
Detecting influential observations in spatial
models using Bregman divergence

Ian Meneghel Danilevicz

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA UFSCar-USP

IAN MENEGHEL DANILEVICZ

**DETECTING INFLUENTIAL OBSERVATIONS IN SPATIAL MODELS USING BREGMAN
DIVERGENCE**

Master dissertation submitted to the Departamento de Estatística - DEs/UFSCar and the Instituto de Ciências Matemáticas e de Computação - ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Statistics.

Advisor: Prof. Dr. Ricardo Sandes Ehlers

**São Carlos
March 2018**

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA UFSCar-USP

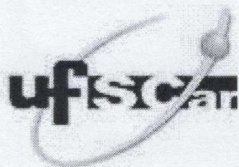
IAN MENEGHEL DANILEVICZ

**DETECÇÃO DE OBSERVAÇÕES INFLUENTES EM MODELOS ESPACIAIS USANDO
DIVERGÊNCIA DE BREGMAN**

Dissertação apresentada ao Departamento de Estatística – Des/UFSCar e ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Estatística - Programa Interinstitucional de Pós-Graduação em Estatística UFSCar-USP.

Orientador: Prof. Dr. Ricardo Sandes Ehlers

**São Carlos
Março de 2018**

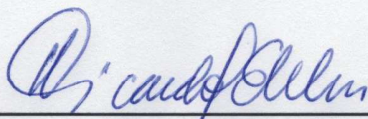


UNIVERSIDADE FEDERAL DE SÃO CARLOS

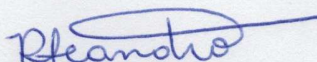
Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

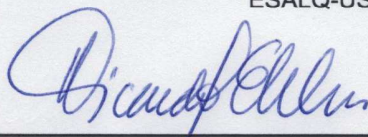
Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Ian Meneghel Danilevicz, realizada em 26/02/2018:



Prof. Dr. Ricardo Sandes Ehlers
USP

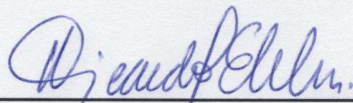


Profa. Dra. Roseli Aparecida Leandro
ESALQ-USP



Prof. Dr. Marcos Oliveira Prates
UFMG

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Marcos Oliveira Prates e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.



Prof. Dr. Ricardo Sandes Ehlers

*Para Cristiano Santos,
em memória.*

ACKNOWLEDGEMENTS

First of all, Temer out!

Once Mandela said that if you talk to a man in a language he understands, that goes to his head. However if you talk in his mother language, that goes to his heart. So my acknowledgments must be in Portuguese.

Gostaria de agradecer aos professores membros da banca por terem aceito o convite de participar desse importante momento da minha trajetória.

Agradeço ao professor Ricardo Ehlers por ter aceito me orientar antes mesmo de me conhecer pessoalmente. Pela sua confiança no meu trabalho, por dividir os seus contatos acadêmicos, por ter estado sempre presente.

À Capes pelo financiamento desses meus dois anos como estudante.

Aos professores Rafael Izbicki e Rafael Stern pelas ótimas aulas que tive com eles na UFSCar. Valeu a pena viajar até aqui para aprender com eles Inferência Bayesiana, Probabilidade e a debater.

Ao professor Gustavo, por ter sido uma pessoa atenta às dificuldades dos alunos recém chegados. O apoio dele foi vital para mim.

Aos meus colegas, em especial ao David, com quem tive a sorte de compartilhar orientação e, dessa forma, compartilhar ideias de HMC, DIC, WAIC entre outras technicalidades de estatísticos.

À Elisa Salengue, pelo artigo do Gelman que ela me presenteou além, é claro, do contato próximo que não perdemos mesmo com a distância.

Ao clã dos historiadores gaúchos que amo: Nathan, João Cé, Daniele Primavera, Balbinot, Ricardo, Raquel, Galo Místico, Gi, Diego, Lu Gomes, Gab Alemão, Pietro, Nanda, Luís Felipe, Si, Krishna e Shambhavi. Aos meus correspondentes internacionais, minha querida Mari, ao malandro do Tomas, o druida digital Silva e o extravagante Peruzzo.

Ao meu primeiro amigo em terras estrangeiras, que conheci no desgastante curso de Verão e com quem compartilhei otimismo e decepções, devaneios e choques de realidade, casca em chamuscas e almoços jambuzianescos. Sanca teria sido uma chatice não fosse meu primo das colinas, Hécio Martinho.

A cada rua desconhecida de uma nova cidade na qual decidimos entrar há mapas que se desdobram. Correr e pedalar me provoca isso. Sendo que pedalar com outras pessoas é como pegar emprestado os diários desses companheiros. Chloe, Paulo Mendes,

Marco Pollo, Luciana Irocssa e Caki são os ciclistas com quem Sanca retorceu-se e resignificou-se com capivaras, corsas e ursos (imaginários ou glauberescos).

Há um conto no qual uma moça do Kansas segue um caminho de tijolos amarelos. Existe uma garota cujos sapatos colorem os caminhos pelos quais ela passa, eu sei porque conheci Helena.

Agradeço aos meus pais. Diz Criolo, que sua mãe ao se mudar abandonou uma mala de roupas para não deixar os livros. Jamais esquecerei do Leitão Leitor, nem de Aquiles, pois estão entre as primeiras leituras que minha mãe me narrou. Igualmente impossível, é esquecer do dever de casa da primeira série: escrever uma fábula alternativa do Lobo mau.

Ao meu primeiro mestre por compartilhar sua sabedoria nos tatames. Antes de entrar na Universidade, eu secretamente prometi que meu mestrado seria dedicado a ele. Finalmente posso honrar minha palavra.

“An ounce of practice is worth a ton of theory.”
Swami Sivananda

ABSTRACT

DANILEVICZ, I. M. **Detecting influential observations in spatial models using Bregman divergence**. 2018. 80 p. Dissertation (Master in Statistics) – Centro de Ciências Exatas e Tecnologia, Universidade Federal de São Carlos, São Carlos – SP, 2018.

How to evaluate if a spatial model is well adjusted to a problem? How to know if it is the best model between the class of conditional autoregressive (CAR) and simultaneous autoregressive (SAR) models, including homoscedasticity and heteroscedasticity cases? To answer these questions inside Bayesian framework, we propose new ways to apply Bregman divergence, as well as recent information criteria as widely applicable information criterion (WAIC) and leave-one-out cross-validation (LOO).

The functional Bregman divergence is a generalized form of the well known Kullback-Leiber (KL) divergence. There is many special cases of it which might be used to identify influential points. All the posterior distributions displayed in this text were estimate by Hamiltonian Monte Carlo (HMC), a optimized version of Metropolis-Hasting algorithm. All ideas showed here were evaluate by both: simulation and real data.

Keywords: Bayesian inference, Bregman divergence, Hamiltonian Monte Carlo, influential points, spatial models, heteroscedasticity.

RESUMO

DANILEVICZ, I. M. Detecção de observações influentes em modelos espaciais usando divergência de Bregman. 2018. 80 p. Dissertação (Mestrado em Estatística) – Centro de Ciências Exatas e Tecnologia, Universidade Federal de São Carlos, São Carlos – SP, 2018.

Como avaliar se um modelo espacial está bem ajustado? Como escolher o melhor modelo entre muitos da classe autorregressivo condicional (CAR) e autorregressivo simultâneo (SAR), homoscedásticos e heteroscedásticos? Para responder essas perguntas dentro do paradigma bayesiano, propomos novas formas de aplicar a divergência de Bregman, assim como critérios de informação bastante recentes na literatura, são eles o *widely applicable information criterion* (WAIC) e validação cruzada *leave-one-out* (LOO).

O funcional de Bregman é uma generalização da famosa divergência de Kullback-Leiber (KL). Há diversos casos particulares dela que podem ser usados para identificar pontos influentes. Todas as distribuições a posteriori apresentadas nesta dissertação foram estimadas usando Monte Carlo Hamiltoniano (HMC), uma versão otimizada do algoritmo Metropolis-Hastings. Todas as ideias apresentadas neste texto foram submetidas a simulações e aplicadas em dados reais.

Palavras-chave: Inferência Bayesiana, divergência de Bregman, Monte Carlo Hamiltoniano, pontos influentes, modelos espaciais, heteroscedasticidade.

LIST OF FIGURES

Figure 1 – Example of strictly convex and identity functions	41
Figure 2 – Model comparison by WAIC and LOO; at the left the true model is a homoscedastic SAR and the proposed models are: I the corrected, II with a wrong covariate, III a heteroscedastic SAR, IV a CAR, V a linear model; at the right the true model is a homoscedastic CAR and the proposed models are: I the corrected, II with a wrong covariate, III a heteroscedastic CAR, IV a SAR, V a linear model.	51
Figure 3 – Kullback-Leiber divergence for simulated data, at the left the true model is a homoscedastic SAR and the proposed models are: I the corrected in black square, II with a wrong covariate in red circle, III a heteroscedastic SAR in green up triangle, IV a CAR in blue rhombus, V a LM in pink down triangle; at the right the true model is a homoscedastic CAR and the proposed models are: I the corrected in black square, II with a wrong covariate in red circle, III a heteroscedastic CAR in green up triangle, IV a SAR in blue rhombus, V a LM in pink down triangle. . .	52
Figure 4 – Normalizing Bregman divergence for simulated homoscedastic SAR data	53
Figure 5 – Normalizing Bregman divergence for simulated SAR data, where black rhombus is no perturbation set, red circle is Perturbation II, green triangle is III, blue square is IV. Solid colors were used to indicate true influential point.	54
Figure 6 – Normalizing Bregman divergence for simulated heteroscedastic SAR data	55
Figure 7 – Normalizing Bregman divergence for simulated homoscedastic CAR data	55
Figure 8 – Normalizing Bregman divergence for simulated CAR data, where black rhombus is no perturbation set, red circle is Perturbation II, green triangle is III, blue square is IV. Solid colors were used to indicate true influential point.	56
Figure 9 – Normalizing Bregman divergence for simulated heteroscedastic CAR data	56
Figure 10 – Brazil 2013, pesticide per cultivated area in kg/hectare	60
Figure 11 – Brazil 2013, comparison of models, where I is a homoscedastic SAR, II is a heteroscedastic SAR, III is a homoscedastic CAR, IV is a heteroscedastic CAR, V is a linear regression.	60
Figure 12 – Brazil 2013, Kullback-Leiber and normalizing Bregman, where black square is a homoscedastic SAR, green triangle is a homoscedastic CAR.	61
Figure 13 – Brazil 2013, Normalizing Bregman divergence.	61
Figure 14 – Europe 2014, average weekly hours of work.	62

Figure 15 – Europe 2014, comparison of models, where I is a homoscedastic SAR, II is a heteroscedastic SAR, III is a homoscedastic CAR, IV is a heteroscedastic CAR, V is a linear regression.	63
Figure 16 – Europe 2014, comparison of models, where black square is a heteroscedastic CAR with GPD, red circle is a homoscedastic CAR with study gender ratio green triangle is a heteroscedastic SAR with GPD.	64
Figure 17 – Europe 2014, Normalizing Bregman.	65

LIST OF TABLES

Table 1 – Brazil 2013, Indexes of Sustainable Development	23
Table 2 – Analysis of Sensitivity to 500 simulated homoscedastic SAR models with $\phi \in \{0.7, 0.9\}$	48
Table 3 – Analysis of Sensitivity to 500 simulated heteroscedastic SAR models with $\phi \in \{0.7, 0.9\}$	48
Table 4 – Analysis of Sensitivity to 500 simulated homoscedastic CAR models with $\phi \in \{0.7, 0.9\}$	49
Table 5 – Analysis of Sensitivity to 500 simulated heteroscedastic CAR models with $\phi \in \{0.7, 0.9\}$	50
Table 6 – LOO mean and (standard deviation) to 500 simulated cases of each model, variance type, prior and value of $\phi \in \{0.7, 0.9\}$	50
Table 7 – Brazil 2013, parameter estimation by model	62
Table 8 – Europe 2014, parameter estimation by model	65

LIST OF ABBREVIATIONS AND ACRONYMS

AC	Acre
AL	Alagoas
AM	Amazonas
AP	Amapá
BA	Bahia
CAR	conditional autoregressive
DIC	deviance information criterion
ELPD	expected log predictive density for a new data point
ELPPD	expected log pointwise predictive density for a new dataset
GPD	gross domestic product
HMC	Hamiltonian Monte Carlo
IBGE	Instituto Brasileiro de Geografia e Estatística
ICMC	Instituto de Ciências Matemáticas e de Computação
ISCED	international standard classification of education
ISD	Indexes of Sustainable Development
LOO	leave-one-out cross-validation
LPPD	log pointwise predictive density
KL	Kullback-Leibler
MCMC	Markov chain Monte Carlo
NATO	North Atlantic Treaty Organization
PPS	purchasing power standards
SAR	simultaneous autoregressive
SDC	Sustainable Development Commission
UFSCar	Universidade de São Carlos

UN	United Nations
UNCED	United Nations Conference on Environment and Development
USP	Universidade de São Paulo
WAIC	widely applicable information criterion

CONTENTS

1	INTRODUCTION	21
2	MODELS AND METHODS	23
2.1	The CAR Model	23
2.1.1	Choice of Priors for homoscedastic CAR	25
2.1.2	Choice of Priors for heteroscedastic CAR	26
2.2	The SAR Model	27
2.2.1	Choice of Priors for homoscedastic SAR	28
2.2.2	Choice of Priors for heteroscedastic SAR	29
2.3	Hamiltonian Monte Carlo	30
3	MODEL SELECTION AND DIAGNOSTIC ANALYSIS	35
3.1	Information Criteria	35
3.1.1	DIC	35
3.1.2	WAIC	36
3.1.3	LOO	37
3.1.4	Standard errors	38
3.2	Bayesian model diagnostics	38
3.2.1	Influential points	38
3.2.2	Functional Bregman divergence	39
3.2.3	Perturbation on dependent models	41
3.2.4	Posterior comparison using Bregman divergence	43
3.2.5	Normalizing Bregman divergence	44
4	SIMULATION	47
4.1	Analysis of Sensitivity	47
4.2	Model selection and misspecification	50
4.3	Perturbation of Likelihood and Bregman divergence	52
5	APPLICATION	59
5.1	Indexes of Sustainable Development in Brazil	59
5.2	Working conditions in Europe Union	62
6	CONCLUSION	67
	BIBLIOGRAPHY	69

APPENDIX	73
APPENDIX A – CONDITIONAL POSTERIOR DISTRIBUTIONS	75

1 INTRODUCTION

Time series and spatial models share the same restriction compared to traditional regression models, their data are dependent. The absence of the assumption of independence implies into theoretical and practical consequences to fit models and to proceed with residual diagnostics, even more in a Bayesian environment. To proceed the analysis of diagnostics, we must detect influential points, which could be defined as a set of observations which modify the parameter estimation. There is a previous literature which proposed to use the Bregman divergence to detect influential observation if they are independent and identically distributed, (GOH; DEY, 2014). We extend their ideas to dependent samples. Furthermore, we suggest to resize the Bregman divergence to a scale between zero and one.

There is a vast literature of temporal and spatial models to study, but to illustrate this procedure, we choose the simultaneous autoregressive (SAR) and conditional autoregressive (CAR) models to analyze. SAR is a model where a spatial structure is inserted within the mean. On the other hand, CAR inserts the spatial pattern in the variance. Both models are largely used in applied statistics as economics, social and environmental analysis. They could be developed to homoscedastic or heteroscedastic cases. SAR and CAR models can be estimated by both schools: classical and Bayesian. We choose the second theoretical framework to follow and, consequently, we must simulated the posterior distribution, because there is no closed analytical form. So it is necessary to resort to Markov chain Monte Carlo (MCMC), Gibbs Sampling or Hamiltonian Monte Carlo (HMC), we adopt the last option and encourage its use.

There are many ways to compare models in a Bayesian framework from Bayes factor to reversible jump MCMC. Even though the information criteria class grew up and there are new and interesting criteria available as widely applicable information criterion (WAIC) and leave-one-out information criterion (LOO). Both criteria generalize the well known deviance information criterion (DIC). Even before model selection, we must define the prior of the models. We also used the LOO as tool to proceed with analysis of sensitivity as well as the traditional way to study bias and variance of each parameter.

The remainder of the dissertation is structured as follows. In chapter 2, we describe the spatial models which we shall use: their likelihood, priors and posteriors. Though, a briefly discussion about HMC, its advantages and computational costs. In the chapter 3 we keep with a theoretical discussion. We present some recent information criteria and introduce the functional Bregman divergence, mainly to realize how to take advantage of it to proceed with Bayesian analysis of diagnostics. This divergence has some difficulties which require simulation to be applicable. All the above propositions were study using simulated

and real data. Chapter 5 is dedicated to test models and techniques by simulation study. Analysis of sensitivity, evaluate models and diagnostics were first check in artificial and controlled data. Only in chapter 6 we try or methods in real application. The applied data correspond to Indexes of Sustainable Development in Brazil and social surveys realized in European Union. We close the discussion in Chapter 6 and there is an appendix, where the reader might check the model posteriors step by step.

2 MODELS AND METHODS

In this chapter we shall present two canonical spatial models: CAR and SAR. Their parameters, likelihood, prior and posterior distributions. The first section is dedicated to discuss about CAR model, which is more frequently applied by statisticians. The second section is reserved to SAR model, which is more commonly used by economists. Neither of these models have a posterior in closed form, consequently, simulation might be use to find an approached parameter distribution. So, the last section is dedicated to discuss about the Hamiltonian Monte Carlo (HMC), a simulation method which we elected to solve our posterior distributions.

Table 1: Brazil 2013, Indexes of Sustainable Development

State	Agrotoxic ¹	GDP per capta	council ²	legislation ³
AC	2.6	14.733	45.5	81.8
AL	2.8	11.277	10.8	40.2
AP	1.9	17.363	75.0	100.0
AM	0.7	21.874	54.8	72.6
BA	5.9	13.578	55.6	68.6

¹ Measured in kg per cultivated hectare.

² % of counties with environmental council.

³ % of counties with environmental legislation.

