UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Julio 2024, Vol. 17, No. 1, pp. 98–115

Bayesian Hierarchical Modeling in Linear Regression Settings

Modelado Jerárquico Bayesiano en Entorno de Regresión Lineal

Juan Sosa[a]
jcsosam@unal.edu.co

Abstract

Considering the flexibility and applicability of Bayesian modeling, main characteristics of a hierarchical model are revised and summarized under the usual assumption of exchangeability: We present the probabilistic structure of the model, all the levels involved in it, and the full conditional distribution of every parameter of the model. In this model, we allow the mean of the second stage of the model to have a linear dependency on a set of covariates by means of a regression approach. In addition, the Gibbs sampling algorithm used to obtain samples from this hierarchical model is fully described and derived. The case study is one in which we characterize in depth the average surface of the sea temperature register by 86 devices in the Mediterranean sea by the type of devise and the location describe by the latitude and the longitude. The hierarchical model fitted considerably well to this data set. Findings derived in this application include the description of the within and between means and variability of the registered temperatures, evidence of similar devise precision, differences among types of devise, and good qualities of prediction of the model. Finally, the prediction ability of the model for each type of devise is tested using data from the National Oceanic and Atmospheric Administration.

***Keywords*:** Gibbs sampling; hierarchical model; regression; surface of the sea temperature.

Resumen

Considerando la flexibilidad y aplicabilidad del modelado Bayesiano, se revisan y resumen las principales características de un modelo jerárquico bajo el supuesto usual de intercambiabilidad: presentamos la estructura probabilística del modelo, todos los niveles involucrados en él y la distribución condicional completa de cada parámetro del modelo. En este modelo, permitimos que la media de la segunda

[a]Departamento de Estadística, Universidad Nacional de Colombia

etapa del modelo tenga una dependencia lineal con un conjunto de covariables mediante un enfoque de regresión. Además, se describe y deriva completamente el algoritmo de muestreo de Gibbs utilizado para obtener muestras de este modelo jerárquico. El estudio de caso es uno en el que caracterizamos en profundidad la superficie media de la temperatura del mar registrada por 86 dispositivos en el mar Mediterráneo, clasificados por el tipo de dispositivo y la ubicación descrita por la latitud y la longitud. El modelo jerárquico se ajustó considerablemente bien a este conjunto de datos. Los hallazgos derivados de esta aplicación incluyen la descripción de las medias y la variabilidad dentro y entre las temperaturas registradas, evidencia de precisión similar entre dispositivos, diferencias entre tipos de dispositivos y buenas cualidades de predicción del modelo. Finalmente, se prueba la capacidad de predicción del modelo para cada tipo de dispositivo utilizando datos de la Oficina Nacional de Administración Oceánica y Atmosférica.

***Palabras clave*:** muestreo de Gibbs; modelo jerárquico; regresión; temperatura de la superficie del mar.

1. Introduction

Bayesian hierarchical modeling is a versatile statistical framework that integrates Bayesian inference with hierarchical modeling techniques, suitable for analyzing data structured into nested or hierarchical levels (Gelman et al., 2013). This approach is particularly advantageous in fields such as education, healthcare, and environmental science, where data naturally exhibit hierarchical relationships, such as students within schools or patients within hospitals. Thus, Bayesian hierarchical modeling provides a powerful framework for analyzing hierarchical data structures, offering flexibility in modeling complex relationships and uncertainty quantification across various scientific disciplines.

At its core, Bayesian hierarchical modeling leverages probabilistic principles to model uncertainty across different levels of data aggregation. It begins with the specification of prior distributions that encapsulate prior beliefs about model parameters. These priors are updated to posterior distributions following the incorporation of observed data, providing a coherent means to quantify uncertainty and make inferences (McElreath, 2018).

Applications of Bayesian hierarchical models are diverse, spanning from predictive modeling and inference to understanding complex relationships within datasets. For instance, researchers use hierarchical models to analyze spatial data in environmental studies or to evaluate educational interventions across different schools (Gelman et al., 2013). Here, we do the former.

Implementing Bayesian hierarchical models often involves sophisticated computational techniques such as Markov Chain Monte Carlo (MCMC) methods (Gamerman and Lopes, 2006), which facilitate sampling from the posterior distribution of parameters. This process allows for the estimation of model parameters and as-

sessment of model fit, crucial for robust statistical analysis Gelman et al. (2013). Despite its advantages, Bayesian hierarchical modeling presents challenges, including the need for careful prior specification that balances prior knowledge with data-driven information. Moreover, the computational intensity of these methods requires adequate resources and expertise in Bayesian statistics (Hoff, 2009).

A feature characteristic of many phenomena in nature is that the observed data, y_{ij}, can be used to estimate aspects of the population distributions of the θ_j even though the values of θ_j are not themselves observed. It is natural to model such a problem hierarchically, with observable outcomes modeled conditionally on certain parameters, which themselves are given a probabilistic specification in terms of further parameters, known as hyperparameters. Hierarchical models can have enough parameters to fit the data well, while using a population distribution to structure some dependence into the parameters, thereby avoiding problems of overfitting (Gelman et al., 2013, p.101). In addition, by establishing hierarchies we are not forced to choose between complete pooling and not at all as the classic analysis of variance does (Gelman et al., 2013, pp.101, 115). In this way, the hierarchical model in the linear regression setting is a conceptually straightforward generalization of the normal hierarchical model. We use an ordinary regression model to describe within-group heterogeneity of observations, then describe between-group heterogeneity using a sampling model for the group-specific regression parameters (Hoff, 2009, p.196).

In this way, considering the flexibility and applicability of Bayesian modeling, main characteristics of a hierarchical model are revised and summarized under the usual assumption of exchangeability: We present the probabilistic structure of the model, all the levels involved in it, and the full conditional distribution of every parameter of the model. In this model, we allow the mean of the second stage of the model to have a linear dependency on a set of covariates by means of a regression approach. In addition, the Gibbs sampling algorithm used to obtain samples from this hierarchical model is fully described and derived.

The case study is one in which we characterize in depth the average surface of the sea temperature register by 86 devices in the Mediterranean sea by the type of devise and the location describe by the latitude and the longitude. The hierarchical model fitted considerably well to this data set. Findings derived in this application include the description of the within and between means and variability of the registered temperatures, evidence of similar devise precision, differences among types of devise, and good qualities of prediction of the model. Finally, the prediction ability of the model for each type of devise is tested using data from the National Oceanic and Atmospheric Administration (NOAA).

