements of Σ^* . As for Figure 6, boxplots above the diagonal refer to covariances and boxplots below the diagonal refer to correlations. A large diagonal element signifies a large variability in the corresponding measurement over time, on average over lakes. A large off-diagonal element signifies a large temporal correlation between the two variables over time, on average over lakes.

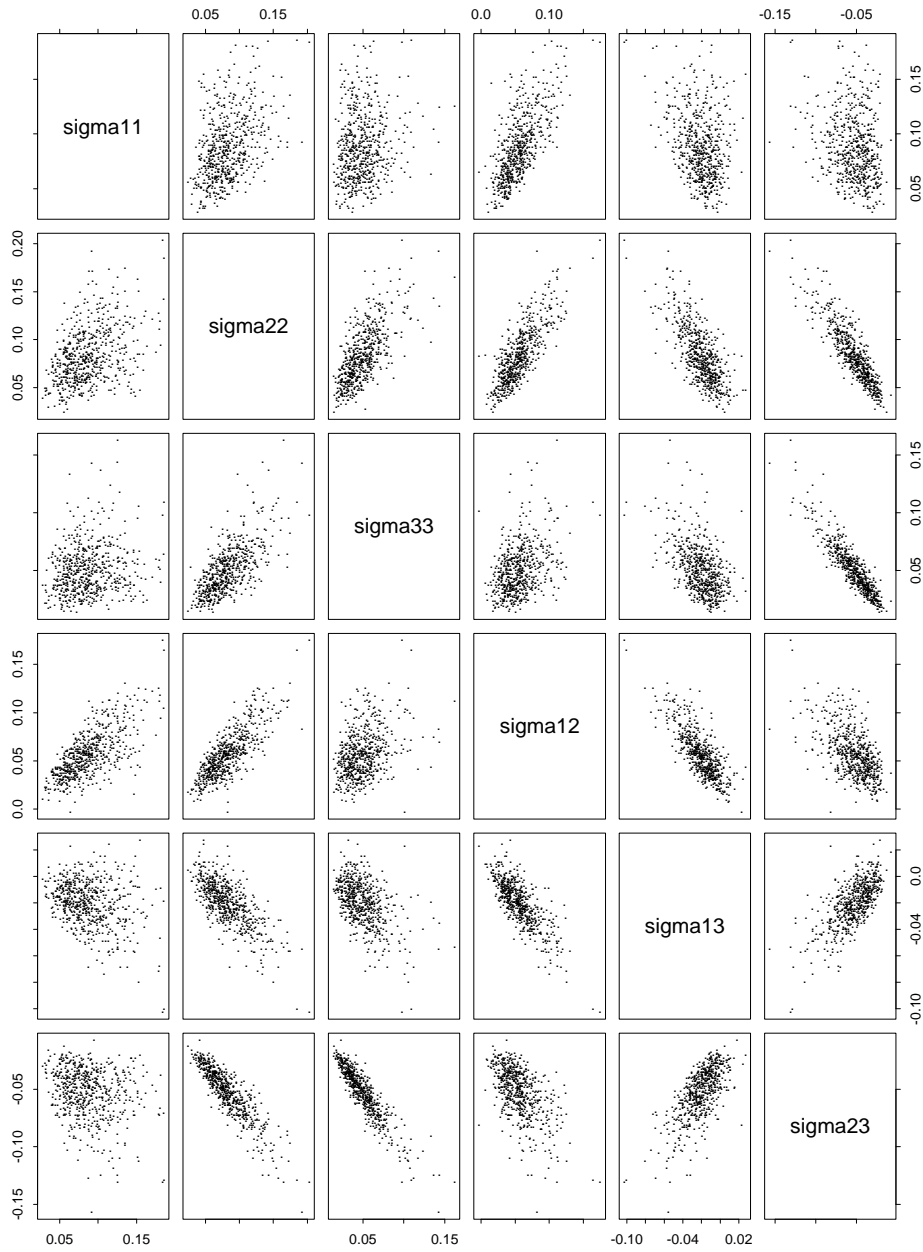


Figure 9: Scatterplots of sample points from all pairwise joint posterior distributions of elements of Σ^* .

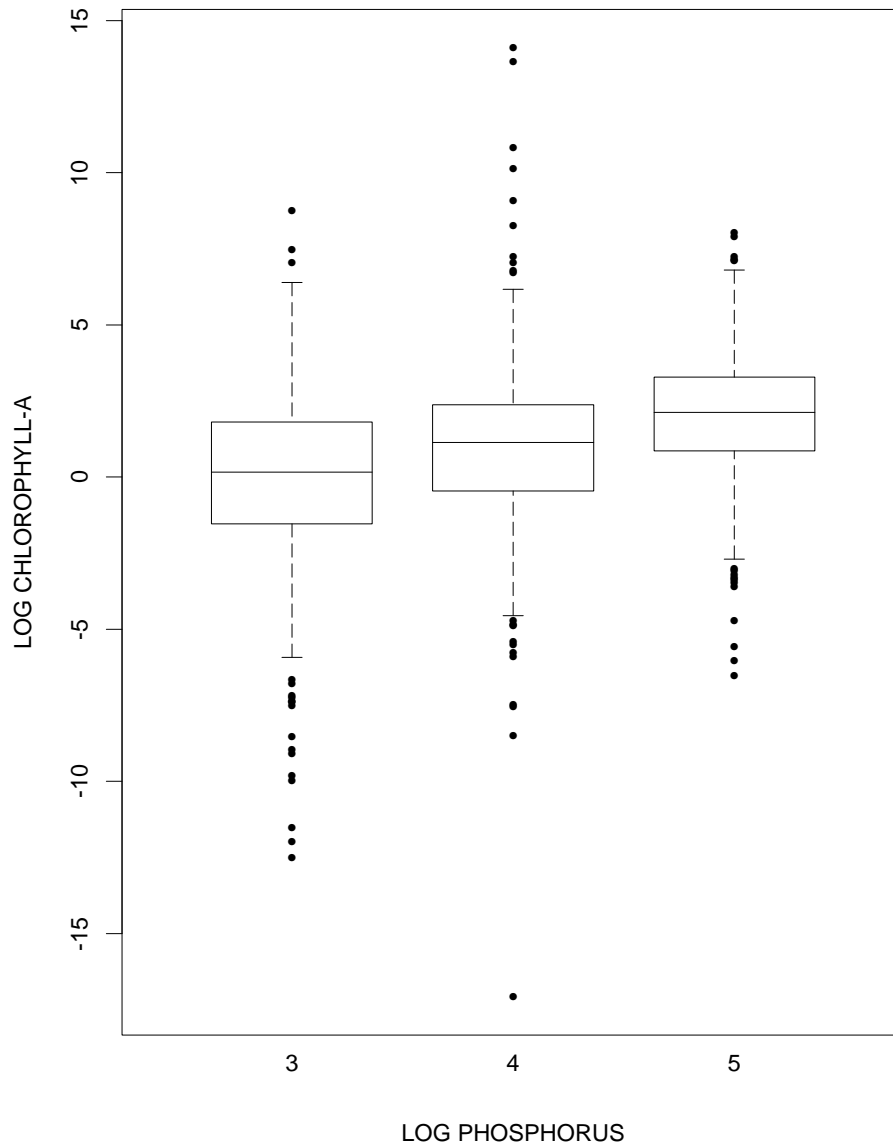


Figure 10: Predictive distribution of log C for three levels of TP in a hypothetical 13-th lake, randomly selected from north temperate lakes. The level of $\log\left(\frac{TN}{TP}\right)$ is set at its mean value 2.71.