Before examining CAR and SAR, let us look at a motivator example. Table 1 displays a sample of Brazilian States and some Indexes of Sustainable Development (ISD). Whether we wish to estimate the level of agrotoxic per cultivated area in a missing State, we can proceed with a linear regression with three predictor variables: GDP per capta, council and legislation. However, if there is spatial information available, i.e., if we know which States sharing boundary then we might build a graph, where neighbor States are connected points, and incorporate this information into analysis. Both CAR and SAR used this graph in their likelihoods as we shall see later.

2.1 The CAR Model

The conditional autoregressive (CAR) model, introduced by (BESAG, 1974), have received the major attention from statistician than SAR, by the reason it is more malleable. CAR model with independent covariates is frequently expressed as in expression 2.1 (OLIVEIRA, 2012):

$$y_i | \mathbf{y}_{(i)} \sim N \left(\mathbf{x}'_i \boldsymbol{\beta} + \sum_{j=1}^n \phi a_{ij} (y_j - \mathbf{x}'_j \boldsymbol{\beta}), \sigma_i^2 \right), \quad (2.1)$$

where $\mathbf{y}'_{(i)} = (y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$ is the response variable, $\mathbf{x}' = (x_1, \dots, x_k)$ is the vector of k covariates and the respectively linear coefficients $\boldsymbol{\beta}' = (\beta_1, \dots, \beta_k)$. The covariance structure is contemplated by $\phi a_{ij} \geq 0$ and $\sigma_i^2 \geq 0, \forall i$. The ϕ is the spatial correlation parameter and a_{ij} shall be explained at Definition 1.

The model could be vector represent as in expression 2.2.

$$\mathbf{y} \sim N(X\boldsymbol{\beta}, (I_n - \phi A)^{-1}M), \quad (2.2)$$

where M is a diagonal matrix responsible by the pure variance. The square matrix A is the adjacent matrix, which summarize the spatial connections between the observations, i.e., it describes the graph suggested in the motivator example. Here we used the terms "adjacent", "neighbor" or "near" to indicate two observations share some spatial connection, which depends of the research context.

Definition 1 *A symmetric matrix A is called an adjacent matrix if each element $a_{ij} = 1$ if positions i and j are spatial connected and $a_{ij} = 0$ otherwise. The diagonal is define as zero, $a_{ii} = 0 \forall i$,*

$$A_{n \times n} = \begin{bmatrix} 0 & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & 0 & \dots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \dots & 0 \end{bmatrix}. \quad (2.3)$$

There are two restrictions to ϕA and M as discussed by (CRESSIE, 1993) and (RUE; HELD, 2005):

1. $\phi M^{-1}A$ must be symmetric;
2. $M^{-1}(I_n - \phi A)$ must be positive defined.

The CAR could be homoscedastic as well as heteroscedastic. These conditions depend of the matrix M , which could be as the next two items, as previously showed by (CRESSIE; PERRIN; THOMAS-AGNAN, 2005) and (OLIVEIRA, 2012):

1. Whether $M = \sigma^2 I_n$, the CAR model owns homoscedasticity;
2. On the other hand, if $M = (\sigma_1^2, \dots, \sigma_n^2)' I_n$, it presents heteroscedasticity.

Thus, if we assume $M = \sigma^2 I_n$ and $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \phi, \sigma^2\}$, CAR has a likelihood as in equation 2.4.

$$f(\mathbf{y}|\boldsymbol{\theta}) = (2\pi\sigma^2)^{-n/2} |I_n - \phi A|^{1/2} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta})'(I_n - \phi A)(\mathbf{y} - X\boldsymbol{\beta}) \right\}, \quad (2.4)$$

although, a more complex frame comes with $M = (\sigma_1^2, \dots, \sigma_n^2)' I_n$. For simplicity, we shall refer to $(\sigma_1^2, \dots, \sigma_n^2)' I_n$ as Σ and $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \phi, \Sigma\}$. So it owns a heteroscedastic likelihood as in equation 2.5,

$$f(\mathbf{y}|\boldsymbol{\theta}) = (2\pi)^{-n/2} |I_n - \phi A|^{1/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{y} - X\boldsymbol{\beta})'(I_n - \phi A)\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta}) \right\}. \quad (2.5)$$

2.1.1 Choice of Priors for homoscedastic CAR

First, we shall solve a homoscedastic problem and so extended it to heteroscedastic case. We assigned the following prior distributions for the parameters and assume that they are independent.

$$\begin{aligned} \phi &\sim \text{Uniform}(0, 1) \\ \boldsymbol{\beta} &\sim N(0, \eta I_k) \\ \sigma^2 &\sim \text{IG}(a, b), \end{aligned} \quad (2.6)$$

where I_k is the identity matrix of dimension $k \times k$, so for a large η , this is a poor informative prior. Although a Beta prior for ϕ was suggested by (BANERJEE; CARLIN; GELFAND, 2015), we follow (OLIVEIRA, 2012) who defends there is no reason to increase the complexity in this parameter. Finally the prior of σ^2 as an inverse Gamma is interesting for pragmatical purpose. However, a good elicitation of the hyperparameters a , shape, and b , scale, might be not so simple. If the hyperparameter $a > 2$, we might guarantee that the prior $\pi(\sigma^2)$ is proper and, consequently, the posterior is proper too. However, even if $0 < a \leq 2$, there is low possibility that the posterior is improper, (OLIVEIRA, 2012).

Using Bayes's theorem and independence of priors, we have the joint posterior density,

$$p(\phi, \boldsymbol{\beta}, \sigma^2 | \mathbf{y}) \propto (\sigma^2)^{-(a+1+n/2)} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta})'(I_n - \phi A)(\mathbf{y} - X\boldsymbol{\beta}) - \frac{1}{2\eta} \boldsymbol{\beta}'\boldsymbol{\beta} - \frac{2b}{2\sigma^2} \right\}. \quad (2.7)$$

The solution step by step of all posterior distributions showed in this work are available in Appendix. Equation 2.7 has no closed form, but can be solve by Hamiltonian Monte Carlo (HMC). We also can derive the conditional posterior distribution of each specific parameter. To begin with β as in expression 2.8, which gives a multivariate Normal,

$$\beta|\mathbf{y} \sim N_k(0, \eta\sigma^2 [\eta X'(I_n - \phi A)X + \sigma^2 I_n]^{-1}). \quad (2.8)$$

In the same way, we may derive the conditional posterior of ϕ , which is a truncate Exponential distribution between zero and one as in 2.9,

$$\phi|\mathbf{y} \sim Exp\left(\frac{1}{2\sigma^2}(\mathbf{y}'A\mathbf{y} + (X\beta)'AX\beta)\right) I_{\phi \in [0,1]}. \quad (2.9)$$

Finally, we might derive the conditional posterior of σ^2 , which correspond to a inverse Gamma with hyperparameter a and b as exhibit in expression 2.10,

$$\sigma^2|\mathbf{y} \sim IG\left(a + \frac{n}{2}, \frac{1}{2}(\mathbf{y} - X\beta)'(I_n - \phi A)(\mathbf{y} - X\beta) + b\right). \quad (2.10)$$

2.1.2 Choice of Priors for heteroscedastic CAR

Whether we cannot assume homoscedasticity, we must alter the variance's prior from an inverse Gamma distribution to a product of inverse Gammas as in 2.11.

$$\sigma_i^2 \sim IG(a_i, b_i) \forall i \in \{1, \dots, n\}. \quad (2.11)$$

Consequently, the posterior might be changed as in equation 2.12,

$$p(\phi, \beta, \Sigma|\mathbf{y}) \propto |\Sigma|^{-1/2} \prod_{i=1}^n ((\sigma_i^2)^{-(a_i+1)}) \exp\left\{-\frac{1}{2}(\mathbf{y} - X\beta)'(I_n - \phi A)\Sigma^{-1}(\mathbf{y} - X\beta) - \frac{1}{2\eta}\beta'\beta - \sum_{i=1}^n \frac{b_i}{\sigma_i^2}\right\}, \quad (2.12)$$

where $\Sigma = (\sigma_1^2, \dots, \sigma_n^2)'I_n$ is a diagonal heteroscedastic variance matrix. Again, we discuss each one of the conditional posteriors. First, the referent to β in 2.13,

$$\beta|\mathbf{y} \sim N_k\left(0, \eta[\eta X'(I_n - \phi A)\Sigma^{-1}X + I_n]^{-1}\right), \quad (2.13)$$

which is clearly a k multivariate Normal, because X' has dimension $k \times n$. The spatial parameter ϕ is checked in expression 2.14, a truncate Exponential distribution,

$$\phi|\mathbf{y} \sim \text{Exp} \left(\frac{1}{2}(\mathbf{y}'A\Sigma^{-1}\mathbf{y} + (X\boldsymbol{\beta})'A\Sigma^{-1}(X\boldsymbol{\beta})) \right) I_{\phi \in [0,1]}. \quad (2.14)$$

For the conditional distribution of variance we must used some different strategy because working with matrix operation would be quite hard in this case. We remind that σ_i^2 and σ_j^2 are independent for all $i \neq j$. So, without loss of generality consider any σ_i^2 and its respective prior a_i and b_i as we developed in expression 2.15,

$$\sigma_i^2|\mathbf{y} \sim IG \left(a_i + 1/2, \frac{1}{2}[(\mathbf{y} - X\boldsymbol{\beta})'(I_n - \phi A)(\mathbf{y} - X\boldsymbol{\beta}) + 2b_i] \right), \quad (2.15)$$

where all σ_j^2 , a_j and b_j for all $j \neq i$ are constants.

2.2 The SAR Model

The simultaneous autoregressive (SAR) models have become widely used by economists (WANG; LUNG-FEI, 2013), and they present a relevant characteristic: additive errors, this implies in easier simulation and estimation. SAR is simpler than CAR and consequently it is more parsimonious, an important quality if or data set is not abundant.

A SAR model with independent covariates could be expressed as in equation 2.16.

$$\mathbf{y} = \phi W\mathbf{y} + X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (2.16)$$

where $\mathbf{y} = (y_1 \ y_2 \ \dots \ y_n)'$ is a n vector of the outcomes, X denotes an $n \times k$ matrix of covariates, i.e., known constants, $\boldsymbol{\beta} = (\beta_1 \ \beta_2 \ \dots \ \beta_k)'$ is a k vector of linear regression coefficients, $\boldsymbol{\epsilon} = (\epsilon_1 \ \epsilon_2 \ \dots \ \epsilon_n)'$ is a n vector of errors, ϕ is the coefficient of spatial effects on \mathbf{y} , and W represents a spatial weight matrix as clarified by Definition 2.

Definition 2 A square matrix W is called a weight matrix if it is the result of an adjacent matrix A , where each row is divided by the row sum, r_i , i.e., $r_i = \sum_{j=1}^n A_{i,j}$. Then $w_{ij} = a_{ij}/r_i$.

$$W_{n \times n} = \begin{bmatrix} 0 & a_{1,2}/r_1 & \dots & a_{1,n}/r_1 \\ a_{2,1}/r_2 & 0 & \dots & a_{2,n}/r_2 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1}/r_n & a_{n,2}/r_n & \dots & 0 \end{bmatrix}, \quad (2.17)$$

therefore it is guaranteed that $\sum_{j=1}^n W_{i,j} = 1 \ \forall i$.

The errors are initially assumed to be independently and normally distributed with mean zero and common variance σ^2 . This is the homoscedastic case where $\epsilon \sim N(0, \sigma^2 I_n)$ and I_n is the identity matrix of dimension n . By using the properties of the multivariate normal distribution, it follows that the likelihood function is given by equation 2.18,

$$f(\mathbf{y}|\boldsymbol{\theta}) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta} - \phi W\mathbf{y})' (\mathbf{y} - X\boldsymbol{\beta} - \phi W\mathbf{y}) \right\}, \quad (2.18)$$

where $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \phi, \sigma^2\}$. We could generalize the model to heteroscedastic context, where $\epsilon \sim N(0, \Sigma)$. Consequently, the likelihood would be as equation 2.19,

$$f(\mathbf{y}|\boldsymbol{\theta}) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{y} - X\boldsymbol{\beta} - \phi W\mathbf{y})' \Sigma^{-1} (\mathbf{y} - X\boldsymbol{\beta} - \phi W\mathbf{y}) \right\}, \quad (2.19)$$

where $\Sigma = (\sigma_1^2, \dots, \sigma_n^2) I_n$.

2.2.1 Choice of Priors for homoscedastic SAR

The SAR model has similar parameters with CAR, so is not extraordinary we decide for resembling prior,

$$\begin{aligned} \phi &\sim \text{Uniform}(0, 1) \\ \boldsymbol{\beta} &\sim N(0, \eta I_k) \\ \sigma^2 &\sim \text{IG}(a, b). \end{aligned} \quad (2.20)$$

Using Bayes's theorem and independence of priors, we have the joint posterior density,

$$p(\phi, \boldsymbol{\beta}, \sigma^2 | \mathbf{y}) \propto \frac{1}{(\sigma^2)^{(a+1)+n/2}} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta} - \phi W\mathbf{y})' (\mathbf{y} - X\boldsymbol{\beta} - \phi W\mathbf{y}) - \frac{1}{2\eta} \boldsymbol{\beta}' \boldsymbol{\beta} - \frac{2b}{2\sigma^2} \right\}. \quad (2.21)$$

Expression 2.21 has no close form, but could be adjusted by HMC. We also can derive the conditional posterior distribution of each specific parameter, just begin with $\boldsymbol{\beta}$ as in expression 2.22,

$$\boldsymbol{\beta} | \mathbf{y} \sim N_k((\mathbf{y}' X - \phi \mathbf{y}' W X)(\eta X' X + \sigma^2 I_k)^{-1}, \sigma^2 \eta (\eta X' X + \sigma^2 I_k)^{-1}). \quad (2.22)$$

In the same way, we may derive the conditional posterior of ϕ as in 2.23,

$$\phi|\mathbf{y} \sim N\left(\frac{\mathbf{y}'W\mathbf{y}}{(W\mathbf{y})'(W\mathbf{y})} - \frac{(X\boldsymbol{\beta})'W\mathbf{y}}{(W\mathbf{y})'(W\mathbf{y})}, \frac{\sigma^2}{(W\mathbf{y})'(W\mathbf{y})}\right) I_{\phi \in [0,1]}. \quad (2.23)$$

Where $I_{\phi \in [0,1]}$ is the indicator of truncation for ϕ inside the real interval from zero to one, which result into a truncate Normal distribution for ϕ . Finally, we might derive the conditional posterior of σ^2 as in 2.24. The variance follows an inverse Gamma, a similar result with the CAR model,

$$\sigma^2|\mathbf{y} \sim IG\left(a + \frac{n}{2}, \frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta} - \phi W\mathbf{y})'(\mathbf{y} - X\boldsymbol{\beta} - \phi W\mathbf{y}) + b\right). \quad (2.24)$$

2.2.2 Choice of Priors for heteroscedastic SAR

When we cannot assume homoscedasticity, we must alter the variance's prior from an inverse Gamma distribution to a product of inverse Gammas as in 2.25.

$$\sigma_i^2 \sim IG(a_i, b_i), \quad \forall i \in \{1, \dots, n\}. \quad (2.25)$$

Consequently, the posterior might be changed as in 2.26,

$$p(\phi, \boldsymbol{\beta}, \Sigma|\mathbf{y}) \propto |\Sigma|^{-1/2} \prod_{i=1}^n ((\sigma_i^2)^{-(a_i+1)}) \exp\left\{-\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta} - \phi W\mathbf{y})'\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta} - \phi W\mathbf{y}) - \frac{1}{2\eta}\boldsymbol{\beta}'\boldsymbol{\beta} - \sum_{i=1}^n \frac{b_i}{\sigma_i^2}\right\}. \quad (2.26)$$

Equation 2.26 has no closed form. Although, the conditional posteriors display a more treatable shape, $\boldsymbol{\beta}|\mathbf{y}$ follows a multivariate Normal as described in 2.27.

$$\boldsymbol{\beta}|\mathbf{y} \sim N_k\left([\mathbf{y}'\Sigma^{-1}X + \phi(W\mathbf{y})'\Sigma^{-1}X][X'\Sigma^{-1}X + \eta^{-1}I_k]^{-1}, [X'\Sigma^{-1}X + \eta^{-1}I_k]^{-1}\right). \quad (2.27)$$

So the conditional of $\phi|\mathbf{y}$ still a truncate Normal between zero and one. See expression 2.28,

$$\phi|\mathbf{y} \sim N\left(\frac{\mathbf{y}'\Sigma^{-1}W\mathbf{y}}{(W\mathbf{y})'\Sigma^{-1}W\mathbf{y}} - \frac{(W\mathbf{y})'\Sigma^{-1}X\boldsymbol{\beta}}{2(W\mathbf{y})'\Sigma^{-1}W\mathbf{y}}, ((W\mathbf{y})'\Sigma^{-1}W\mathbf{y})^{-1}\right) I_{\phi \in [0,1]}. \quad (2.28)$$

For the conditional distribution of variance we must use the same strategy applied in the heteroscedastic CAR model. We remind that σ_i^2 and σ_j^2 are independent for all $i \neq j$. So, without loss of generality consider any σ_i^2 and its respective prior a_i and b_i as we developed in expression 2.29,

$$\sigma_i^2 | \mathbf{y} \sim IG \left(a_i + \frac{1}{2}, \frac{1}{2} [(\mathbf{y} - X\boldsymbol{\beta} - \phi W \mathbf{y})'(\mathbf{y} - X\boldsymbol{\beta} - \phi W \mathbf{y}) + 2b_i] \right). \quad (2.29)$$

where all σ_j^2 , a_j and b_j for all $j \neq i$ are constants.

2.3 Hamiltonian Monte Carlo

In the last section we define the posterior distribution of SAR and CAR models. Though we must adjust them, because there are no closed forms for them. The most obvious strategy is to simulate by random walk Gibbs sampling, since all the conditional posteriors are already defined. However, the elevated rate of intrinsic autocorrelation within Markov chains implies into a high tax of sampling discard to fix the dependence between each iteration. Consequently, this methodology demands a waste of time and computational effort.