This article is structure as follows: Section 2 provides all the details about the the hierarchical normal model including model specification and posterior inference. Then, Section 3 presents a real-world application where we fully characterize in depth the average surface of the sea temperature register by 86 devices in the Mediterranean sea. Finally, Section 4 discusses our main findings and shows some ideas for future research.

2. The hierarchical normal model

Here we present the treatment of a hierarchical model (with a regression level on it) based on the normal distribution, in which observed data is assumed to be normally distributed with a different mean and variance for each group or experiment. Here we follow the same notation as inGelman et al. (2013, Chap. 5).

2.1. The model

We take into account m independent groups, each group with n_i independent normally distributed data points, y_{ij}, each with unknown mean (group effect) μ_i and unknown variance σ_i^2; that is, $y_{ij} \sim \mathsf{N}(\mu_i, \sigma_i^2)$, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$. This means that $\bar{y}_{i\cdot} \sim \mathsf{N}(\mu_i, \sigma_i^2/n_i)$ where $\bar{y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ is the sample mean of the i-th group, $i = 1, \ldots, m$.

We consider a hierarchical model in which the hierarchies are given by:

$$
\begin{aligned}
p(\bar{y}_{i\cdot} \mid \mu_i, \sigma_i^2) &= \mathsf{N}(\bar{y}_{i\cdot} \mid \mu_i, \sigma_i^2/n_i) && \text{Likelihood} \\
p(\mu_i \mid \boldsymbol{\beta}, \tau^2) &= \mathsf{N}(\mu_i \mid \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}, \tau^2) && \text{Stage I} \\
p(\sigma_i^2 \mid \xi^2) &= \mathsf{IG}(\sigma_i^2 \mid \alpha + 1, \alpha\xi^2) && \text{Stage I} \\
p(\boldsymbol{\beta}, \tau^2) &\propto 1 && \text{Stage II} \\
p(\xi^2) &= \mathsf{G}(\xi^2 \mid a, b) && \text{Stage II}
\end{aligned}
$$

where the μ_i and the σ_i^2 are parameters of the model, $\boldsymbol{\beta}$, τ^2, and ξ^2 are the corresponding hyperparameters (in particular $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^{\mathsf{T}}$ is a vector of p

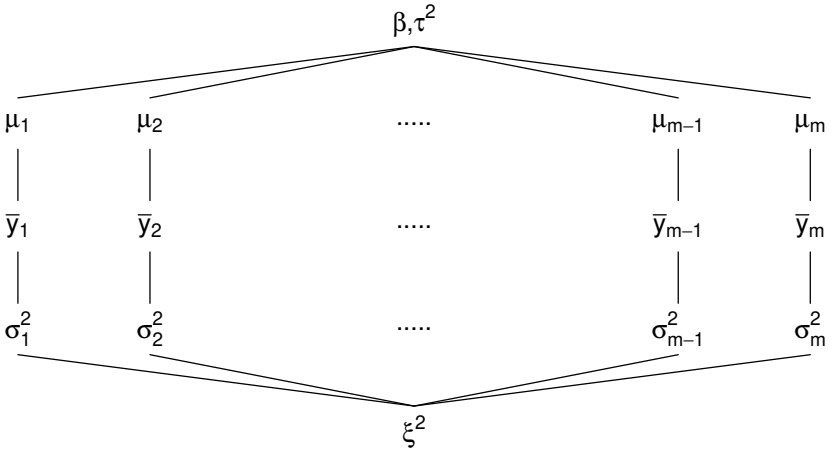

Figure 1: Graphical representation of a hierarchical model.

regressors), $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{ip}]^\mathsf{T}$ is a vector of p explanatory variables, and α, a, and b are fixed constants picked according to external information of the data set (this hierarchical model is illustrated in figure 1). Thus, the unknown quantities in this model ($2m+p+2$ in total) include the group-specific means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ and variances $\boldsymbol{\sigma^2} = (\sigma_1^2, \ldots, \sigma_m^2)$, the mean and variance $(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}, \tau^2)$ of the population of group-specific means, and the variance ξ^2 of the population of group-specific variances. This model is susceptible of generalization; for instance, we could allow the model to have a proper prior for $(\boldsymbol{\beta}, \tau^2)$, and we also could include more hyperparameters in the model (α does not need to be fixed). See for example Hoff (2009, p. 143).

2.2. Posterior inference

Joint posterior inference for the parameters can be made by construction a Gibbs sampler which requires iteratively sampling each parameter from its full conditional distribution.

2.2.1. Posterior distribution

Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})$ be the measurements of group i, $i = 1, \ldots, m$, $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)$ be the entire set of observations, and $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{\beta}, \tau^2, \xi^2)$ be the full parameter-hyperparameter vector. The posterior distribution of $\boldsymbol{\theta}$ is then given by

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = p(\boldsymbol{\beta}, \tau^2)\, p(\boldsymbol{\mu} \mid \boldsymbol{\beta}, \tau^2)\, p(\xi^2) p(\boldsymbol{\sigma^2} \mid \xi^2)\, p(\boldsymbol{y} \mid \boldsymbol{\theta})\,,$$

which leads to

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{y}) &= \prod_{i=1}^{m} \mathsf{N}(\mu_i \mid \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}, \tau^2) \times \mathsf{G}(\xi^2 \mid a, b) \times \prod_{i=1}^{m} \mathsf{IG}(\sigma_i^2 \mid \alpha+1, \alpha\xi^2) \times \prod_{i=1}^{m} \mathsf{N}(\bar{y}_{i\cdot} \mid \mu_i, \sigma_i^2/n_i) \\ &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\tau} \exp\left\{-\frac{1}{2\tau^2}(\mu_i - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})^2\right\} \times \frac{b^a}{\Gamma(a)}(\xi^2)^{a-1} e^{-b\xi^2} \\ &\quad \times \prod_{i=1}^{m} \frac{(\alpha\xi^2)^{\alpha+1}}{\Gamma(\alpha+1)}(\sigma_i^2)^{-((\alpha+1)+1)} e^{-(\alpha\xi^2)\sigma_i^2} \times \prod_{i=1}^{m} \frac{n_i}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{n_i}{2\sigma_i^2}(\bar{y}_{i\cdot} - \mu_i)^2\right\}. \end{aligned}$$

Although this is an abuse of standard mathematical notation, the *full conditional distribution* (fcd) of parameter ϕ given the rest of the parameters and the data $\boldsymbol{y}$ is denoted by $p(\phi \mid \text{rest}, \boldsymbol{y})$. We derived these distributions looking at the dependencies in the full posterior distribution. Thus, we have that:

- The fcd of μ_i, $i = 1, \ldots, m$ is

$$p(\mu_i \mid \text{rest}) = \mathsf{N}\left(\mu_i \,\middle|\, \frac{\frac{n_i\bar{y}_{i\cdot}}{\sigma_i^2} + \frac{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}}{\tau^2}}{\frac{n_i}{\sigma_i^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n_i}{\sigma_i^2} + \frac{1}{\tau^2}}\right).$$

- The fcd for σ_i^2, $i = 1, \ldots, m$, is
$$p(\sigma_i^2 \mid \text{rest}) = \mathsf{IG}\left(\sigma_i^2 \mid \alpha + 3/2, \alpha\xi^2 + n_i(\bar{y}_{i\cdot} - \mu_i)^2/2\right) .$$
- The fcd of ξ^2 is
$$p(\xi^2 \mid \text{rest}) = \mathsf{G}\left(\xi^2 \,\middle|\, a + m(\alpha+1), b + \alpha\sum_{i=1}^{m} 1/\sigma_i^2\right) .$$
- The fcd of τ^2 is
$$p(\tau^2 \mid \text{rest}) = \mathsf{IG}\left(\tau^2 \mid m/2 - 1, \|\boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}\|^2/2\right) ,$$
where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m]^\mathsf{T}$.
- The fcd for $\boldsymbol{\beta}$ is
$$p(\boldsymbol{\beta} \mid \text{rest}) = \mathsf{N}\left(V_{\boldsymbol{\beta}}\boldsymbol{X}^\mathsf{T}\boldsymbol{\mu}, \tau^2 V_{\boldsymbol{\beta}}\right) ,$$
where $V_{\boldsymbol{\beta}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}$.

2.2.2. Gibbs sampling algorithm

Let $\phi^{(m)}$ denote the state of parameter ϕ in the m-th iteration of the algorithm. The algorithm is as follows:

1. Choose an initial configuration for each parameter in the model, say $\mu_1^{(0)}, \ldots, \mu_m^{(0)}$, $(\sigma_1^2)^{(0)}, \ldots, (\sigma_m^2)^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $(\tau^2)^{(0)}$, and $(\xi^2)^{(0)}$.
2. Update $\mu_1^{(m-1)}, \ldots, \mu_m^{(m-1)}$, $(\sigma_1^2)^{(m-1)}, \ldots, (\sigma_m^2)^{(m-1)}$, $\boldsymbol{\beta}^{(m-1)}$, $(\tau^2)^{(m-1)}$, and $(\xi^2)^{(m-1)}$ until convergence:
   a) Sample $\mu_i^{(m)}$, $i = 1, \ldots, m$, from
   $$p\left(\mu_i \middle| (\sigma_i^2)^{(m-1)}, \boldsymbol{\beta}^{(m-1)}, (\tau^2)^{(m-1)}, \boldsymbol{y}\right) .$$
   b) Sample $(\sigma_i^2)^{(m)}$, $i = 1, \ldots, m$, from
   $$p\left(\sigma_i^2 \middle| \mu_i^{(m)}, (\xi^2)^{(m-1)}, \boldsymbol{y}\right) .$$
   c) Sample $(\xi^2)^{(m)}$ from
   $$p\left(\xi^2 \middle| (\sigma_1^2)^{(m)}, \ldots, (\sigma_m^2)^{(m)}, \boldsymbol{y}\right) .$$
   d) Sample $(\tau^2)^{(m)}$ from
   $$p\left(\tau^2 \middle| \mu_1^{(m)}, \ldots, \mu_m^{(m)}, \boldsymbol{\beta}^{(m-1)}, \boldsymbol{y}\right) .$$
   e) Sample $\boldsymbol{\beta}^{(m)}$ from
   $$p\left(\boldsymbol{\beta} \middle| \mu_1^{(m)}, \ldots, \mu_m^{(m)}, (\tau^2)^{(m)}, \boldsymbol{y}\right) .$$
3. Cycle until achieve convergence.

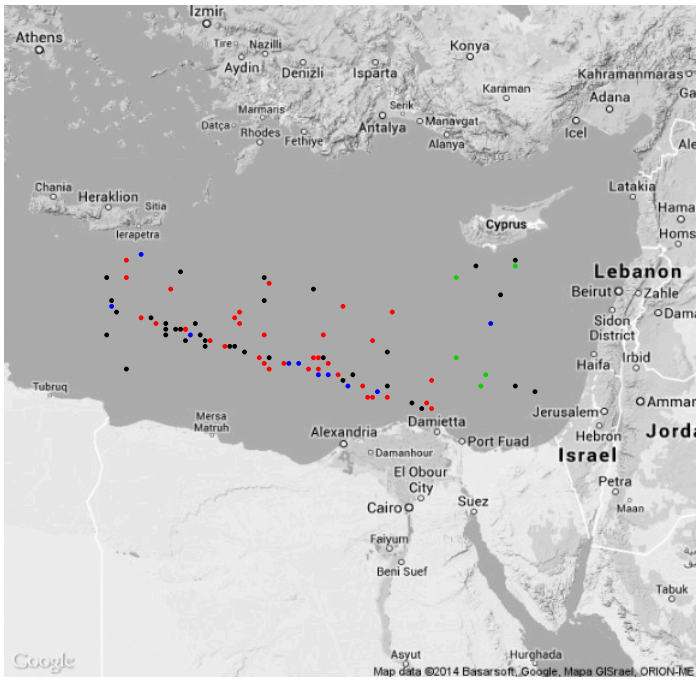

Figure 2: Location of each device in the Mediterranean Sea: bucket (black), eri (red), f.buoy (green), and d.buoy (blue).

3. Application

The data set contains the mean of the observations recorded by $m = 86$ different devices on the *temperature of the sea* (SST) along with the type of device (four categories: bucket, eri, f.buoy, and d.buoy) in a number of locations (for which the latitude and longitude is available) in the Mediterranean in December 2003 (see Figure 2). Thus, the data available of the $j-th$ device, $j = 1, \ldots, 86$, is given by the array

$$(\text{latitud}_i, \text{latitud}_i, \bar{y}_{i\cdot}, n_i, \text{type}_i)$$

where $\bar{y}_{i\cdot}$ is the average of the n_i temperatures (in Celsius) recorded by device i. This way, we have that $\boldsymbol{y} = [\bar{y}_1, \ldots, \bar{y}_{86}]$, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_{12}]^T$, and $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{i12}]^T$ is a p-dimensional vector of covariates, $p = 12$, with $x_{i1} = 1$ for each device i, $x_{i2} = 1$ if $\text{type}_i = \textbf{eri}$, and 0 otherwise, $x_{i3} = 1$ if $\text{type}_i = \textbf{f.buoy}$, and 0 otherwise, $x_{i4} = 1$ if $\text{type}_i = \textbf{d.buoy}$, and 0 otherwise, $x_{i5} = \text{latitude}_i$, $x_{i6} = x_{i2} \times x_{i5}$, $x_{i6} = x_{i3} \times x_{i5}$, $x_{i6} = x_{i4} \times x_{i5}$, $x_{i9} = \text{longitude}_i$, $x_{i10} = x_{i2} \times x_{i9}$, $x_{i11} = x_{i3} \times x_{i9}$, and $x_{i12} = x_{i4} \times x_{i9}$, which under the assumptions of the model implies that