A representative mixing of the parametric space Θ is indispensable for all ensuing Bayesian analysis. This epistemological reason and the respect to computational time justify the seeking for efficient methods of simulation. The Hamiltonian Monte Carlo (HMC) comes as a recent and powerful simulation technique whether all the parameters of interest are continuous. What is exactly our case in both models: SAR and CAR. The HMC could be seen as a kind of Metropolis-Hastings, (NEAL, 2011), which is much more known and used. We show the symmetric Metropolis-Hastings algorithm, (GELMAN et al., 2003), and its three basic steps:

1. Draw a starting point $\boldsymbol{\theta}^0$, sample a proposal $\boldsymbol{\theta}^*$ vector for the parameter at the time t ;
2. determine the ratio r between densities,

$$r = \frac{p(\boldsymbol{\theta}^* | \mathbf{y})}{p(\boldsymbol{\theta}^{t-1} | \mathbf{y})}; \quad (2.30)$$

3. set

$$\boldsymbol{\theta}^t = \left\{ \begin{array}{ll} \boldsymbol{\theta}^* & , \text{ with probability } \min(r, 1) \\ \boldsymbol{\theta}^{t-1} & , \text{ otherwise} \end{array} \right\}. \quad (2.31)$$

HMC consist of take advantage of the gradient of the posterior instead of a naive random walk at the time of roam the parametric space Θ , i.e., at the first step we propose a smart θ^* . Using the gradient of the log probability function to increase convergence time and to find the stationary distribution, (NEAL, 2011).

The trick of HMC is to develop a Metropolis algorithm where the invariant distribution is not the posterior, $p(\theta|\mathbf{y})$, but a Hamiltonian density, $p(H(\theta, \varphi))$, which is a mix of the log posterior, $U(\theta)$, and an artificial, $K(\varphi)$, distribution. The density of $H(\cdot)$ is associate with an energy equation and could be simplify to

$$p(H(\theta, \varphi)) = \frac{1}{c} \exp(-H(\theta, \varphi)), \quad (2.32)$$

where c is just a normalization constant. We may decompose the Hamiltonian density into two independent pieces, i.e., $H(\theta, \varphi) = U(\theta) + K(\varphi)$, where $\theta = \{\theta_1, \dots, \theta_d\}$ are the "position" or "kinetic" variables and $\varphi = \{\varphi_1, \dots, \varphi_d\}$ are the "momentum" or "potential" variables in physics literature.

Let us consider the "kinetic" variables as minus log prior and likelihood, $U(\theta) = -\log(\pi(\theta)f(\mathbf{y}|\theta))$, and the "potential" variables as artificial random variables, commonly, log of a multivariate normal without correlation and zero mean, (NEAL, 2011).

$$K(\varphi) = \sum_{i=1}^d \frac{\varphi_i^2}{2v_i}, \quad (2.33)$$

where v_i is a variance hyperparameter. Until this point, we just increase the complexity of the problem without give any advantage, we shall proceed with the leapfrog technique to finally understand the point. One leapfrog iteration consist of:

$$\varphi^{(t)} \sim N_d(0, V) \quad (2.34a)$$

$$\varphi^{(t+\epsilon/2)} = \varphi^{(t)} - \frac{\epsilon}{2} \nabla U(\theta^{(t)}) \quad (2.34b)$$

$$\theta^{(t+\epsilon)} = \theta^{(t)} + \epsilon(\varphi^{(t+\epsilon/2)} V^{-1}) \quad (2.34c)$$

$$\varphi^{(t+\epsilon)} = \varphi^{(t+\epsilon/2)} - \frac{\epsilon}{2} \nabla U(\theta^{(t+\epsilon)}), \quad (2.34d)$$

where t indicates the time, $V = \text{diag}(v_1, \dots, v_d)$ is the diagonal matrix of variance, ϵ is a constant called step size which define the time discretization. $\nabla U(\theta)$ is the gradient of minus logarithm prior times likelihood. As the multivariate normal in 2.34a fills all the domain and is symmetric, we could see the leapfrog as a valid propose to Metropolis algorithm, (METROPOLIS et al., 1953) and (HASTINGS, 1970), which consist of a sum

of random and deterministic components. Though, the step 2 is subtly different, i.e., r would be update to

$$r = \frac{p(H(\boldsymbol{\theta}^{(t+\epsilon)}, \boldsymbol{\varphi}^{(t+\epsilon)})}{p(H(\boldsymbol{\theta}^{(t)}, \boldsymbol{\varphi}^{(t)})}, \quad (2.35)$$

where t is the current time and $\boldsymbol{\theta}^{(t+\epsilon)}, \boldsymbol{\varphi}^{(t+\epsilon)}$ are the new propose. The probability of acceptance a new propose is again equal to $\min(r, 1)$. Furthermore, instead of run just one leapfrog we might do L steps before proceed with judgment of acceptance. Consequently, we may run a time length of $L \times \epsilon$. For example, if L is equal to 2, the algorithm in 2.34 will be replicated at steps 2.34b and 2.34c.

In our discussion it should be evident that Hamiltonian Monte Carlo involves three tuning arguments or, in Bayesian words, three hyperparameters. They are the L number of leapfrogs by iteration, the ϵ step size length and the initial distribution of the "momentum" variable $\boldsymbol{\varphi}$. To choice an appropriated number of L , which associated with ϵ will not produce a constant periodicity may be done using the No-U-Turn (NUTS) sampler, (HOFFMAN; GELMAN, 2014).

NUTS is an algorithm created to avoid the need to hand-tune L and ϵ . During the warmup the algorithm will test different values of leapfrogs and step size and automatically judges the best range to sample.

The basic strategy is to double L until increase the leapfrog will no more enlarge the distance between initial value of θ and a proposed value θ^* . The criterion is the derivative with respect to time of half the squared distance between the θ and θ^* , as in equation 2.36,

$$\frac{\partial}{\partial t} \left[\frac{(\theta^* - \theta) \cdot (\theta^* - \theta)}{2} \right] = (\theta^* - \theta) \frac{\partial}{\partial t} (\theta^* - \theta) = (\theta^* - \theta) \cdot \varphi, \quad (2.36)$$

where φ is the momentum variable. To guarantee the reversibility of the process we must create an auxiliary variable z , which follows an Uniform distribution. This variable is responsible for backward and forward the Hamiltonian trajectory.

To define an efficient value to ϵ is even more simple. During the warmup NUTS would constantly check if the acceptance probability is sufficiently high, i.e., at least greater than half. If it is not, the algorithm just shorten the step size at next iteration (NESTEROV, 2009) and (HOFFMAN; GELMAN, 2014).

Three obvious limitations of this technique are that a stationary distribution must exist, we must be capable to write the prior and likelihood and the gradient might be declared. The Stan programming language provides the numerical gradient, what solves the third requirement. Finally, the distribution of $\boldsymbol{\varphi}$ could be a multivariate normal with a diagonal variance or a full covariance matrix structure. The former used to be elected

because precision increase is almost irrelevant compare with the computational memory costs, ([Stan Development Team, 2016](#)).

3 MODEL SELECTION AND DIAGNOSTIC ANALYSIS

Model Selection and Diagnostic Analysis are two topics with many challenges to solve if we adopted a Bayesian perspective. The main reason is the difficulty to evaluate a model which already has incorporated subjective judgments as prior. How would be possible to an analyst to judge her or his own beliefs?

As an unconsolidated field, there is a lot of propositions to model checking and selection. We briefly remind some canonical approaches as posterior predictive (RUBIN, 1981), Bayes factor (KASS; RAFTERY, 1995), cross-validation (STONE, 1974), reversible jump Markov chain Monte Carlo (GREEN, 1995) and (HASTIE; GREEN, 2012), and all the vast family of information criteria. The later is the subject of our next section.

The frame of diagnostic analysis depends of the selected model and its theoretical assumption. However, influential points are a constant concern to any analyst. By this reason, our second section is dedicated to influential points, an important aspect of model diagnostic.

3.1 Information Criteria

An objective way to choice models is an indispensable tool for any statistician. Here we present three recent information criteria: DIC, WAIC and LOO.

One of the best articles to understand the hide discussion about information criteria is presented by (GELMAN; HWANG; VEHTARI, 2013). Where the authors define two seminal quantities: the expected log predictive density for a new data point (ELPD) and the expected log pointwise predictive density for a new dataset (ELPPD). Where DIC is a way to estimate the first and WAIC is a path to estimate the second. It follows the respective equations:

$$elpd = E_f(\log p_{post}(\tilde{y}_i)) = \int (\log p_{post}(\tilde{y}_i)) f(\tilde{y}_i) d\tilde{y}. \quad (3.1)$$

$$elppd = \sum_{i=1}^n E_f(\log p_{post}(\tilde{y}_i)) = \sum_{i=1}^n \int (\log p_{post}(\tilde{y}_i)) f(\tilde{y}_i) d\tilde{y}. \quad (3.2)$$

3.1.1 DIC

The deviance information criterion (DIC) is the first authentic Bayesian criteria (SPIEGELHALTER et al., 2002) and (LINDE, 2005), which generalized the frequentist idea of AIC, (AKAIKE, 1973), to Bayesian framework. The core of DIC is just the $elpd_{DIC}$, a Bayesian version of AIC with two relevant changes, replacing the maximum likelihood

estimate $\hat{\boldsymbol{\theta}}$ with the posterior mean $\hat{\boldsymbol{\theta}}_{Bayes} = E(\boldsymbol{\theta}|\mathbf{y})$ and replacing k with a data-based bias correction. The new measure of predictive accuracy is,

$$\widehat{elpd}_{DIC} = \log f(\mathbf{y}|\hat{\boldsymbol{\theta}}_{Bayes}) - p_{DIC}, \quad (3.3)$$

where p_{DIC} is the effective number of parameters, defined as,

$$p_{DIC} = 2 \log f(\mathbf{y}|\hat{\boldsymbol{\theta}}_{Bayes}) - E_{post}(\log f(\mathbf{y}|\boldsymbol{\theta})), \quad (3.4)$$

where the expectation in the second term is an average of $\boldsymbol{\theta}$ over its posterior distribution. DIC core is multiplied by minus two to displays an AIC similar format as in equation 3.5.

$$DIC = -2 \log f(\mathbf{y}|\hat{\boldsymbol{\theta}}_{Bayes}) + 2p_{DIC}. \quad (3.5)$$

3.1.2 WAIC

To explain the WAIC we must expend one more intermediate step. In practice, we do not know the parameter $\boldsymbol{\theta}$, so we cannot know the log predictive density $\log p(\mathbf{y}|\boldsymbol{\theta})$. By this reason we would work with the posterior distribution, $p_{post}(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y})$, and summarize the predictive accuracy of the fitted model to data by the log pointwise predictive density (LPPD).

$$lppd = \log \prod_{i=1}^n p(y_i) = \sum_{i=1}^n \int f(y_i|\boldsymbol{\theta})p_{post}(\boldsymbol{\theta})d\boldsymbol{\theta}, \text{ calculated as } \sum_{i=1}^n \log \left(\frac{1}{S} \sum_{s=1}^S f(y_i|\boldsymbol{\theta}^s) \right), \quad (3.6)$$

where $\boldsymbol{\theta}^s$ are the sampled parameters and S is the number of iterations of the Markov chain. WAIC is consider a fully Bayesian approach because it takes into account all the posterior distribution instead of a single point estimation as DIC, which uses an average of $\boldsymbol{\theta}$ over its posterior distribution, (WATANABE, 2010). The core of WAIC is like,

$$\widehat{elpd}_{WAIC} = lppd - p_{WAIC}. \quad (3.7)$$

where lppd was define in equation 3.6 and p_{WAIC} is the correspondent bias correction for

effective number of parameters to avoid overfitting,

$$p_{WAIC} = \sum_{i=1}^n \log(E_{post} f(y_i|\boldsymbol{\theta})) - E_{post} \log(f(y_i|\boldsymbol{\theta})), \text{ calculated as} \quad (3.8)$$

$$2 \sum_{i=1}^n \left(\log \left(\frac{1}{S} \sum_{s=1}^S f(y_i|\boldsymbol{\theta}^s) \right) - \frac{1}{S} \sum_{s=1}^S \log f(y_i|\boldsymbol{\theta}^s) \right).$$

We define WAIC as -2 times the expression 3.7 to be analogous with DIC.

3.1.3 LOO

In Bayesian cross-validation, the data are split into a training \mathbf{y}_{train} and a test set \mathbf{y}_{test} , after the model is fit to \mathbf{y}_{train} we check the fitting quality with the outstanding test data. The main idea of LOO is to repeat n times the cross-validation procedure each time with $n - 1$ observations into the training set (GELMAN; HWANG; VEHTARI, 2013). This cross-validation strategy allows us to solve a log pointwise predictive density (LPPD) by LOO, i.e.,

$$lppd_{LOO} = \sum_{i=1}^n \log p_{post}(-i)(y_i) = \sum_{i=1}^n \log \int p_{pred}(y_i|\boldsymbol{\theta}) p_{post}(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (3.9)$$

$$\text{calculated as } \sum_{i=1}^n \log \left(\frac{1}{S} \sum_{s=1}^S f(y_i|\boldsymbol{\theta}^{i,s}) \right),$$

where p_{pred} correspond to the predictive distribution. Each prediction is conditioned on $n - 1$ data points, which causes an underestimation of $lppd_{LOO}$. (BURMAN, 1989) proposed a first order bias correction, he argued that the bias is equivalent to $b = lppd - \overline{lppd}$, where

$$\overline{lppd}_{-i} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \log p_{post}(-i)(y_j), \text{ calculated as } \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \log \left(\frac{1}{S} \sum_{s=1}^S f(y_j|\boldsymbol{\theta}^{i,s}) \right). \quad (3.10)$$

Consequently, the bias-corrected lppd LOO is equivalent to $lppd_{LOO_e} = lppd_{LOO} + b$. Differently from the previous criteria, LOO does not require a model dimension penalization, but an effective number of parameters can be approximated by $p_{LOO} = lppd - lppd_{LOO}$.

The LOO is the most computational costly, but it is considered the most precise too. Because it is the asymptotic convergence of WAIC. An obvious LOO disadvantage is it requires sample to be divided into disjoint, ideally conditionally independent, pieces. However, strictly talking both previous criteria have the same restriction. This is a clear limitation to structured models, as our SAR and CAR. Although, our simulation study reveals some empirical hope that these criteria seems to still applicable to spatial models.

3.1.4 Standard errors

To improve computational speed of WAIC and LOO, they must be estimate by Pareto importance sampling. This technique requires to declare a log pointwise predictive inside the HMC code.

This strategy implies in violate an important assumption: the dependence structure of the likelihood. However, if we proceed despite of this obstruction, we can define $WAIC_{(-i)}$ as the WAIC of a subsample $\mathbf{y}_{(-i)} = \{y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n\}$, which could be estimated by Pareto importance sampling as $\widehat{WAIC}_{(-i)}$ (VEHTARI; GELMAN; GABRY, 2016).

A second advantage of applying Pareto importance sampling consist of do a "sample" of n $WAIC_{(-i)}$. Then we can obtain the standard error of our estimated WAIC,

$$SE(\widehat{WAIC}) = \sqrt{Var_{i=1}^n \widehat{WAIC}_{(-i)}/n}. \quad (3.11)$$

A similar result might be construct to LOO criterion. By Pareto importance sampling we can estimate $LOO_{(-i)}$ and its standard error,

$$SE(\widehat{LOO}) = \sqrt{Var_{i=1}^n \widehat{LOO}_{(-i)}/n}. \quad (3.12)$$

The availability of standard errors could help us at the time of interpret if a difference between two or many criteria is or is not relevant. Once information criteria are not a statistical test and there is no significance inference associated at them, any graphical tool is useful.

3.2 Bayesian model diagnostics

A main procedure in analysis of diagnostics consist of be able to identify if any observation may strongly influence the model parameter estimation and prediction. In other words, what would be changed if a single observation could be removed or displays a quite different value?

3.2.1 Influential points

An influential point consist of some observation which strongly changes parameter estimations. The classical example is a point which drastically alters the slope parameter in a linear regression. In Bayesian inference, our focus lies into all posterior distribution instead of a single parameter estimating.

The reason of a point to be or not to be an influential observation consist in a complex issue. Many times, outliers could be responsible for strongly influence in the model.

However, outliers and influential points are two different concepts. An outlier consist of a large lack between observation value, y_i , and its prediction, \hat{y}_i (COOK; WEISBERG, 1982). On the other hand, an influential point is an observation which if it would be remove from the data it would cause a drastic modification in the posterior distribution.

In linear regression background an influential observation used to be called as a point of leverage. Furthermore, it is easy to separate outliers from influential points in that context. Pearson error measures the lack of fit in each observation and the Cook's distance measures the influence (COOK; WEISBERG, 1982). Although these concepts were proposed to independent data, they were generalized to dependent models, see (FOX, 1972).

Influential observations role a main place in any model diagnostic. Because all the conclusions are directly affect if the estimation of parameters has problems. However, complex models require a different approach. Seeking for leverage effect in all parameters which we must estimate in a complex model seems unfeasible.

Further, dependent models have an additional obstacle, which consist of an impossibility of split the sample. An intuitive way to check about the stability of estimations is to divide the sample and separately estimate the parameters. If each partition produces a similar estimation we have arguments to accept the hypothesis of low influence. This appealing idea is hard to extend to dependent data for a reason, how to split data without crashing the characteristics of likelihood?

Fortunately, Bayesian inference should produce a posterior distribution. So, if there is a function which measures the distance between two probability densities we can measure distance between two posterior distributions or between a Bayesian model and its perturbed version.

We can use a well know function as Kullback-Leiber to estimate the divergence between two posterior distributions. The following subsections are dedicated to present functional Bregman divergence, a generalization of Kullback-Leiber and how to measure perturbed models.

3.2.2 Functional Bregman divergence

The functional Bregman measures a divergence between functions, which could be probability densities. We do not used to describe Bregman as a distance because there is no symmetry. We define (Ω, X, ν) as a finite measure space and $f_1(x)$ and $f_2(x)$ as two non-negative functions, (GOH; DEY, 2014).

Definition 3 *Let us consider $\psi : (0, \infty) \rightarrow \mathbb{R}$ be a strictly convex and differentiable function on \mathbb{R} . Then the functional Bregman divergence D_ψ is defined under the marginal*

density $\nu(x)$ as

$$D_\psi(f_1, f_2) = \int \psi(f_1(x)) - \psi(f_2(x)) - \psi'(f_2(x))[f_1(x) - f_2(x)]d\nu(x), \quad (3.13)$$

where ψ' represents the derivative of ψ .

This divergence has some already proofed proprieties, which used to be present in the following sequence:

1. (Nonnegativity) $D_\psi(f_1, f_2) \geq 0$ for any non-negative measurable functions and equality holds if and only if $f_1 = f_2$;
2. (Convexity) $D_\psi(f_1, f_2)$ is convex with respect to f_1 , but not necessarily with respect to f_2 ;
3. (Linearity) $D_{c_1\psi^*+c_2\psi^{**}}(f_1, f_2) = c_1D_{\psi^*}(f_1, f_2) + c_2D_{\psi^{**}}(f_1, f_2)$ for any positive c_1 and c_2 as well as for any ψ^* and ψ^{**} which respect the Definition 3;
4. (Equivalent classes) Whether $\psi(f_1) = \psi^*(f_1) + bf_1 + c$, where $b, c \in \mathbb{R}$, thus $D_\psi(f_1, f_2) = D_{\psi^*}(f_1, f_2)$. Therefore, the set of strictly convex functions, ψ , might be partitioned into equivalent classes such that $[\psi^*] = \{\psi | D_\psi(f_1, f_2) = D_{\psi^*}(f_1, f_2)\}$;
5. (Linear separation) The geometric place of the non-negative measurable function f which has the same distance from two fixed functions f_1 and f_2 is a hyperplane;
6. (Dual divergence) Let us consider ψ as a Legendre function and ψ^* as its conjugate, thus $D_\psi(f_1, f_2) = D_{\psi^*}(f_1^*, f_2^*)$, where f_1 and f_2 are respectively related to f_1^* and f_2^* by Legendre transformation;
7. (Generalized Pythagorean inequality) For any non-negative measurable functions f_1, f_2 and f_3 the functional Bregman divergence satisfies the equation: $D_\psi(f_1, f_3) = D_\psi(f_1, f_2) + D_\psi(f_2, f_3) + \int \psi'(f_2) - \psi'(f_3)[f_1 - f_2]d\nu$.

All the above proprieties were debate by (FRIGYIK; SRIVASTAVA; GUPTA, 2008a) in context of functional divergences, and they were already proofed in a technical report of the same authors, see (FRIGYIK; SRIVASTAVA; GUPTA, 2008b).

Let us discuss about the limit of a strictly convex function to a convex one, i.e., what happens if we elected the identity as our ψ ? Then $\psi(f(x)) = f(x) \forall f(x)$ and $\psi'(f(x)) = 1 \forall f(x)$. Consequently, the functional Bregman would be zero to any $f_1(x)$ and $f_2(x)$.

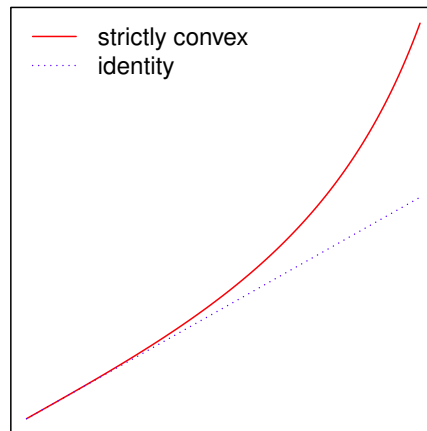


Figure 1: Example of strictly convex and identity functions

Whether we choose a strictly convex ψ as illustrated in Figure 1, then Bregman divergence would be always greater than zero, except for trivial case where $f_1(x) = f_2(x)$. Furthermore, ψ works as a tuning parameter and if we increase its distance from the identity we might have $D_\psi(f_1, f_2)$ so big as we could desire, it does not matter the functions $f_1(x)$ and $f_2(x)$.

The choice of the convex function ψ presents some degree of freedom. Here we follow the suggestion of (GOH; DEY, 2014) and restricted to the class of convex functions defined by (EGUCHI; KANO, 2001), i.e., $\psi_\alpha(x)$, where $\alpha \in \mathbb{R}$.

$$\psi_\alpha(x) = \begin{cases} x \log x - x + 1, & \alpha = 1 \\ -\log x + x - 1, & \alpha = 0 \\ (x^\alpha - \alpha x + \alpha - 1)/(\alpha^2 - \alpha), & \text{otherwise.} \end{cases} \quad (3.14)$$

Even though the choice of α has infinity degree of freedom, different values reserves distinct qualities, as we shall discuss later. Three of the possible choices of α are more studied and already known as specific divergences. If $\alpha = 0$ reduces to Itakura-Saito distance, $\alpha = 1$ becomes Kullback-Leibler divergence, $\alpha = 2$ is the squared Euclidean distance or $L^2/2$.

3.2.3 Perturbation on dependent models

Here we extend the ideas from (GOH; DEY, 2014), who define some perturbation for models which fitting independent and identically distributed observations to dependent

models. Let $\{f(\mathbf{y}|\boldsymbol{\theta}, X)|\boldsymbol{\theta} \in \Theta\}$ be a class of statistical models, (GOH; DEY, 2014) define a general perturbation as follows:

$$\delta(\boldsymbol{\theta}, \mathbf{y}, X) = \frac{f_{\delta}(\mathbf{y}|\boldsymbol{\theta}, X)\pi_{\delta}(\boldsymbol{\theta})}{f(\mathbf{y}|\boldsymbol{\theta}, X)\pi(\boldsymbol{\theta})}, \quad (3.15)$$

where δ indicates that likelihood or prior suffers some perturbation. In the outlier detection context, the perturbation may be restricted just inside the likelihood, or even in the \mathbf{y} variable, once the prior was unaltered. To avoid misinterpretation, the authors define a specific perturbation, δ_1 , restricted to likelihood.

$$\delta_1(\boldsymbol{\theta}, \mathbf{y}, X) = \frac{f(\mathbf{y}_{(i)}|\boldsymbol{\theta}, X)\pi(\boldsymbol{\theta})}{f(\mathbf{y}|\boldsymbol{\theta}, X)\pi(\boldsymbol{\theta})} = \frac{f(\mathbf{y}_{(i)}|\boldsymbol{\theta}, X)}{f(\mathbf{y}|\boldsymbol{\theta}, X)}, \quad (3.16)$$

where $\mathbf{y}_{(i)}$ was consider by the authors as \mathbf{y} vector without the i-th case . We cannot proceed as them, because we can not exclude an observation without modify A or W matrix and consequently the likelihood is not comparable. Even more, as (WANG; LUNG-FEI, 2013) highlighted the worst way to work with model estimation on dependence framework is case deletion. Instead, they obtained much better results by data imputation.

For all these reasons, we propose a different way to express $\mathbf{y}_{(i)}$. Instead of remove an observation, we inputted an observation with its prediction, $\mathbf{y}_{(i)} = (y_1, \dots, \hat{y}_i, \dots, y_n)'$, what means y_i case suffers an imputation which does not depend of it. The specific point y_i to be inputted in the vector $\mathbf{y}_{(i)}$ is define as in equation 3.17 for the SAR case.

$$\hat{y}_i = \mathbf{x}'_i \hat{\boldsymbol{\beta}} + \hat{\rho} W \mathbf{y}, \quad (3.17)$$

where $\hat{\rho}$ is a previous estimation of ρ , as well as $\hat{\boldsymbol{\beta}}$ from $\boldsymbol{\beta}$. Again, y_i was not required to estimate \hat{y}_i , because $w_{i,j} = 0$ when $i = j$. And for the CAR, we suggest the procedure in equation 3.18.

$$\hat{y}_i = \mathbf{x}'_i \hat{\boldsymbol{\beta}} + \sum_{j=1}^n \hat{\phi} a_{ij} (y_j - \mathbf{x}'_j \hat{\boldsymbol{\beta}}), \quad (3.18)$$

where $\hat{\phi}$ and $\hat{\boldsymbol{\beta}}$ are the previous estimation of ϕ and $\boldsymbol{\beta}$, respectively. In both cases, SAR and CAR, \hat{y}_i just depend of i-th observation in the covariates, \mathbf{x}'_i . In this way, we have a technique to judge the local influence of i-th point, which consist of proceed with a divergence between the likelihood and its inputted analogous as in 3.19,

$$d_{\psi,i} = D_{\psi}(f(\mathbf{y}_{(i)}|\boldsymbol{\theta}, X), f(\mathbf{y}|\boldsymbol{\theta}, X)). \quad (3.19)$$

Another motivator objective is to proceed with an analysis of sensitivity, which means in Bayesian inference to understand the weight between select different priors. To

solve this problem, some authors like (GOH; DEY, 2014), (DEY; BIRMIWAL, 1994) and (MCCULLOCH, 1989) follow the scheme displayed in equation 3.20. However, it is hard to conclude something, because a divergence between two priors always will compare a prior with another prior, and there is no way to judge the better one if there is no previous information available.

$$\delta_2(\boldsymbol{\theta}, \mathbf{y}, X) = \frac{f(\mathbf{y}|\boldsymbol{\theta}, X)\pi_\xi(\boldsymbol{\theta})}{f(\mathbf{y}|\boldsymbol{\theta}, X)\pi(\boldsymbol{\theta})} = \frac{\pi_\xi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}, \quad (3.20)$$

where $\pi(\boldsymbol{\theta})$ is a default prior and $\pi_\xi(\boldsymbol{\theta})$ is the same prior with an additional noise. This is an interesting argument, but if we do not have a default prior just becomes useless. For this reason, we propose to follow another strategy as to use information criteria.

This last idea consist of concentrate all model information in a single value, which could be the DIC as used by (RODRIGUES; ASSUNÇÃO, 2012), but might easily be generalized to WAIC or even LOO. If these criteria still around the same values with different priors we may conclude that these priors are not so informative.

3.2.4 Posterior comparison using Bregman divergence

Our main objective is to proceed with a direct comparison between two posteriors $p(\boldsymbol{\theta}|\mathbf{y})$ and $p_\delta(\boldsymbol{\theta}|\mathbf{y})$ using functional Bregman divergence. However, equation 3.13 has not a general closed form. There are two techniques to obtain \hat{D}_ψ . First a Gaussian approximation and later a particular importance sampling, the importance weighted marginal density estimation (IWMDE). The first method is easier to implement, but is not accurate for hierarchical models, and spatial models belong to them. So, for lack of alternatives we continue with IWMDE, (GOH; DEY, 2014).

We are working with many operations which involve posteriors, so it is convenient to define the normalizing constant for $p(\boldsymbol{\theta}|\mathbf{y})$ as

$$m^{-1}(\mathbf{y}) = \int \frac{\omega(\boldsymbol{\theta})}{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}, \quad (3.21)$$

where $m(\mathbf{y})$ is the marginal density $\int f(\boldsymbol{\theta}|\mathbf{y})\pi(\boldsymbol{\theta})$ and $\omega(\cdot)$ is any probability density function, do not confuse with a particular $w_{i,j}$ element of our W matrix. Let $\{\boldsymbol{\theta}^s\}_{s=1}^S$ be samples from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$, where the samples could be generated by HMC. We can sample the inverse of the constant 3.21 using the simulated $\boldsymbol{\theta}^s$ and choosing an appropriate ω .

$$\tilde{m}^{IW}(\mathbf{y}) = \left[\frac{1}{S} \sum_{s=1}^S \frac{\omega(\boldsymbol{\theta}^s)}{f(\mathbf{y}|\boldsymbol{\theta}^s)\pi(\boldsymbol{\theta}^s)} \right]^{-1}. \quad (3.22)$$

By this way we might define the posterior distribution $\tilde{p}^{IW}(\boldsymbol{\theta}|\mathbf{y})$ as the result of the IWMDE processes,

$$\tilde{p}^{IW}(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\tilde{m}^{IW}(\mathbf{y})}. \quad (3.23)$$

From equation 3.15, the estimate of perturbed posterior is given by

$$\tilde{p}_\delta^{IW}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\tilde{p}^{IW}(\boldsymbol{\theta}|\mathbf{y})\delta(\boldsymbol{\theta}, \mathbf{y}, X)}{\frac{1}{S} \sum_{s=1}^S \delta(\boldsymbol{\theta}^s, \mathbf{y}, X)}, \quad (3.24)$$

where $\tilde{p}^{IW}(\boldsymbol{\theta}|\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/\tilde{m}^{IW}(\mathbf{y})$. Consequently, we can approximate the functional Bregman divergence between $p(\boldsymbol{\theta}|\mathbf{y})$ and $p_\delta(\boldsymbol{\theta}|\mathbf{y})$ by

$$\hat{D}_\psi^{IW} = \frac{1}{S} \sum_{s=1}^S \left\{ \frac{\psi(\tilde{p}^{IW}(\boldsymbol{\theta}^s|\mathbf{y})) - \psi(\tilde{p}_\delta^{IW}(\boldsymbol{\theta}^s|\mathbf{y})) - (\tilde{p}^{IW}(\boldsymbol{\theta}^s|\mathbf{y}) - \tilde{p}_\delta^{IW}(\boldsymbol{\theta}^s|\mathbf{y}))\psi'(\tilde{p}_\delta^{IW}(\boldsymbol{\theta}^s|\mathbf{y}))}{\tilde{p}^{IW}(\boldsymbol{\theta}^s|\mathbf{y})} \right\}. \quad (3.25)$$

Whether the convex function $\psi = \psi_\alpha(x)$ as viewed at 3.14, we could simplify the above equation 3.25.

$$\hat{D}_{\psi_\alpha}^{IW} = \frac{1}{S} \sum_{s=1}^S \left\{ \frac{1 - \alpha \{\delta(\boldsymbol{\theta}^s, \mathbf{y}, X)/\bar{\delta}\}^{\alpha-1} + (\alpha - 1) \{\delta(\boldsymbol{\theta}^s, \mathbf{y}, X)/\bar{\delta}\}^\alpha}{\alpha(\alpha - 1) \{\tilde{p}^{IW}(\boldsymbol{\theta}^s|\mathbf{y})\}^{1-\alpha}} \right\}, \quad (3.26)$$

where $\bar{\delta} = \frac{1}{S} \sum_{s=1}^S \delta(\boldsymbol{\theta}^s, \mathbf{y}, X)$. Furthermore, if we restrict $\alpha = 1$, we simplify a lot the above expression and might not care about the ω weights choice.

$$\hat{D}_{\psi_\alpha}^{IW} = \frac{1}{S} \sum_{s=1}^S \left\{ -\log\{\delta(\boldsymbol{\theta}^s, \mathbf{y}, X)/\bar{\delta}\} \right\}, \text{ if } \alpha = 1. \quad (3.27)$$

3.2.5 Normalizing Bregman divergence

Any result from a functional Bregman divergence belongs to positive real domain, $d_{\psi,i} \in [0, \infty)$ and $md_\psi \in [0, \infty)$. A scale where it is hard to have some intuition about what is a high or a low value. To facilitate this comparison, (MCCULLOCH, 1989) proposed a calibration of Kullback-Leiber, which compresses the scale between 0.5 and 1. Naturally, his results could be extend to Bregman divergence.

He made an analogy with two Bernoulli distribution. Considering d_{kl} as any Kullback-Leiber divergence of interest like $KL(P, Q) = d_{kl}$. The main idea is to compare d_{kl} with a known density like a Bernoulli with probability of success equal to half. Though

the KL between two probability distribution of Bernoulli is known and is easily expressed as in equation 3.28,

$$KL(B(0.5), B(q)) = -\frac{1}{2} \log(4q(1-q)), \quad (3.28)$$

where $q = c(d_{kl})$. After this, he just inverted the equation to isolate $c(d_{kl})$ like showed in Proposition 1.

Proposition 1 *Given d_{kl} a Kullback-Leiber divergence of interest,*

$$c(d_{kl}) = \frac{1 + \sqrt{1 - \exp(-2d_{kl})}}{2}, \quad (3.29)$$

$c(\cdot)$ is the McCulloch's calibration.

The McCulloch's idea is a good start point, but it has a weak theoretical element. Because it requires compare any probability distribution with a Bernoulli what do not make sense in several contexts.

We also suggest another way to compare the Bregman divergence between two densities, which consist of normalizing Bregman divergence. Suppose we have a full density probability f_0 and we wish to compare it with each likelihood without i -th element as f_1, \dots, f_n to check local influence. We already know that the sum of all divergences belongs to positive real domain,

$$\sum_{i=1}^n d_{\psi}(f_0, f_i) = k, \quad (3.30)$$

where k is a positive constant. The value of k could be zero if and only if ψ is the identity, but also could be arbitrarily high conform ψ becomes more and more convex, in particular k may be one.

By generalized Pythagorean inequality (FRIGYIK; SRIVASTAVA; GUPTA, 2008b), it is natural to suppose the order maintenance as $d_{\psi^*}(f_1, f_2) > d_{\psi^*}(f_1, f_3) \implies d_{\psi^{**}}(f_1, f_2) > d_{\psi^{**}}(f_1, f_3)$ to any ψ^* and ψ^{**} under the restriction of strictly convexity. If the above order relation is maintained then we can guarantee that all Bregman divergence with any ψ consisting of the same divergence just with a different location scale.

Now, we are gathering these two arguments: a finite sum of divergences is finite and ψ just tuning the scale but not the order of Bregman. So there is a special case of ψ , let us call it as ψ_{χ} , which the sum of a set of n f_i results in one, and this divergence $d_{\psi_{\chi}}$ is equal to any normalizing Bregman divergence $B(d_{\psi})$. Because all Bregman divergence contains the same information about the order.

Proposition 2 Given $d_\psi(f_0, f_q)$ any Bregman divergence of interest in a set of n divergences,

$$B(d_\psi(f_0, f_q)) = d_{\psi_x}(f_0, f_q) = \frac{d_\psi(f_0, f_q)}{\sum_{i=1}^n d_\psi(f_0, f_i)}, \quad (3.31)$$

where $B(\cdot)$ is normalizing Bregman.

Any normalizing Bregman divergence belongs to range between zero and one, $0 \leq d_{\psi_x}(f_0, f_q) \leq 1, \forall q \in \{1, \dots, n\}$. Consequently, there is no other divergence so intuitive to work like the normalizing Bregman. For example, an output of 0.9 has a large possibility to be an influential point. However, what it is a high and low result it still a hard task, because this scale depends of the sample size.