- If $\text{type}_i = \textbf{bucket}$ then

$$\mathsf{E}\left(\mu_i | \boldsymbol{\beta}, \boldsymbol{x}_i\right) = \boldsymbol{\beta_1} + \boldsymbol{\beta_5}\text{lat}_i + \boldsymbol{\beta_9}\text{lon}_i\,,$$

- If $\text{type}_i = \textbf{eri}$ then

$$\mathsf{E}\,(\mu_i|\boldsymbol{\beta}, \boldsymbol{x}_i) = (\beta_1 + \boldsymbol{\beta_2}) + (\beta_5 + \boldsymbol{\beta_6})\text{lat}_i + (\beta_9 + \boldsymbol{\beta_{10}})\text{lon}_i\,,$$

- If $\text{type}_i = \textbf{f.buoy}$ then

$$\mathsf{E}\,(\mu_i|\boldsymbol{\beta}, \boldsymbol{x}_i) = (\beta_1 + \boldsymbol{\beta_3}) + (\beta_5 + \boldsymbol{\beta_7})\text{lat}_i + (\beta_9 + \boldsymbol{\beta_{11}})\text{lon}_i\,,$$

- If $\text{type}_i = \textbf{d.buoy}$ then

$$\mathsf{E}\,(\mu_i|\boldsymbol{\beta}, \boldsymbol{x}_i) = (\beta_1 + \boldsymbol{\beta_4}) + (\beta_5 + \boldsymbol{\beta_8})\text{lat}_i + (\beta_9 + \boldsymbol{\beta_{12}})\text{lon}_i\,.$$

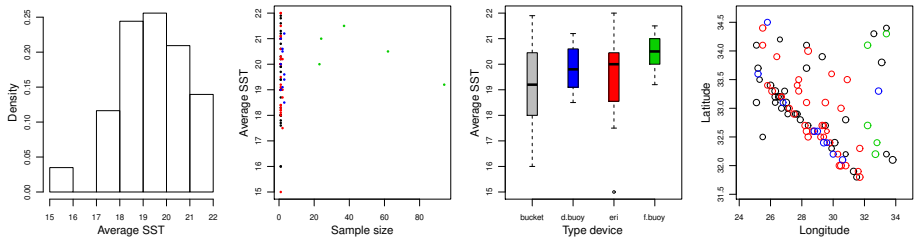

Figure 3: Descriptive plots: bucket (black), eri (red), f.buoy (green), and d.buoy (blue).

3.1. Exploratory data analysis

A histogram of the sample averages is shown in the first panel of 3. The range of average SSTs is from 15° to 22° which seems quite large in comparison with the mean SST in 2003 (20.5° according to Skliris et al. 2011). The second panel shows the relationship between the sample average and the sample size. This plot shows a peculiarity of this data set en relation to the sample sizes: $n_i \leq 3$ for 94 % of the devices, whereas for just a few devices the sample size is bigger, namely, $n_5 = 23$, $n_6 = 24$ $n_7 = 37$, $n_8 = 62$, and $n_9 = 94$, which are f.buoy devices. In addition, as expected, this plot points out that in general devices with very high or very low sample averages tend to be those devices with low sample sizes.

Furthermore, the third panel of figure 3 summarizes the main features about the distribution of the mean temperature by type of devise; taking into account this plot we perceive some differences among types of devise (also capture by the model) in location and maybe scale. On the other hand, the fourth panel shows the location of each devise. In this plot the center of each circle corresponds to the location of the device according to its longitude and latitude, whereas the radius is proportional to the mean SST. Colors in this figure correspond to the type of device. It's very interesting to be aware of two main facts: f.buoy devices (the ones with largest sample sizes) are located in the east of the Mediterranean Sea (see figure 2 to see the location in more detail); and all the circles seem to have similar radius, which strongly suggests that variances of population of group-specific means are similar (this fact will be relevant to set the fixed constants in the model (α, a, and b) according to the external information).

3.2. Prior distributions

To run the Gibbs sampling algorithm described in section 2.2, we need to pick appropriate values for a, b, and α. Hyperparameters a and b control the the prior distribution for the scale parameter (ξ^2) that influences the distribution of the population of group-specific variances. According to the SSTs given in Figure 2 of Skliris et al. (2011), in 2003 the variability in the temperature of Mediterranean Sea is roughly 1.08°. On the other hand, according to the data set, the variability of the sample mean temperatures is about 1.96 (which is clearly an overestimate of the variance reported in Skliris et al. 2011).

Having this in mind, we only weakly concentrate the prior distribution around this value by taking $\alpha = 9$, $a = 10$ and $b = a/2$: doing it in this way is really helpful mainly because $\mathsf{E}\left(\xi^2 \mid a, b\right) = 2$ and $\mathsf{Var}\left(\xi^2 \mid a, b\right) = 0.4$ which makes that $\mathsf{E}\left(\sigma_i^2 \mid \xi^2\right)$ is around 2 and $\mathsf{Var}\left(\sigma_i^2 \mid \xi^2\right) = 0.6$, i.e., a weakly concentration of σ_i^2 around 2 ($cv = 35\,\%$). Running the Gibbs sampler algorithm produces a $S \times (2m + p + 2)$ matrix containing a value of the SST mean and variance for each device (μ_i, σ_i^2), and values of $\boldsymbol{\beta}$, τ^2 and ξ^2 at each iteration of the Markov chain.

3.3. MCMC diagnostics

Before doing inference using the MCMC samples we should determine if there might be any problems with the Gibbs sampler. The first thing we want to do is to see if there are any indications that the chain is not stationary, i.e., if the simulated parameter values are moving in a consistent direction. Standard practice is to plot to produce boxplots of sequential groups of samples (Hoff, 2009, p. 139). We do this in the first row of Figure 4. There does not seem to be any evidence that the chain has not achieved stationarity.