Under the null hypothesis that there is no influential observation in the sample, a reasonable expected normalizing Bregman would be $1/n$.

$$E(d_{\psi_x}(f_0, f_i)) = \frac{1}{n} \forall i \in \{1, \dots, n\},$$

i.e., we expected each observation would present the same divergence. Therefore, this bound becomes our start point to identify influential observation. If any observation returns a higher normalizing Bregman value than $1/n$, then it is a natural candidate to be an influential point, which we must investigate.

However, a better cutter point than $1/n$ still as a open problem to future researches. Yet we have an useful graphical and visual tool to seek for influence, but not a theoretical constant which can segregate influential from not influential cases.

4 SIMULATION

This chapter is dedicated to check how well our theoretical assumptions work in many different computational simulated settings. The first section is developed to attend analysis of sensitivity, because it is reasonable to advance with the best possible prior to each model. We proposed sixteen settings, which consist in the combination of two models (CAR and SAR), two variance patterns (homoscedastic and heteroscedastic), two spatial coefficients ($\phi = 0.7$ and $\phi = 0.9$) and two priors. We replicate five hundred times each simulation setting.

We wish to test how well our information criteria select the models. To proceed with this challenge we create artificial SAR and CAR vectors and evaluate them with the correct and incorrect models to check if WAIC and LOO catch misspecification. Then we have two settings (CAR and SAR) each one with five models. The adjusted models to SAR data were: I the corrected, II with a wrong covariate, III a heteroscedastic SAR, IV a CAR, V a linear model. The adjusted models to CAR data were quite similar.

We realize the Bregman divergence behavior for sixteen settings of disturbed likelihoods. They consist in the combination of two models (CAR and SAR), two variance patterns (homoscedastic and heteroscedastic) and four sets of perturbation (absent, one point, another point and two points).

All models which have run at this chapter were tuning with two chains and ten thousand of samples with burn in of five thousand. Both criteria of convergence, Geweke (GEWEKE, 1992) and Gelman (GELMAN; RUBIN, 1992) were satisfied, as well as the graphical analysis.

4.1 Analysis of Sensitivity

A traditional analysis of sensitivity consist of to fit a model with many different priors and to evaluate their impact in the parameters. We proceed with the study of two meaningful priors: an extremely flat prior and another with more concentrated probability density. This choice was motivated by the reason of a divergence between authors about which of the priors is the most appropriated, (Stan Development Team, 2016). The prior 1, the flat, has the following hyperparameters: $\alpha = 0.01, \beta = 0.01, \eta = 100$. The prior 2, the concentrated, has the following hyperparameters: $\alpha = 2.01, \beta = 1.01, \eta = 100$.

The first model which we are studying is the homoscedastic SAR with the follow parameters: $\phi = 0.7, \beta_0 = 0.2, \beta_1 = -0.3, \sigma^2 = 1$ and another very similar model with a different $\phi = 0.9$. We may observe in Table 2 that both priors return almost the same results for any parameter. Consequently, the SAR seems robust to these priors choice.

Table 2: Analysis of Sensitivity to 500 simulated homoscedastic SAR models with $\phi \in \{0.7, 0.9\}$

prior 1 = $\{\alpha = 0.01, \beta = 0.01, \eta = 100\}$								
	$\phi = 0.7$	$\beta_0 = 0.2$	$\beta_1 = -0.3$	$\sigma = 1$	$\phi = 0.9$	$\beta_0 = 0.2$	$\beta_1 = -0.3$	$\sigma = 1$
mean	0.752	0.154	-0.295	0.992	0.891	0.224	-0.307	1.016
bias	0.052	-0.046	0.005	-0.008	-0.009	0.024	-0.007	0.016
var	0.018	0.041	0.029	0.022	0.006	0.092	0.037	0.021
mse	0.021	0.044	0.029	0.022	0.006	0.092	0.037	0.022
prior 2 = $\{\alpha = 2.01, \beta = 1.01, \eta = 100\}$								
	$\phi = 0.7$	$\beta_0 = 0.2$	$\beta_1 = -0.3$	$\sigma = 1$	$\phi = 0.9$	$\beta_0 = 0.2$	$\beta_1 = -0.3$	$\sigma = 1$
mean	0.752	0.153	-0.295	0.991	0.894	0.219	-0.306	0.993
bias	0.052	-0.047	0.005	-0.009	-0.006	0.019	-0.006	-0.007
var	0.018	0.040	0.029	0.021	0.006	0.088	0.037	0.020
mse	0.021	0.043	0.029	0.021	0.006	0.088	0.037	0.020

The next model which we are studying is the heteroscedastic SAR with the follow parameters $\phi = 0.7, \beta_0 = 0.2, \beta_1 = -0.3, \sigma_1 = 1, \sigma_2 = 4, \sigma_3 = 1, \dots, \sigma_{26} = 1, \sigma_{27} = 4$ and another very similar model with a different $\phi = 0.9$. We proposed the same two priors. To save space, we just display the main parameters, i.e., the spatial parameter and four dispersion, just for illustrate the general pattern. We might observe in Table 3 that the priors have a strong impact in parameter estimation. With the first prior almost all mean values of σ_i are greater than 7, even close to 40. However with the second prior we observe that σ_1 and σ_{27} are softly greater than 3 and all the others are between 0.9 and 1.5.

Consequently, we might not put any prior in a heteroscedastic SAR model. This kind of model has too much parameters and we must put some restriction on their estimation, a proper prior with mean equal to one could be an interesting way, as the inverse Gamma with $\alpha = 2.01, \beta = 1.01$ would fit it.

Table 3: Analysis of Sensitivity to 500 simulated heteroscedastic SAR models with $\phi \in \{0.7, 0.9\}$

prior 1 = $\{\alpha = 0.01, \beta = 0.01, \eta = 100\}$										
	$\phi = 0.7$	$\sigma_1 = 1$	$\sigma_2 = 4$	$\sigma_{26} = 1$	$\sigma_{27} = 4$	$\phi = 0.9$	$\sigma_1 = 1$	$\sigma_2 = 4$	$\sigma_{26} = 1$	$\sigma_{27} = 4$
mean	0.708	13.637	35.013	16.269	32.874	0.859	9.960	39.313	7.285	31.201
bias	0.008	12.637	31.013	15.269	28.874	-0.041	8.960	35.313	6.285	27.201
var	0.038	22665.6	11126.6	28080.3	11221.4	0.015	5869.6	45436.5	277.8	3541.7
mse	0.039	22825.3	12088.4	28313.5	12055.1	0.017	5949.8	46683.5	317.3	4281.5
prior 2 = $\{\alpha = 2.01, \beta = 1.01, \eta = 100\}$										
	$\phi = 0.7$	$\sigma_1 = 1$	$\sigma_2 = 4$	$\sigma_{26} = 1$	$\sigma_{27} = 4$	$\phi = 0.9$	$\sigma_1 = 1$	$\sigma_2 = 4$	$\sigma_{26} = 1$	$\sigma_{27} = 4$
mean	0.728	1.390	3.068	1.390	3.197	0.898	0.955	2.670	0.971	2.959
bias	0.028	0.390	-0.932	0.390	-0.803	-0.002	-0.045	-1.330	-0.029	-1.041
var	0.019	0.149	3.234	0.134	3.415	0.008	0.192	3.269	0.203	4.110
mse	0.019	0.301	4.102	0.286	4.060	0.008	0.194	5.038	0.203	5.195

We keep our study with the homoscedastic CAR with the follow parameters: $\phi = 0.7, \beta_0 = 0.2, \beta_1 = -0.3, \sigma = 1$ and another very similar just with a different $\phi = 0.9$. We proposed the same two priors already discussed for SAR to both cases. We may observe in Table 4 that both priors return almost the same results for any parameter. Consequently, the CAR seems robust to these priors choice. However, there is a underestimation of ϕ in all the cases, this phenomenon was already warned, (BANERJEE; CARLIN; GELFAND, 2015). We tried to fix this point by the following schemes: we increase the number of iterations from 20,000 to 100,000, we have reparameterized the model, we add a new hierarchical level with the inclusion a hyperprior. None of these proposes have success, even change the Uniform prior of ϕ to a Beta one had no effect if this prior was not extremely informative. This ultimate appeal is not useful, because a so strong prior would estimate ϕ as a fixed value.

Table 4: Analysis of Sensitivity to 500 simulated homoscedastic CAR models with $\phi \in \{0.7, 0.9\}$

prior 1 = $\{\alpha = 0.01, \beta = 0.01, \eta = 100\}$									
	$\phi = 0.7$	$\beta_0 = 0.2$	$\beta_1 = -0.3$	$\sigma = 1$	$\phi = 0.9$	$\beta_0 = 0.2$	$\beta_1 = -0.3$	$\sigma = 1$	
mean	0.294	0.200	-0.302	0.792	0.201	0.197	-0.298	0.974	
bias	-0.406	0.000	-0.002	-0.208	-0.699	-0.003	0.002	-0.026	
var	0.026	0.001	0.006	0.047	0.020	0.004	0.009	0.092	
mse	0.191	0.002	0.006	0.090	0.509	0.004	0.009	0.093	
prior 2 = $\{\alpha = 2.01, \beta = 1.01, \eta = 100\}$									
	$\phi = 0.7$	$\beta_0 = 0.2$	$\beta_1 = -0.3$	$\sigma = 1$	$\phi = 0.9$	$\beta_0 = 0.2$	$\beta_1 = -0.3$	$\sigma = 1$	
mean	0.293	0.199	-0.302	0.780	0.199	0.196	-0.297	0.954	
bias	-0.407	-0.001	-0.002	-0.220	-0.701	-0.004	0.003	-0.046	
var	0.026	0.002	0.006	0.043	0.021	0.004	0.010	0.085	
mse	0.192	0.002	0.006	0.092	0.512	0.004	0.010	0.087	

We have build a heteroscedastic CAR with the same structure as heteroscedastic SAR, i.e., $\Theta = \{\phi = 0.7, \beta_0 = 0.2, \beta_1 = -0.3, \sigma_1 = 1, \sigma_2 = 4, \sigma_3 = 1, \dots, \sigma_{26} = 1, \sigma_{27} = 4\}$ and another very similar just with a different $\phi = 0.9$. Unsurprising, this model presents similar problems as the heteroscedastic SAR. We may see in Table 5 that the first prior almost all mean values of σ_i are greater than 5. Although in the second prior we observe that all σ_i are around 1 when $\phi = 0.7$. What is a problem, because we expected to see greater values for σ_1 and σ_{27} . However when $\phi = 0.9$ the bias of prior 2 to σ_i is lower.

Another way to proceed with analysis of sensitivity consist of reduces all model information in a single measure and evaluate it. This is not a so traditional technique, but it was already applied by researchers as (RODRIGUES; ASSUNÇÃO, 2012). They observed the DIC, if its value does not change too much to each prior then they would conclude that model does not strongly depend of priors. Here we used the LOO instead of the DIC, because LOO owns some advantages, but the argument is similar to the

Table 5: Analysis of Sensitivity to 500 simulated heteroscedastic CAR models with $\phi \in \{0.7, 0.9\}$

prior 1 = $\{\alpha = 0.01, \beta = 0.01, \eta = 100\}$										
	$\phi = 0.7$	$\sigma_1 = 1$	$\sigma_2 = 4$	$\sigma_{26} = 1$	$\sigma_{27} = 4$	$\phi = 0.9$	$\sigma_1 = 1$	$\sigma_2 = 4$	$\sigma_{26} = 1$	$\sigma_{27} = 4$
mean	0.296	21.963	7.608	9.705	10.265	0.179	13.613	43.234	9.445	213.487
bias	-0.404	20.963	3.608	8.705	6.265	-0.721	12.613	39.234	8.445	209.487
var	0.032	2.1×10^4	138.1	291.1	329.3	0.023	530.3	3.2×10^4	178.8	3.3×10^6
mse	0.195	2.1×10^4	151.1	366.9	368.5	0.543	689.3	3.4×10^4	250.1	3.3×10^6
prior 2 = $\{\alpha = 2.01, \beta = 1.01, \eta = 100\}$										
	$\phi = 0.7$	$\sigma_1 = 1$	$\sigma_2 = 4$	$\sigma_{26} = 1$	$\sigma_{27} = 4$	$\phi = 0.9$	$\sigma_1 = 1$	$\sigma_2 = 4$	$\sigma_{26} = 1$	$\sigma_{27} = 4$
mean	0.253	1.102	1.008	0.981	1.037	0.136	1.249	2.499	1.098	1.841
bias	-0.447	0.102	-2.992	-0.019	-2.963	-0.764	0.249	-1.501	0.098	-2.159
var	0.022	0.378	0.270	0.180	0.337	0.015	0.388	2.805	0.316	1.203
mse	0.222	0.388	9.224	0.181	9.115	0.598	0.450	5.058	0.325	5.865

previous authors. In the Table 6 we summarize all models and priors which we already have discussed above, but just with the mean value of LOO and its standard deviation.

Table 6: LOO mean and (standard deviation) to 500 simulated cases of each model, variance type, prior and value of $\phi \in \{0.7, 0.9\}$

model	$\phi = 0.7$		$\phi = 0.9$	
	prior 1	prior 2	prior 1	prior 2
Homoscedastic SAR	78.65 (8.36)	78.60 (8.38)	79.86 (8.02)	79.78 (8.03)
Heteroscedastic SAR	128.64 (13.31)	86.99 (10.08)	130.02 (13.69)	86.90 (10.00)
Homoscedastic CAR	65.37 (15.25)	65.34 (15.27)	75.66 (17.21)	75.58 (17.21)
Heteroscedastic CAR	122.01 (17.31)	73.53 (15.64)	136.48 (19.12)	89.22 (19.51)

We might see in Table 6 that the homoscedastic SAR displays similar mean values of LOO to the both prior and likely values of standard deviation. So, we have a robust model as well as the homoscedastic CAR, which present similar results. Remark, we just compare row results and with the same ϕ parameter in Table 6, other parallels do not make sense because they involve distinct data.

The heteroscedastic SAR and CAR have very distinct mean values of LOO conform we change the priors. Consequently, we can not fit that models without a rigorous balance. The conclusions after to analysis Table 6 are not so different from the previous tables. But it is quite simple to analysis just the LOO despite of all parameters.

4.2 Model selection and misspecification

In the last section, the LOO criterion was used to compare priors and proceed with analysis of sensitivity. However it is mainly and originally applied to model selection as we shall develop in this section.

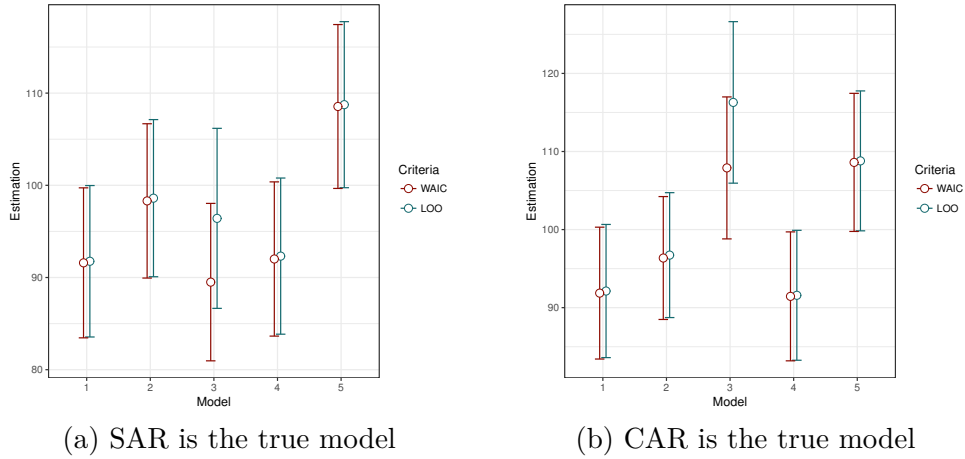


Figure 2: Model comparison by WAIC and LOO; at the left the true model is a homoscedastic SAR and the proposed models are: I the corrected, II with a wrong covariate, III a heteroscedastic SAR, IV a CAR, V a linear model; at the right the true model is a homoscedastic CAR and the proposed models are: I the corrected, II with a wrong covariate, III a heteroscedastic CAR, IV a SAR, V a linear model.

First we create a vector of constants as covariate to an artificial homoscedastic SAR vector. The parameters were $\Theta = \{\phi = 0.8, \beta_0 = 0.4, \beta_1 = -0.3, \sigma^2 = 1\}$. Then we proposed five models to this simulated data set and compare which better fit it by WAIC and LOO. The fitted models were respectively: I the correct, II another homoscedastic SAR with a wrong covariate, III a heteroscedastic SAR with the corrected covariate, IV a homoscedastic CAR with the true covariate, V a simple linear regression with the correct predictor variable.

We might see in Figure 2 (left) that the corrected model presents the lower values of both WAIC and LOO, 91.5 and 91.7, respectively. The same model with the wrong covariate displays the worst result with WAIC of 98.3 and LOO of 98.6. The heteroscedastic SAR is the unique where there is some disagreement between WAIC 89.4 and LOO 96.4, the WAIC is even lower than the true model, but LOO is more trustworthy as we shall see by Kullback-Leiber. The CAR was not bad with 92.0 and 92.3 WAIC and LOO. Finally, no spatial regression was the worst model.

Figure 3 displays the point divergence of each model. We used the Kullback-Leiber instead of the normalizing Bregman, because we want to compare all models with the same scale. For example, a model where all observations have a bad fitting could display a constant result to each point and consequently small values of normalizing Bregman.

Figure 3 (left) shows that regression, pink down triangle, has the worst fit in position 22. But the heteroscedastic SAR, green up triangle, was bad in observations 22 and 27. Remind that homoscedastic SAR was worst than heteroscedastic SAR with respect to WAIC. Though plot 3 (left) helps us to decide if we must trust in WAIC or

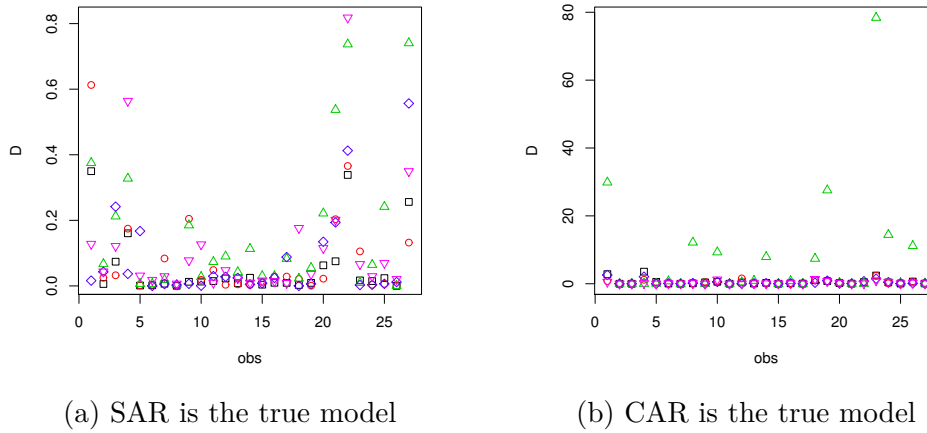


Figure 3: Kullback-Leiber divergence for simulated data, at the left the true model is a homoscedastic SAR and the proposed models are: I the corrected in black square, II with a wrong covariate in red circle, III a heteroscedastic SAR in green up triangle, IV a CAR in blue rhombus, V a LM in pink down triangle; at the right the true model is a homoscedastic CAR and the proposed models are: I the corrected in black square, II with a wrong covariate in red circle, III a heteroscedastic CAR in green up triangle, IV a SAR in blue rhombus, V a LM in pink down triangle.

LOO, because homoscedastic SAR, black square, just shows two influential observations by KL: 1 and 22. However both were lower than the concurrent model.

We also create a vector of constants as covariate to an artificial homoscedastic CAR vector. The parameters were the same used to SAR model. Then we proposed many models to this simulated exercise. The fitted models were respectively: I the correct, II another homoscedastic CAR with a wrong covariate, III a heteroscedastic CAR with the correct covariate, IV a homoscedastic SAR with the true covariate, V a simple linear regression with the correct predictor variable.

In Figure 2 (right) the model evaluation is clear with exception of model I, the corrected with 91.8 and 92.1, and IV, a SAR with 91.4 and 91.5, WAIC and LOO. So the SAR looks subtly better than the true model: a CAR. Shortly, the Figure 3 (right) just confirm what we already know, the linear regression has strongly problems of fitting and compare with it all the other models are quite good.

4.3 Perturbation of Likelihood and Bregman divergence

To check if Bregman divergence is a valid way to seek influential observations we must test it by simulation. We create a homoscedastic SAR vector from the following parameters $\phi = 0.8, \beta_0 = 0.2, \beta_1 = -0.3, \sigma^2 = 1$. Then we proceed with the four tuning operations: I, a SAR vector without any alteration; II, a SAR model with an addition noise in a single observation; III, an equivalent to previous one, but putting a noise in a

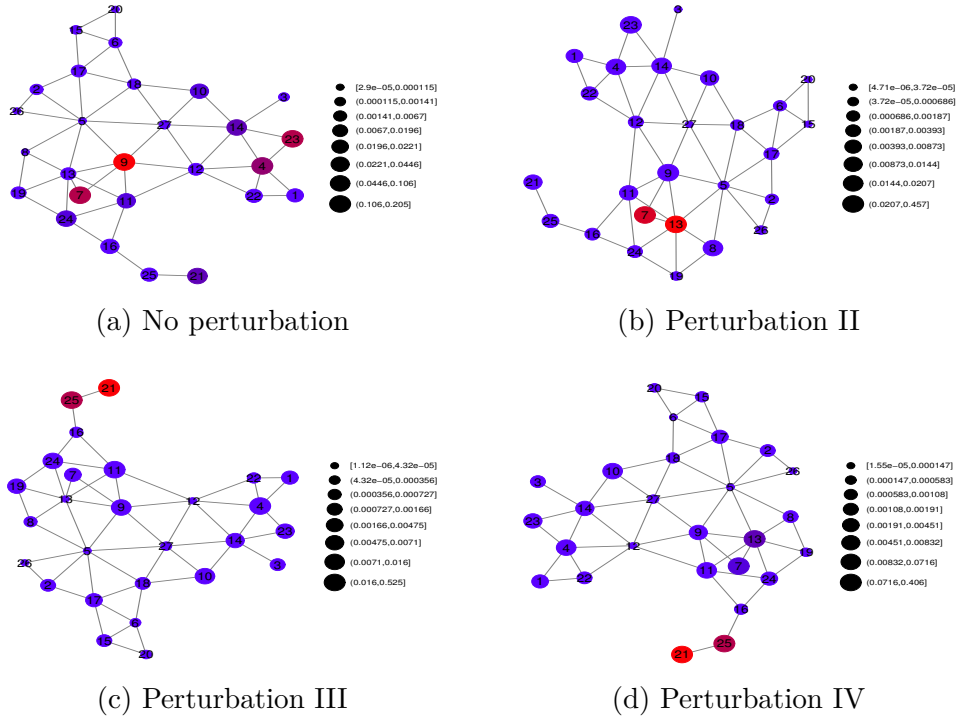


Figure 4: Normalizing Bregman divergence for simulated homoscedastic SAR data

different observation; IV a SAR vector with two perturbations at the same time.

We use the same contamination intensity that (GOH; DEY, 2014) suggest, i.e., to add in the process a quantity q , where q is $1 - 10^{-8}$ quantile of a standard normal. This quantity is roughly the same as suggest by (HAO; LIN, 2016), as well as by (CHO et al., 2009). We choose the position 13 to set II, observation 21 to III and the points 13 and 21 to IV. The adjacent matrix is the Brazilian map, which has 27 states.

Figure 4 shows each set in a respective graph. In the subfigure (b) we see the perturbed observation 13 in red color, what indicates it is an influential point candidate. It is followed by 7, which is not a real influential one, but it is affected by 13, because they are neighbors. This is an advantage of see information by net. In subfigure (c) we see observation 21 in red, what is coherent with its real status, again it is followed by its neighbor the observation 25 in purple. In subfigure (c) 21 is red and 25, which is its border connection is purple. Although 13 presents blue color scale, we observe it is a big circle. In subfigure (a) the observation 9 is red although there is no influential point, this is a false positive. In next figure we will fix this misperception.

Figure 5(a) shows the four sets at the same plot, we lose the net perception, but we gain scale precision. It becomes clear that any candidate to influential point must be greater than the dotted line, $1/n$. The observation 13 was higher than $1/n$, dotted line, when there are two noise points and greater than 0.3, dashed line, when it is alone. On the other hand observation 21 was higher than 0.3 and higher than 0.5, respectively. 7

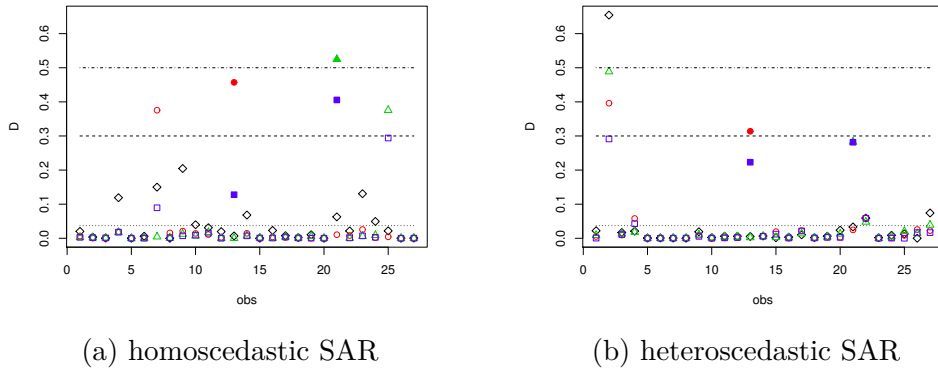


Figure 5: Normalizing Bregman divergence for simulated SAR data, where black rhombus is no perturbation set, red circle is Perturbation II, green triangle is III, blue square is IV. Solid colors were used to indicate true influential point.

appears as a misunderstood of the metrics, with a single value higher than 0.3.

An important remark is that the choice of this lines in Figure 5 and 8 are simple exposure feature to better describe the plots around the text.

Now we must progress with a more complex model, which is the heteroscedastic SAR. We create a heteroscedastic SAR vector with one covariate with the known parameters: $\beta_0 = 0.2$, $\beta_1 = -0.3$, $\phi = 0.8$, and each $\sigma_i^2 = 1$ for all i , except for $\sigma_4^2 = 16$. The scheme of perturbation still the same, even the same observations 13 and 21.

The Figure 5(b) displays observations 2, 13 and 21 as the stronger candidates to be influential points in any set. Unfortunately, 2 is the most outstanding though it is a false positive. At least 13 and 21 are always higher than $1/n$ and around 0.3. Observe that the square and triangle of 21 are so close they seems a single point which it hides the triangle.

Figure 6 reminds us the point positions in the net. However, there is no directly connection between observation 4, which it was infected, and 2. So, the explanation could not be build by this way. Just the increase of model complexity turn the Bregman divergence a not so precise metric to analyze influence.

We wish to check the same patterns for CAR model. Thus, we simulated a similar set to it. We created a homoscedastic CAR vector with the same SAR parameters. The infection scheme is analogous, i.e., we select the same observations and intensity to put an additional noise.

In Figure 7 we can see all the four sets separately, in topleft we observe the observation 7 in red, topright the 13, bottomleft the 21, bottomright the 13 and the 21 in purple. What is exactly the corrected infected positions except for 7, which belongs to a set without perturbation.

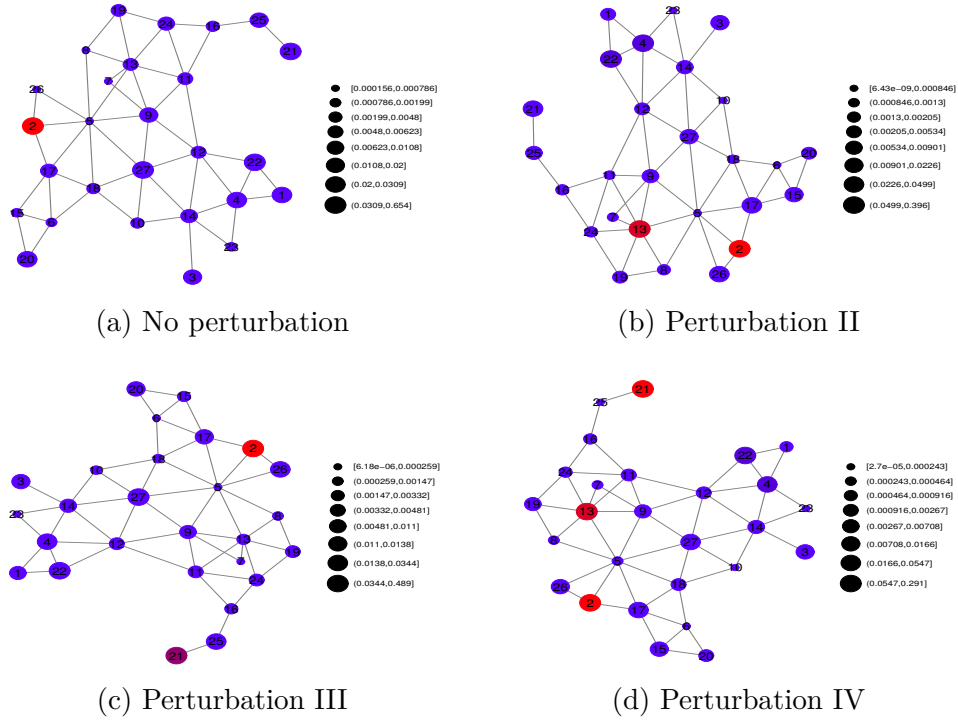


Figure 6: Normalizing Bregman divergence for simulated heteroscedastic SAR data

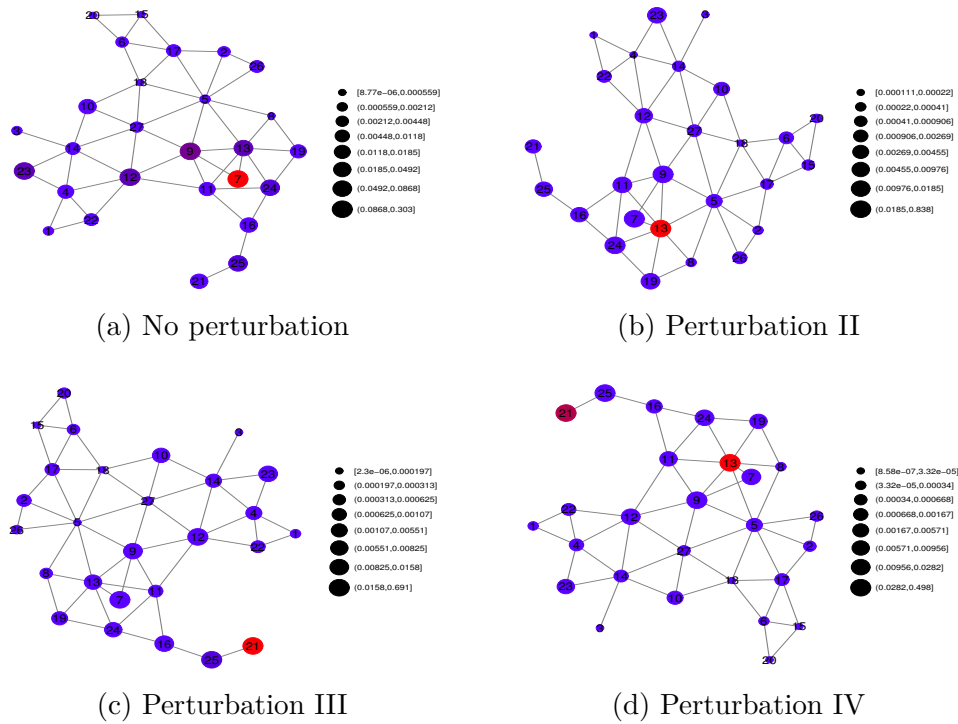


Figure 7: Normalizing Bregman divergence for simulated homoscedastic CAR data

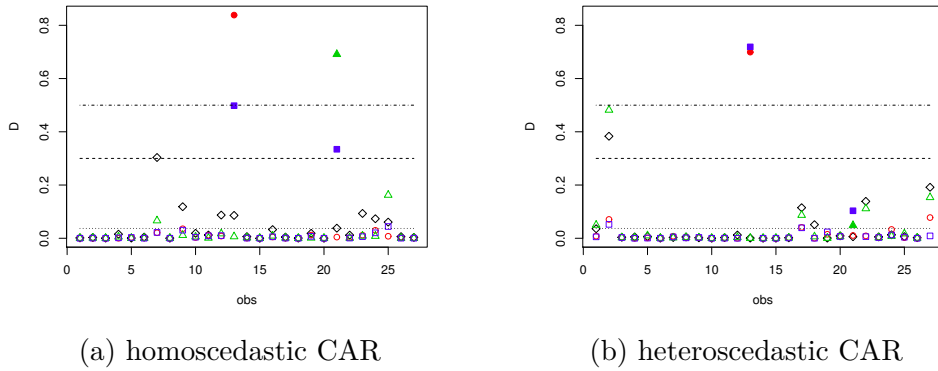


Figure 8: Normalizing Bregman divergence for simulated CAR data, where black rhombus is no perturbation set, red circle is Perturbation II, green triangle is III, blue square is IV. Solid colors were used to indicate true influential point.

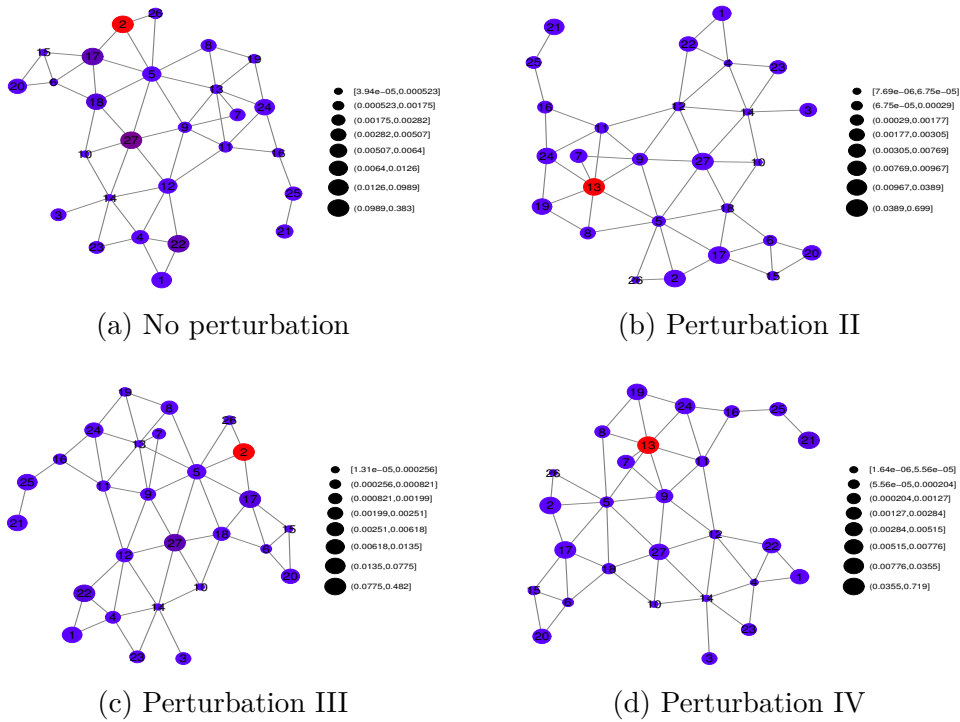


Figure 9: Normalizing Bregman divergence for simulated heteroscedastic CAR data

By Figure 8(a) we realize that position 7, the greatest in no perturbation set, is subtly greater than 0.3, the dashed line, while in Perturbation IV set both observations are around 0.3 level. Furthermore, when there is only one perturbation, 13 is between 0.3 and 0.5 despite 21 shows a greater estimation than 0.5. Though, position 25 also appears as a false positive in a single set.

In Figure 9 we can observe all the four sets for heteroscedastic CAR, in topleft we might note the observation 2 in red, topright the 13, bottomleft the 2 and in bottomright just

the 13. Consequently, we may temporarily conclude that we can identify the perturbation in 13, but not in 21 point.

As in heteroscedastic SAR, the CAR with complex variability presents a strong false positive, which is the observation 2. Figure 8(b) displays clearly point 2 as a dominant position in Bregman divergence. However, the true infected observations 13 and 21 are the following high positions. Both around the bound of 0.3 in any set.