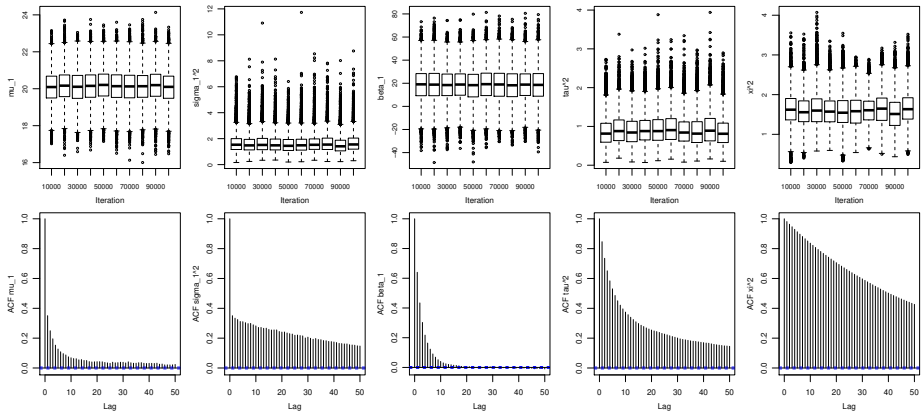

Figure 4: Stationary plots and autocorrelation plots for some parameters and hyperparameters of the model.

Second row of Figure 4 shows autocorrelation plots for the same parameters. Even though there does seem strong dependence in the chain for and τ^2 and ξ^2 (the plot for the rest looks well), we are very confident of the samples we got since effective sample sizes range from 887 to 94,144. In addition, approximate Monte Carlo standard errors can be obtained by dividing the approximated posterior standard deviations by the square root of the effective sample sizes, giving values ranging from 0.000002 to 0.0015. These are small compared to the scale of the approximated posterior expectations of these parameters.

3.4. Posterior summaries

In what follows we summarize the posterior distributions of the parameters of the model. Panels 1 an 2 of Figure 5 show Monte Carlo summaries to the posterior densities of each μ_i and σ_i^2, respectively, by type of device; whereas panels 3 and 4 of this figure show Monte Carlo approximations to the the posterior densities of τ^2 and ξ^2. From the first panel, we see that the 95 % (posterior) credible intervals for bucket and eri have approximately the same coverture; the credible interval length for d.bouy devices seems to be smaller than the previous two; and finally, there is a relevant difference in the coverture shown by f.bouy devices. Here we see the effect of the sample size: when we have more data available in each device, the credible intervals are more precise. This plot also suggests that the SST detected by f.buoy devices is slightly greater in comparison with the rest. On the other hand, the second panel of this figure exhibits how the specific-device variances are around the same vale (about 1.6) which approximately agrees with the value reported by Skliris et al. (2011), so we feel confident with the use of the external information we did. In summary, the model is able to detect a slightly difference in the mean SST due to a f.dbuoy effect, may be masked by the amount of measurements available for those devices; group-specific variability seem stable and close to each other, non perturbed among devices and type of devices.

We present Figure 6 to illustrate the shrinkage, e.i, the fact that the posterior expected value of each μ_i is pulled a bit from $\bar{y}_{i\cdot}$ towards $\boldsymbol{x}_i T\boldsymbol{\beta}$ by an amount

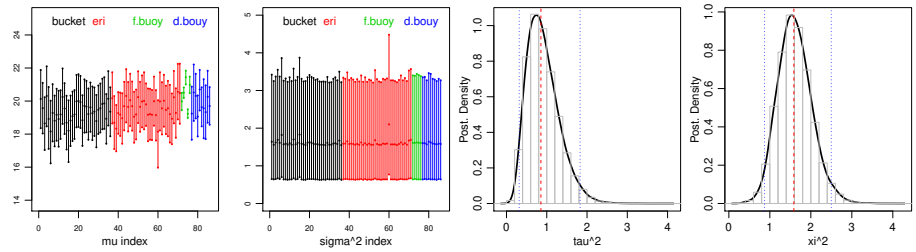

Figure 5: Panel 1 and 2 show the posterior mean and the 95 % credible interval of each μ_i and σ_i^2 by type of device, respectively: bucket (black), eri (red), f.buoy (green), and d.buoy (blue). Panel 3 and 4 show the marginal posterior distribution of τ^2 and ξ^2, respectively.

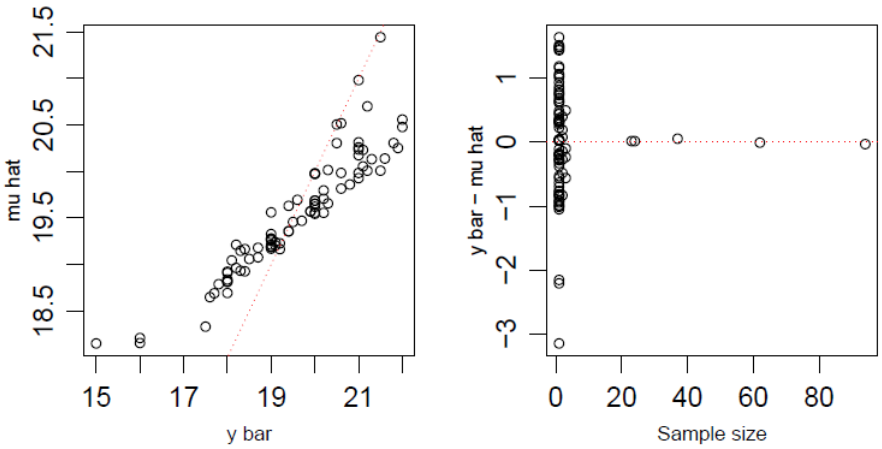

Figure 6: Shrinkage for μ_i.

depending of n_i (Hoff 2009, p. 140). Notice that the relationship in the first panel of Figure 6 roughly follows a line with a slope that is less than one, indicating that high values of $\bar{y}_{i\cdot}$ correspond to slightly less high values of $\hat{\mu}_i$, and low values of $\bar{y}_{i\cdot}$ correspond to slightly less low values of $\hat{\mu}_i$. The second panel of 6 shows the amount of shrinkage as a function of the group-specific sample size. Devises with low sample sizes get shrunk the most, whereas groups with large sample sizes hardly get shrunk at all.