5 APPLICATION

In this chapter we illustrated our methods with two real data sets, a Brazilian and an European examples. To each data set we seek for covariates, they are: GPD per capita, and two legislation covariates to Brazilian study and GPD per capita and educational gender gap to European study. To each different covariate we proposed five models, they are a homoscedastic SAR, heteroscedastic SAR, homoscedastic CAR, heteroscedastic CAR and a linear model. To all twenty five models, we use the same procedure, which consist of two chains, each one with 20,000 iterations; half of them were destine to warmup, given a total post-warmup of 20,000 draws. All models have converged as the $\hat{R} = 1$ indicates, see (GELMAN; RUBIN, 1992), as well as by graphical analysis which we omitted to save space.

5.1 Indexes of Sustainable Development in Brazil

The second United Nations Conference on Environment and Development (UNCED) was held in Rio de Janeiro in 1992. One of its products was the Agenda 21, which propose new ideas about development and encourages governments to produce policy and data according to sustainable paradigm, (SMARDON, 2008).

The Instituto Brasileiro de Geografia e Estatística (IBGE) coordinates an effort to produce more than 20 indexes each year after 2009. The sustainable development indexes enlarge the dichotomy frame which split economic developing and social environmental maintenance. The objective is to produce data with which is possible to do new scientific questions. The prescriptive mark of 2001 wrote by the Sustainable Development Commission (SDC) from UN was a starting point to define the body from the research, which shows four dimensions: environmental, social, economic and institutional (IBGE, 2015).

Between all these indexes there are some already canonical as child mortality, gross domestic product (GDP) per capita and literacy. Despite the importance of these famous indexes, we choose as target variable something newer as the use of pesticides in cultivated fields. This variable is measured in kilogram per hectare and is related with people's health, development of agricultural techniques and market, see (CAPOBIANGO; CARDEAL, 2005) and (PINGALI, 1995). Figure 10 displays a Brazilian map with the target variable. Where regions like Midwest and South present higher levels of agrototoxic use.

From all available covariates we elected the following to build our models. First a canonical covariate as GDP per capita (in thousands of Reais), a new SDI as percent of counties with environmental council, and another similar SDI as percent of counties with environmental legislation. To each covariate we propose five models: I. a homoscedastic

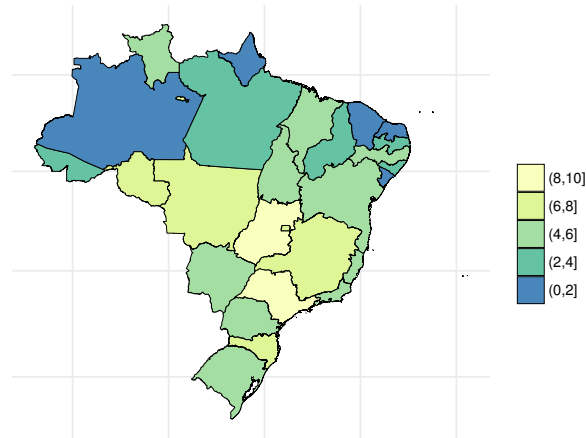


Figure 10: Brazil 2013, pesticide per cultivated area in kg/hectare

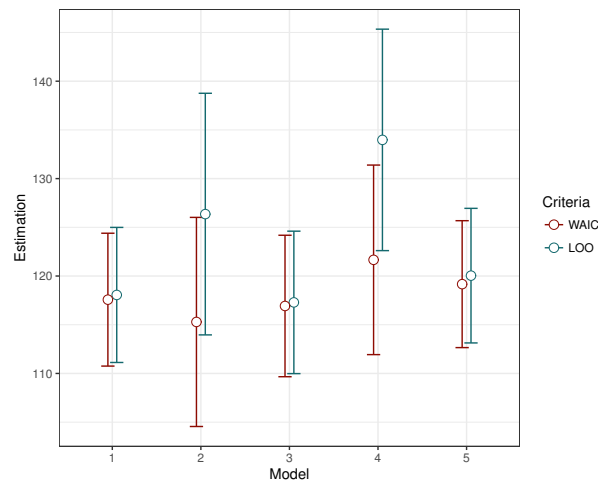


Figure 11: Brazil 2013, comparison of models, where I is a homoscedastic SAR, II is a heteroscedastic SAR, III is a homoscedastic CAR, IV is a heteroscedastic CAR, V is a linear regression.

SAR, II. a heteroscedastic SAR, III. a homoscedastic CAR, IV. a heteroscedastic CAR, V. a linear regression.

The GDP per capita completely dominate the set and displays better fit in any situation. Consequently we show in Figure 11 the results only for GPD. We might observe that the SAR and CAR, both homoscedastic, are the best candidates.

By Figures 12 and 13 we remark that both models are quite similar about their influential points. The difference is minimal. SAR has Amazonas, Roraima and Distrito Federal as influential points candidates as well as CAR presents Amazonas and São Paulo. However, neither have big values of divergence, so possible both models are appropriate to

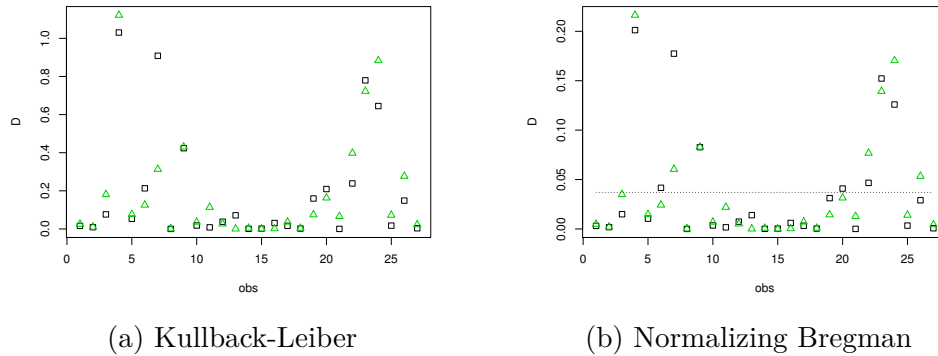


Figure 12: Brazil 2013, Kullback-Leiber and normalizing Bregman, where black square is a homoscedastic SAR, green triangle is a homoscedastic CAR.

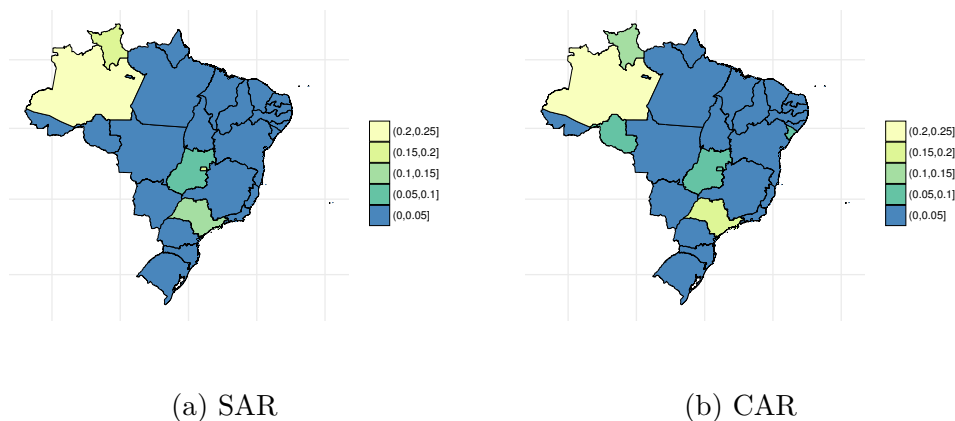


Figure 13: Brazil 2013, Normalizing Bregman divergence.

fit Brazilian agricultural practices.

In Table 7 we present the main parameters of the better models: CAR and SAR homoscedastic. Both present weak spatial estimation, ϕ around 0.18 and 0.5. And an even worst estimation of predictor coefficient, which is 0.05 to both, but the 95% posterior probability interval flows between -0.03 and 0.13. These results do not give us a directly and strongly conclusion about the variables relationship.

However, there is a hide point: Brazilian GPD also has a strong spatial pattern. Consequently, an analysis which include this variable already inputs an indirect geographical structure. Furthermore, the signal of $\hat{\beta}_1$ was positive in the linear regression, so if there is a "collinearity" between ϕ and β_1 it do not invert the relationship between agrotoxic and GPD. After these appointments, we may say that pesticide use follows spatial proximity and that wealthy states tend to more intensively use it.

Table 7: Brazil 2013, parameter estimation by model

model	parameter	mean	SD	2.5%	25%	50%	75%	97.5%	ESS	\hat{R}
CAR	β_0	2.71	1.27	0.29	1.85	2.67	3.51	5.32	7558	1
	β_1	0.05	0.04	-0.03	0.02	0.05	0.08	0.13	6796	1
	ϕ	0.18	0.06	0.04	0.14	0.18	0.22	0.30	7311	1
	σ	1.94	0.28	1.49	1.74	1.91	2.10	2.57	9738	1
SAR	β_0	0.96	1.00	-0.97	0.30	0.96	1.63	2.91	10786	1
	β_1	0.05	0.04	-0.03	0.03	0.05	0.08	0.13	10441	1
	ϕ	0.50	0.24	0.05	0.32	0.50	0.68	0.94	9650	1
	σ	1.95	0.28	1.49	1.75	1.92	2.12	2.60	11666	1

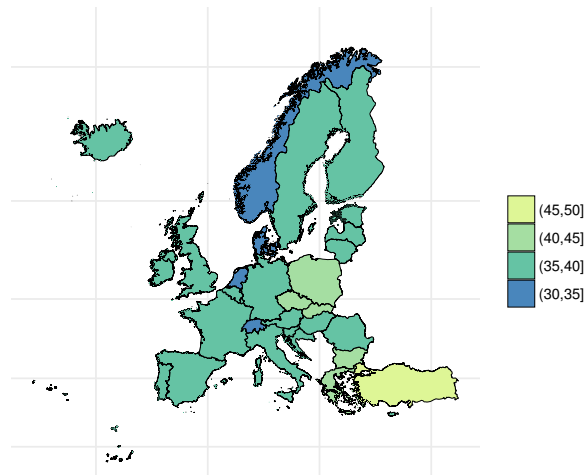


Figure 14: Europe 2014, average weekly hours of work.

Finally, we conclude that local environmental council and legislation do not have already effect the Brazilian pesticide policy. Even a simple SAR without covariates produces a better WAIC and LOO than a linear regression with that variables.

5.2 Working conditions in Europe Union

The average weekly hours of work does not available all the complexity of working conditions, but it may roughly indicate labor dignity and quality of life. Because, less time of work means more time to enjoy the family, culture and friendship (BURGOON; BAXANDALL, 2004).

In this section we investigate the work conditions in the Europe context. Though we used data provenient from Eurostats, the statistical bureau of the Euro Commission. However, to build an adjacent matrix we must to known what is Europe or European Union. Which countries should be include in our analysis and how to objectively build their connections?

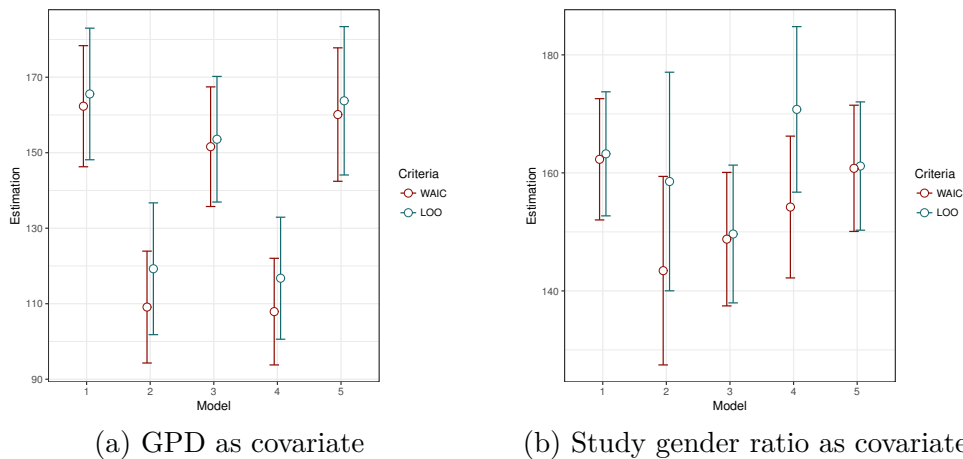


Figure 15: Europe 2014, comparison of models, where I is a homoscedastic SAR, II is a heteroscedastic SAR, III is a homoscedastic CAR, IV is a heteroscedastic CAR, V is a linear regression.

Currently, European Union has 28 associate countries, but it makes sense to exclude of a spatial analysis some countries like Switzerland, which is located in the hearth of the continent? Or Turkey, which is a important partner and member of NATO for so many years?

Furthermore, the connections between these countries are not so objective as in Brazilian example. Now we have to make some arbitrary judgments as link Ireland and United Kingdom, Cyprus with Turkey and Greece, Denmark with Sweden and United Kingdom with France. Some times there is a sea between this places, but also there are bridges between them.

To choice which covariates to include in our study we seek for authors who previously discuss this theme as (BURGOON; BAXANDALL, 2004). They summarize the previous literature as a long relationship between time of work and struggle of the unions, gender equality and main political ideology. They indicate that the unions participation is ambiguous and poorly effective. Ideological view is something that is hard to precisely extend to each European country.

We choose as covariates some gender indicators and GDP per capita in PPS¹, which is a classical economic variable. We also include a educational variable: percent of population which has highest ISCED (International Standard Classification of Education). About gender equality, we investigate the gender overall earnings gap and the ratio between male and female percent of population which has highest ISCED.

From all above listed covariates, the better results in terms of LOO information

¹ Purchasing Power Standards (PPS) is expressed in relation to the European Union (Euro 28) average set to equal 100

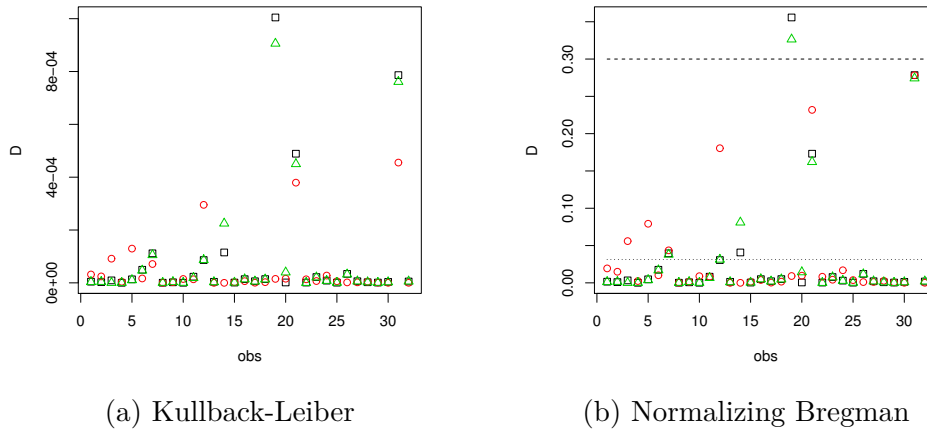


Figure 16: Europe 2014, comparison of models, where black square is a heteroscedastic CAR with GPD, red circle is a homoscedastic CAR with study gender ratio green triangle is a heteroscedastic SAR with GPD.

come from GPD per capta and gender educational attainment level ratio. We just plot the results of this two covariates in Figure 15 at left and right, respectively. When the covariate is the GPD per capta the best models are clearly the heteroscedastic form of SAR and CAR. When the covariate is the gender educational attainment level ratio the best model is the homoscedastic CAR.

Before decide what is the most appropriate model, we check the influential points. In Figure 16 we observe these three models through Kullback-Leiber and normalizing Bregman divergence. The heteroscedastic CAR for GPD per capta shows the following countries as the most influential: 19, Luxembourg, 31, Turkey, and 21, Netherlands. The heteroscedastic SAR displays the exactly same countries in the first positions. On the other hand, homoscedastic CAR with a gender equality has a different order: 31, Turkey, 21, Netherlands, and 12, Greece.

The Kullback-Leiber, a between model comparable scale, in Figure 16, shows gender equality as the covariate with lower local influence. So, despite the better results of LOO for the GPD covariate, the equality gender variable presents lower KL than the GPD per capta. For these reasons, the three models are competitive and fair choices.

Figure 17 shows the same normalizing Bregman results already presented in Figure 16 (b), but in the map. Where we can identify some patterns in the normalizing Bregman. A north cluster of Netherlands, Denmark and Luxembourg. A southeast group of Turkey, Greece and Bulgaria. And an isolated island like Iceland.

Table 8 summarize the parameter estimation of the three main models. The heteroscedastic models present lower average values of ϕ because the complex variance structure explain almost all the variability. Both models where the covariate is the GPD

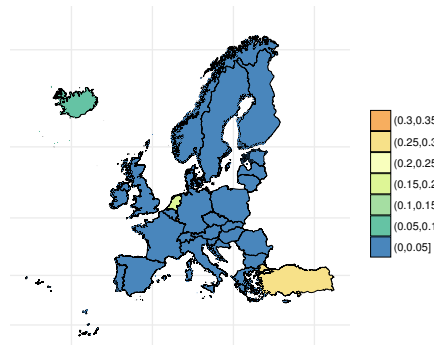
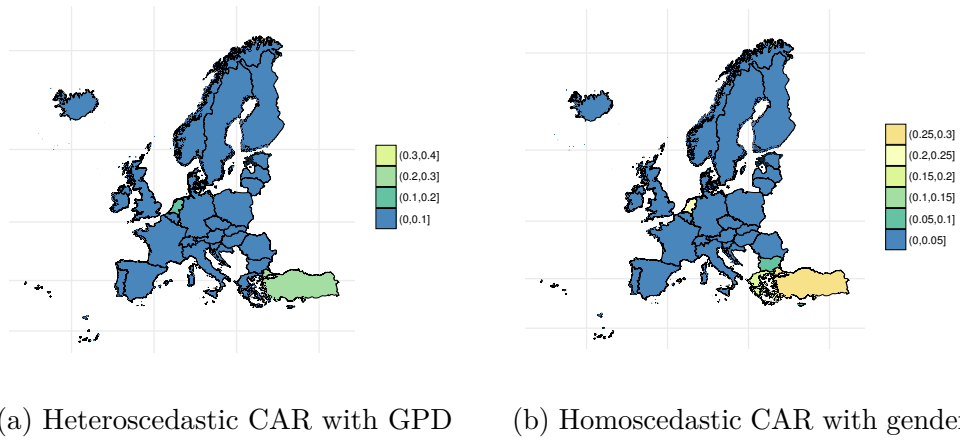


Figure 17: Europe 2014, Normalizing Bregman.