Now, we present the main findings about the regressors in the second stage of

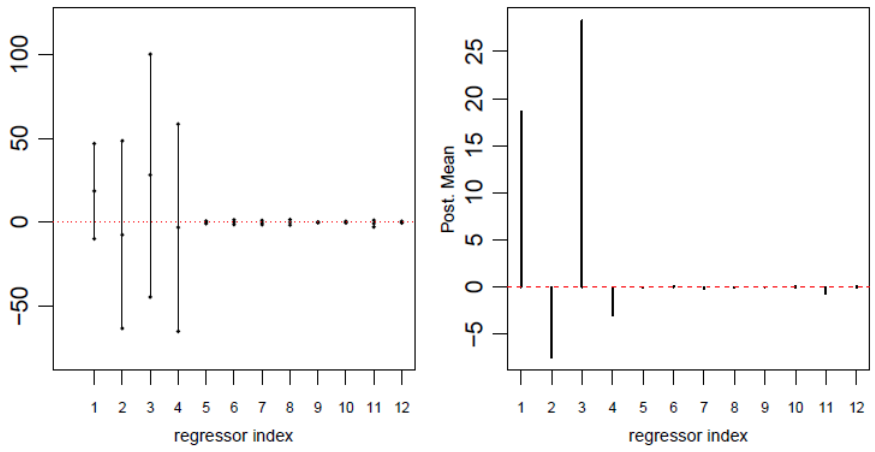

Figure 7: The first panel shows the posterior mean and the 95 % credible interval of each β_i; the second panel also presents the posterior mean of these parameters.

the model. Figure 8 display the main features of the posterior distribution of each β_i including the posterior mean and a 95 % credible interval. These plots strongly suggest that the first four regressors are the only ones that have an impact over the group-specific SST means since their credible intervals are not centered at zero and have the largest coverture. The impact of the rest of the coefficients, the location effect of each type of device, is almost imperceptible. This feature is also illustrated in Figure 8 by looking at the support of the joint posterior distribution of some pairs of parameters. In summary, the covariates that significantly contribute to explain the group-specific SST are the ones associated with mean effect due to the type of device, whereas the impact recorded by the location (latitude and longitude) of the device is almost null.

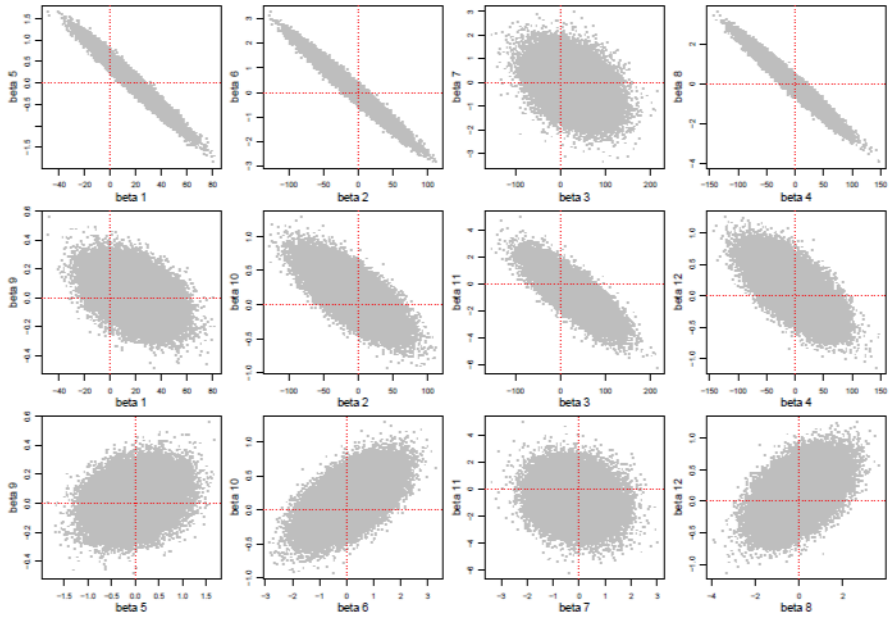

Figure 8: Bivariate posterior distribution for some pairs of $\boldsymbol{\beta}$.

Finally, recalling that β_1 is the mean influence of bucket, we have that $\beta_1 + \beta_2$, $\beta_1 + \beta_3$, and $\beta_1 + \beta_4$, correspond to the mean impact of eri, f.buoy, and d.bbuoy, respectively, over the device-specific mean of the average SST (which means that β_2, β_3, and β_4, represent the difference of the mean effect of eri, f.buoy, and d.bbuoy, respectively, with respect to the effect of bucket). Having this in mind we have that the standardized scores of each type of device are: 1.29 (bucket), 0.46 (eri), 1.39 (f.buoy), and 0.56 (d.buoy). Now, we can give us a precise idea of the influence of each type of device over the device-specific mean of the average SST.

3.5. Model checking

Before proceeding with predicting some temperature predictions, we check the performance of the model by replicating new data (Gelman et al., 2013, Chap. 6) and calculating some statistics and then compere them with actual values given in the data. Figure 9 displays the empirical distribution of the IQR, median, mean, and standard deviation of replicated data along with the observed value (vertical line in red) in actual data set and the corresponding. In the corner of each panel is also shown the value of the corresponding PPP (posterior predictive p-vale) which can be calculated as

$$p = \mathsf{P}(T(\boldsymbol{y}^{\text{rep}}) > T(\boldsymbol{y}) \mid \boldsymbol{y})$$

where $\boldsymbol{y}^{\text{rep}}$ is the predictive data and T is the so called test statistic. About the PPPs there is nothing to be worry about since they are as calibrated as any other model-based probability (Gelman et al., 2013, p.152). We clearly see that there is no anomaly in those distributions that make us doubt about the goodness of fit of our model since the PPPs do not take extreme values and the observed values are always in the range of the the predicted ones, which are well behaved.

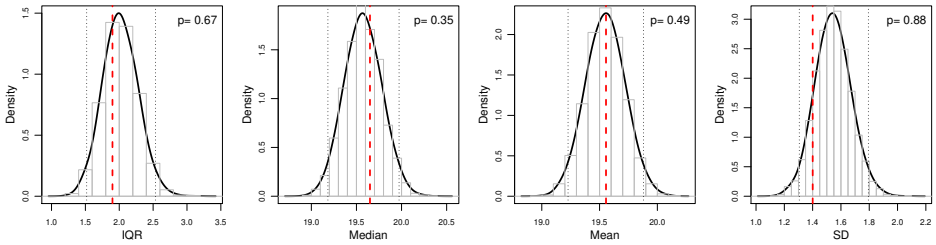

Figure 9: Model checking using empirical distributions of statistics of predicted data.

3.6. Model prediction

Now we use the model to predict temperatures from a new device. In such a case we first randomly select a sample in the chain from which we take the corresponding values of $\boldsymbol{\beta}$, τ^2, and ξ^2. With those values in mind and a given set of covariates, and considering again the sample distributions, we now sample from σ_i^2 and μ_i.

Figure 10 shows the posterior predictive distribution of the SST on a grid 25×25 points that contains all the observational sites from the smallest value of latitude

Table 1: Mean predicted SST and MSE in the grid by type of device.

	bucket	eri	f.buoy	d.buoy
Mean	19.28	19.65	22.87	20.03
MSE	3.07	3.12	21.90	2.91

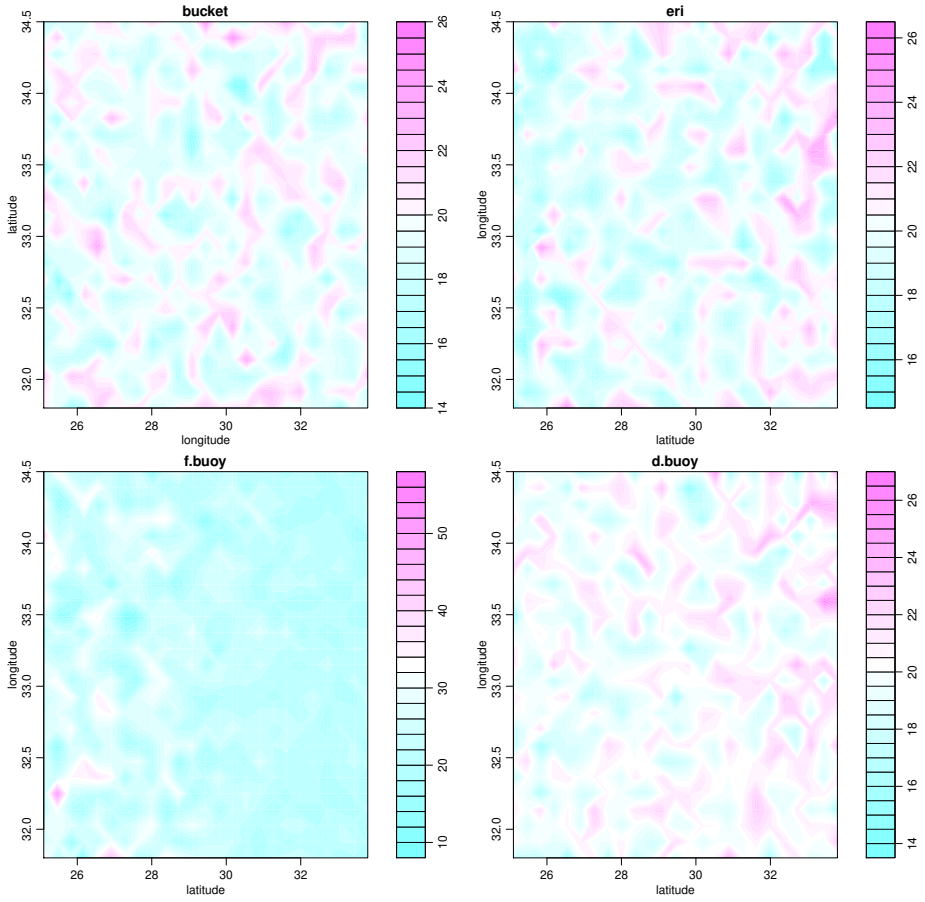

Figure 10: Posterior predictive distribution of the SST on a grid that contains all the observational sites, conditioning on each type device.

and longitude to the largest ones, conditioning on each type device. In addition, Table 1 presents the mean predicted SST and MSE in the grid by each type of device. In order to assess the quality of our predictions we consider to references: First, the mean SST of the 336 measurements reported by the 86 devices, which is 19.96°C; and second, the mean SST in the same grid reported by the *National Oceanic and Atmospheric Administration* (NOAA) in 2012[1], which is 22.44°C. According to this information, type of devices bucket, eri, and d.buoy (types with more devices spread along the grid) have a better performance in terms of prediction since all of them are reporting vales close to 19.96°C, and also in terms of uncertainty since the MSE associated with those prediction is small in comparison to the MSE associated with f.buoy. On the other hand, the prediction given by type of device f.buoy (with data from just 5 devices and therefore location in the

[1]Data available in `ftp://ftp.nodc.noaa.gov/pub/data.nodc/pathfinder/Version5.2/2012/`

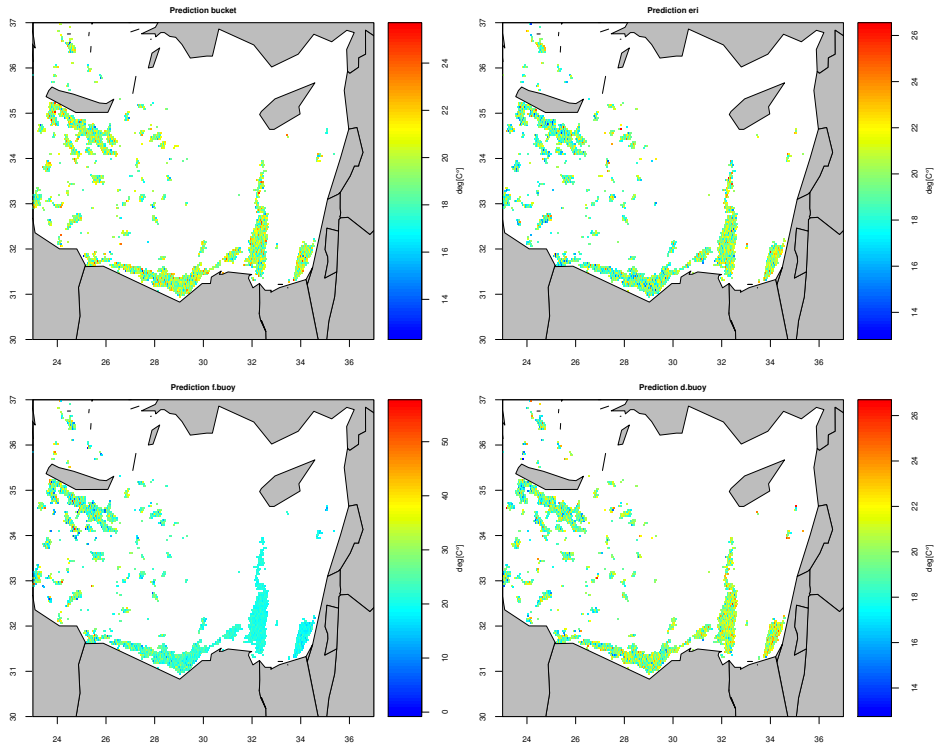

Figure 11: Posterior predictive distribution of the SST on a extended grid that contains all the observational sites, conditioning on each type device.

grid) is the closest one to the mean SST reported by the NOAA with drawback of a large MSE.

Finally, to explore the predictive capacity of the model and the data available, we extend the grid as follows: (1) latitude: from (31.8, 34.5) to (30.0 37.0); and longitude: from (25.1, 33.8) to (23.0, 37.0). This new grid has 336×168 points on it. Figure 12 shows the actual SST in part of that grid (we just have records for 5.2 % of them) according to the NOAA. Now, Figure 11 shows the posterior predictive distribution of the SST on the same grid, conditioning on each type device; in addition, Table 2 presents the mean predicted SST and MSE in the grid by each type of device. On part of this extended grid (see Figure 12), according to data of the NOAA in 2012, the mean SST is 20.15°C. We see how in this case

Table 2: Mean predicted SST and MSE in the extended grid by type of device.

	bucket	eri	f.buoy	d.buoy
mean	19.39	19.45	23.62	19.76
MSE	3.06	3.53	37.82	3.42

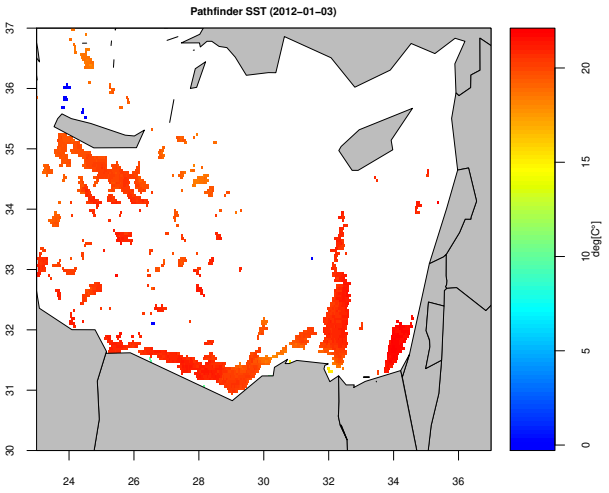

Figure 12: Mean SST on the extended grid according to the NOAA in 2012.

the best prediction is given by the d.buoy type of device in terms of the mean and the MSE. We confirm that the best predictions are given by the types of devices with more devices along the Mediterranean ocean (bucket, eri, and d.buoy), and how the the f.buoy is overestimating the mean SST (we suspect that the reason for that is because the devices of this type are mostly located at the north-east of the sea where sea temperature slightly increases in comparison with the other location where the other devices are placed).

4. Discussion

Hierarchical models provide a strong alternative to analyze complex and realistic settings. Their parameter flexibility allow us to describe many characteristics of a given data set that a regular single-level model does not provide. The ability to model within and between means and variances yields to better knowledge of the problem (even if we want to predict future values in any stage). For even more detail and deep kinds of complexity, this type of models give us many alternatives for generalizations (stages, hyperparameters, probability assumptions, or even covariates influence as in the model developed here). Given the computational tools and methods to get samples from the posterior distribution that are available (Gibbs sampling, direct sampling, Metropolis-Hastings sampling), hierarchical models are indeed a tool almost mandatory for any analyst.

For future analysis of the application, we can always thing in more complex models in which we take into account more factors to model the temperature, like interactions with other variables, say for example the interoceanic wind.

Statements and Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this art

Recibido:
Aceptado:

Referencias

D. Gamerman and H. F. Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC, 2006.

A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013. ISBN 9781439840955. URL `http://books.google.com/books?id=ZXL6AQAAQBAJ`.

P. Hoff. *A First Course in Bayesian Statistical Methods*. Springer Texts in Statistics. Springer, 2009. ISBN 9780387924076. URL `http://books.google.com/books?id=V8jT2SimGR0C`.

R. McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC, 2018.

N. Skliris, S. Sofianos, A. Gkanasos, A. Mantziafou, V. Vervatis, P. Axaopoulos, and A. Lascaratos. Decadal scale variability of sea surface temperature in the mediterranean sea in relation to atmospheric variability. *Ocean Dynamics*, '10.1007, September 2011.

A. Notation

The cardinality of a set A is denoted by $|A|$. If P is a logical proposition, then $\mathbf{1}\{\mathrm{P}\} = 1$ if P is true, and $\mathbf{1}\{\mathrm{P}\} = 0$ if P is false. $\lfloor x \rfloor$ denotes the floor of x, whereas $[n]$ denotes the set of all integers from 1 to n, i.e., $\{1, \ldots, n\}$. The Gamma function is given by $\Gamma(x) = \int_0^\infty u^{x-1}\, e^{-u}\, \mathrm{d}u$. Matrices and vectors with entries consisting of subscripted variables are denoted by a boldfaced version of the letter for that variable. For example, $\boldsymbol{x} = (x_1, \ldots, x_n)$ denotes an $n \times 1$ column vector with entries $x_1, \ldots, x_n$. We use $\mathbf{0}$ and $\mathbf{1}$ to denote the column vector with all entries equal to 0 and 1, respectively, and $\mathbf{I}$ to denote the identity matrix. A subindex in this context refers to the corresponding dimension; for instance, $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. The transpose of a vector $\boldsymbol{x}$ is denoted by $\boldsymbol{x}^\mathsf{T}$; analogously

for matrices. Moreover, if $\mathbf{X}$ is a square matrix, we use $\text{tr}(\mathbf{X})$ to denote its trace and $\mathbf{X}^{-1}$ to denote its inverse. The norm of $\boldsymbol{x}$, given by $\sqrt{\boldsymbol{x}^\mathsf{T}\boldsymbol{x}}$, is denoted by $\|\boldsymbol{x}\|$.

Now, we present the form of some standard probability distributions used in this article:

- Multivariate normal:

  A $d \times 1$ random vector $\boldsymbol{X} = (X_1 \ldots, X_d)$ has a multivariate Normal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, denoted by $\boldsymbol{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathsf{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if its density function is

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2} \, |\boldsymbol{\Sigma}|^{-1/2} \, \exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}.$$

- Gamma:

  A random variable X has a Gamma distribution with parameters $\alpha, \beta > 0$, denoted by $X \mid \alpha, \beta \sim \mathsf{G}(\alpha, \beta)$, if its density function is

$$p(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \, x^{\alpha-1} \, \exp\{-\beta x\}, \quad x > 0.$$

- Inverse Gamma:

  A random variable X has an Inverse Gamma distribution with parameters $\alpha, \beta > 0$, denoted by $X \mid \alpha, \beta \sim \mathsf{IG}(\alpha, \beta)$, if its density function is

$$p(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \, x^{-(\alpha+1)} \, \exp\{-\beta/x\}, \quad x > 0.$$