Table 8: Europe 2014, parameter estimation by model

model	parameter	mean	SD	2.5%	25%	50%	75%	97.5%	ESS	\hat{R}
CAR (GPD)	β_0	43.10	0.39	42.37	42.84	43.08	43.35	43.90	19861	1
	β_1	-0.05	0.00	-0.06	-0.06	-0.05	-0.05	-0.05	21620	1
	ϕ	0.01	0.01	0.00	0.00	0.01	0.02	0.05	40000	1
	σ_1	0.77	0.66	0.23	0.42	0.60	0.89	2.31	25640	1
CAR (gender)	β_0	24.96	3.87	17.21	22.39	25.02	27.55	32.43	15023	1
	β_1	11.33	3.51	4.54	8.99	11.26	13.65	18.36	14861	1
	ϕ	0.28	0.06	0.15	0.24	0.28	0.32	0.41	21177	1
	σ	2.23	0.29	1.74	2.02	2.20	2.40	2.89	21298	1
SAR (GPD)	β_0	41.36	1.79	37.00	40.62	42.06	42.62	43.31	13104	1
	β_1	-0.05	0.00	-0.06	-0.05	-0.05	-0.05	-0.04	12006	1
	ϕ	0.04	0.04	0.00	0.01	0.02	0.06	0.14	19871	1
	σ_1	0.69	0.62	0.21	0.38	0.54	0.79	2.22	19231	1

per capita, the β_1 estimation is negative what indicates people from wealthy countries used to work fewer hours per week. When the covariate is gender educational attainment level ratio the estimation of β_1 is positive what means places where men have more study opportunity than women the population used to work more hours per week.

6 CONCLUSION

SAR and CAR models can be estimated by HMC and are useful techniques to predicted spatial observations. Where Stan platform plays an important role to implement HMC. WAIC and LOO are information criteria which were proposed to independent and identically distributed variables, but they well worked in our spatial dependent models. Also, functional Bregman divergence were originally proposed to compare perturbed and unperturbed likelihoods when they are iid. However, our strategy of data imputation has well worked.

From our sensitivity study we recommend to use a concentrated prior like $a = 2.01$ and $b = 1.01$ as the hyperparameter to σ and σ_i in both models: SAR and CAR. Furthermore, the heteroscedastic case requires an informative prior, because the model already is poorly identifiable. This prior is more conservative and stable than other more flat priors. About model selection: WAIC and LOO tend to agree, but in more complex models as heteroscedastic cases LOO tends to judge better than WAIC. The standard error of information criteria require iid assumption, but in practice it was not required. Although it was a great way to evaluate if the mean estimation of WAIC and LOO were relevant or not. Our suggestion is to select models with near LOO and proceed with the final election after observing the Bregman divergence. Where points higher than $1/n$ and 0.3 are weak and strong candidates to be influential observations, respectively.

We checked our techniques in empirical data from two different continents. Spatial models present better results than linear models for information criteria and Bregman divergence in both data set. We realize the connection between agrototoxic and GPD as well as hours of work and GPD and gender educational ratio despite the spatial pattern from Brazil and Europe.

Our main contribution is to extend a recent theoretical proposal from iid context to dependent models. We observe that Bregman divergence works quite well in our Bayesian spatial models, then we suggest its use for analysis of diagnostics.

BIBLIOGRAPHY

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: PETROV, B. N.; CSAKI, F. (Ed.). **Proceedings of the Second International Symposium on Information Theory**. [S.l.]: Budapest: Akademiai Kiado. Reprinted in *Breakthroughs in Statistics*, New York: Springer (1992), 1973. p. 610–624.

BANERJEE, S.; CARLIN, B. P.; GELFAND, A. E. **Hierarchical Modeling and Analysis for Spatial Data**. 2. ed. [S.l.]: Chapman and Hall Book, New York, 2015.

BESAG, J. Spatial interaction and the analysis of lattice systems. **Journal of the Royal Statistical Association, Series B**, v. 36, p. 192–236, 1974.

BURGOON, B.; BAXANDALL, P. Three worlds of working time: The partisan and welfare politics of work hours in industrialized countries. **Politics & Society**, v. 32, n. 4, p. 439–473, 2004. Disponível em: <https://doi.org/10.1177/0032329204269983>.

BURMAN, P. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. **Biometrika**, v. 76, p. 503–514, 1989.

CAPOBIANGO, H.; CARDEAL, Z. A solid-phase microextraction method for the chromatographic determination of organophosphorus pesticides in fish, water, potatoes, guava and coffee. **Journal of Brazilian Chemical Society**, v. 16, n. 5, p. 907–914, 2005.

CHO, H. et al. Bayesian case influence diagnostic for survival models. **Biometrics**, v. 65, n. 1, p. 116–124, 2009.

COOK, R.; WEISBERG, S. **Residuals and Influence in Regression**. [S.l.]: Chapman and Hall Book, New York, 1982.

CRESSIE, N.; PERRIN, O.; THOMAS-AGNAN, C. Likelihood-based estimation for gaussian mrfs. **Statistical Methodology**, v. 2, p. 1–16, 2005.

CRESSIE, N. A. C. **Statistics for spatial data**. [S.l.]: Wiley, New York, 1993.

DEY, D. K.; BIRMIWAL, L. R. Robust bayesian analysis using divergence measures. **Statistics and Probability Letters**, v. 20, n. 4, p. 287–294, jul. 1994. Disponível em: <http://www.sciencedirect.com/science/article/B6V1D-45DB1M3-7V/1/f1c2f939d635af139132caa85c85c9be>.

EGUCHI, S.; KANO, Y. Robustifying maximum likelihood estimation. **Institute of Statistical Mathematics**, 2001. Technical Report, Tokyo Japan.

FOX, A. Outliers in time series. **Journal of the Royal Statistical Society**, v. 34, p. 350–363, 1972.

FRIGYIK, B.; SRIVASTAVA, S.; GUPTA, M. Functional bregman divergence and bayesian estimation of distributions. **IEEE Trans. Inform. Theory**, v. 54, p. 5130–5139, 2008.

_____. **An introduction to functional derivatives**. [S.l.], 2008. v. 0001.

- GELMAN, A. et al. **Bayesian Data Analysis**. 2. ed. [S.l.]: Chapman and Hall/CRC Texts in Statistical Science, 2003.
- GELMAN, A.; HWANG, J.; VEHTARI, A. Understanding predictive information criteria for Bayesian models. **Statistics and Computing**, v. 6, p. 1–24, 2013.
- GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical Science**, v. 7, p. 457–511, 1992.
- GEWEKE, J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: BERNADO, J. M. et al. (Ed.). **Bayesian Statistics 4**. Oxford, UK: Clarendon Press, 1992.
- GOH, G.; DEY, D. K. Bayesian model diagnostics using functional bregman divergence. **Journal of Multivariate Analysis**, v. 124, p. 371–383, 2014.
- GREEN, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. **Biometrika**, v. 82, n. 4, 1995.
- HAO, H.-X.; LIN, J.-G. Bayesian case influence analysis for GARCH models based on Kullback-Leiber divergence. **Journal of the Korean Statistical Society**, v. 45, p. 595–609, 2016.
- HASTIE, D. I.; GREEN, P. J. Model choice using reversible jump Markov chain Monte Carlo. **Statistica Neerlandica**, v. 66, n. 3, p. 309, 2012.
- HASTINGS, W. Monte Carlo sampling methods using Markov chains and their applications. **Biometrika**, v. 57, n. 1, p. 97–109, 1970.
- HOFFMAN, M. D.; GELMAN, A. The no-u-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. **Journal of Machine Learning Research**, v. 15, p. 1351–1381, 2014.
- IBGE. **Estudos e Pesquisas - Informação Geográfica**. [S.l.]: IBGE, Rio de Janeiro, 2015. ISSN 1517-1450. ISBN 978-85-240-4347-5, (printed).
- KASS, R. E.; RAFTERY, A. E. Bayes factors and model uncertainty. **Journal of the American Statistical Association**, v. 90, p. 773–795, 1995.
- LINDE, A. van der. DIC in variable selection. **Statistica Neerlandica**, v. 1, p. 45–56, 2005.
- MCCULLOCH, R. E. Local model influence. **Journal of American Statistical Association**, v. 84, n. 406, p. 473–478, 1989.
- METROPOLIS, N. et al. Equations of state calculations by fast computing machines. **Journal of Chemical Physics**, v. 21, p. 1087–1092, 1953.
- NEAL, R. MCMC using Hamiltonian dynamics. In: BROOKS, S. et al. (Ed.). **Handbook of Markov Chain Monte Carlo**. [S.l.]: Chapman and Hall/CRC, 2011. p. 116–162.
- NESTEROV, Y. Primal-dual subgradient methods for convex problems. **Mathematical Programming**, v. 120, n. 1, p. 221–259, 2009.

-
- OLIVEIRA, V. D. Bayesian analysis of conditional autoregressive models. **Ann Inst Stat Math**, v. 64, p. 107–133, 2012.
- PINGALI, P. L. **Impact of Pesticides on Farmer Health and the Rice Environment**. [S.l.]: Springer, Netherlands, 1995.
- RODRIGUES, E. C.; ASSUNÇÃO, R. Bayesian spatial models with a mixture neighborhood structure. **Journal of Multivariate Analysis**, v. 109, p. 88–102, 2012.
- RUBIN, D. B. Estimation in parallel randomized experiments. **Journal of Educational Statistics**, v. 6, p. 377–401, 1981.
- RUE, H.; HELD, L. **Gaussian Markov random fields: Theory and applications**. [S.l.]: Boca Raton: Chapman and Hall, 2005.
- SMARDON, R. C. A comparison of Local Agenda 21 implementation in North American, European and Indian cities. **Management of Environmental Quality: An International Journal**, v. 19, n. 1, p. 118–137, 2008.
- SPIEGELHALTER, D. et al. Bayesian measures of model complexity and fit (with discussion). **Journal of the Royal Statistical Society, Series B**, v. 64, p. 1–34, 2002.
- Stan Development Team. **Stan Modeling Language Users Guide and Reference Manual**. 2016. Version 2.14.0. Disponível em: <<http://mc-stan.org/>>.
- STONE, M. Cross-validatory choice and assesment of statistical preditions (with discussion). **Journal of the Royal Statistical Society B**, v. 36, p. 111–147, 1974.
- VEHTARI, A.; GELMAN, A.; GABRY, J. Practical Bayesian model evaluation using leave-one-out cross-validation and waic. **Statistics and Computing**, 2016. ArXiv preprint: <http://arxiv.org/abs/1507.04544/>.
- WANG, W.; LUNG-FEI, L. Estimation of spatial autoregressive models with randomly missing data in the dependent variable. **Econometrics Journal**, v. 16, p. 73–102, 2013.
- WATANABE, S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. **Journal of Machine Learning Research**, v. 11, p. 3571–3594, 2010.

Appendix

APPENDIX A – CONDITIONAL POSTERIOR DISTRIBUTIONS

Here we show all the conditional posterior distributions present at the text step by step. First the homoscedastic SAR model, where $\boldsymbol{\theta} = \{\rho, \boldsymbol{\beta}, \sigma^2\}$.

$$\begin{aligned}
p(\rho, \boldsymbol{\beta}, \sigma^2 | \mathbf{y}) &\propto \pi(\boldsymbol{\beta}, \rho, \sigma^2) f(\mathbf{y} | \boldsymbol{\theta}) \\
&= \frac{1}{2} \frac{1}{(2\pi\eta)^{k/2}} \exp \left\{ -\frac{1}{2\eta} (\boldsymbol{\beta})' (\boldsymbol{\beta}) \right\} \frac{b^a}{\Gamma(a)} (\sigma^2)^{-(a+1)} \exp \{ -b/\sigma^2 \} \\
&\quad \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y})' (\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y}) \right\} \\
&\quad \propto \exp \left\{ -\frac{1}{2\eta} (\boldsymbol{\beta})' (\boldsymbol{\beta}) \right\} (\sigma^2)^{-(a+1)} \exp \{ -b/\sigma^2 \} \\
&\quad \frac{1}{(\sigma^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y})' (\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y}) \right\} \\
&\propto \frac{1}{(\sigma^2)^{(a+1)+n/2}} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y})' (\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y}) - \frac{1}{2\eta} \boldsymbol{\beta}' \boldsymbol{\beta} - \frac{2b}{2\sigma^2} \right\}.
\end{aligned} \tag{A.1}$$

$$\begin{aligned}
p(\boldsymbol{\beta} | \mathbf{y}) &= \int \int p(\boldsymbol{\beta}, \rho, \sigma^2 | \mathbf{y}) d\sigma^2 d\rho \\
&\propto \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y})' (\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y}) - \frac{1}{2\eta} \boldsymbol{\beta}' \boldsymbol{\beta} \right\} \\
&\propto \exp \left\{ -\frac{1}{2\sigma^2} (-2\mathbf{y}' X\boldsymbol{\beta} + (X\boldsymbol{\beta})' (X\boldsymbol{\beta}) + 2\rho (X\boldsymbol{\beta})' W\mathbf{y}) - \frac{1}{2\eta} \boldsymbol{\beta}' \boldsymbol{\beta} \right\} \\
&= \exp \left\{ -\frac{1}{2\sigma^2} (-2\mathbf{y}' X\boldsymbol{\beta} + \boldsymbol{\beta}' X' X\boldsymbol{\beta} + 2\rho \mathbf{y}' W X\boldsymbol{\beta}) - \frac{1}{2\eta} \boldsymbol{\beta}' \boldsymbol{\beta} \right\} \\
&= \exp \left\{ -\frac{\boldsymbol{\beta}' X' X\boldsymbol{\beta}}{2\sigma^2} - \frac{\boldsymbol{\beta}' \boldsymbol{\beta}}{2\eta} - \frac{1}{2\sigma^2} (-2\mathbf{y}' X\boldsymbol{\beta} + 2\rho \mathbf{y}' W X\boldsymbol{\beta}) \right\} \\
&= \exp \left\{ -\frac{1}{2\sigma^2 \eta} \boldsymbol{\beta}' (\eta X' X + \sigma^2 I_k) \boldsymbol{\beta} - \frac{2}{2\sigma^2 \eta} (-\mathbf{y}' X + \rho \mathbf{y}' W X) \boldsymbol{\beta} \right\} \\
&= \exp \left\{ -\frac{1}{2\sigma^2 \eta} [\boldsymbol{\beta}' (\eta X' X + \sigma^2 I_k) \boldsymbol{\beta} - 2(\mathbf{y}' X - \rho \mathbf{y}' W X) \boldsymbol{\beta}] \right\} \\
&= \exp \left\{ -\frac{1}{2\sigma^2 \eta} [\boldsymbol{\beta}' (\eta X' X + \sigma^2 I_k) \boldsymbol{\beta} - 2(\mathbf{y}' X - \rho \mathbf{y}' W X) (\eta X' X + \sigma^2 I_k) (\eta X' X + \sigma^2 I_k)^{-1} \boldsymbol{\beta}] \right\} \\
&\quad \therefore \boldsymbol{\beta} | \mathbf{y} \sim N_k((\mathbf{y}' X - \rho \mathbf{y}' W X) (\eta X' X + \sigma^2 I_k)^{-1}, \sigma^2 \eta (\eta X' X + \sigma^2 I_k)^{-1}).
\end{aligned} \tag{A.2}$$

$$\begin{aligned}
p(\rho|\mathbf{y}) &= \int \int p(\boldsymbol{\beta}, \rho, \sigma^2|\mathbf{y})d\boldsymbol{\beta}d\sigma^2 \\
&\propto \exp \left\{ -\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y})'(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y}) \right\} I_{\rho \in [-1,1]} \\
&\propto \exp \left\{ -\frac{1}{2\sigma^2} [-2\rho\mathbf{y}'W\mathbf{y} + 2\rho(X\boldsymbol{\beta})'W\mathbf{y} + \rho^2(W\mathbf{y})'(W\mathbf{y})] \right\} I_{\rho \in [-1,1]} \\
&= \exp \left\{ -\frac{(W\mathbf{y})'(W\mathbf{y})}{2\sigma^2} \left[-\frac{2\rho\mathbf{y}'W\mathbf{y}}{(W\mathbf{y})'(W\mathbf{y})} + \frac{2\rho(X\boldsymbol{\beta})'W\mathbf{y}}{(W\mathbf{y})'(W\mathbf{y})} + \rho^2 \right] \right\} I_{\rho \in [-1,1]} \\
&= \exp \left\{ -\frac{(W\mathbf{y})'(W\mathbf{y})}{2\sigma^2} \left[\rho^2 - 2\rho \left(\frac{\mathbf{y}'W\mathbf{y}}{(W\mathbf{y})'(W\mathbf{y})} - \frac{(X\boldsymbol{\beta})'W\mathbf{y}}{(W\mathbf{y})'(W\mathbf{y})} \right) \right] \right\} I_{\rho \in [-1,1]} \\
\therefore \rho|\mathbf{y} &\sim N \left(\frac{\mathbf{y}'W\mathbf{y}}{(W\mathbf{y})'(W\mathbf{y})} - \frac{(X\boldsymbol{\beta})'W\mathbf{y}}{(W\mathbf{y})'(W\mathbf{y})}, \frac{\sigma^2}{(W\mathbf{y})'(W\mathbf{y})} \right) I_{\rho \in [-1,1]}.
\end{aligned} \tag{A.3}$$

$$\begin{aligned}
p(\sigma^2|\mathbf{y}) &= \int \int p(\boldsymbol{\beta}, \rho, \sigma^2|\mathbf{y})d\boldsymbol{\beta}d\rho \\
&\propto (\sigma^2)^{-(a+1+n/2)} \exp \left\{ -\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y})'(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y}) - \frac{2b}{2\sigma^2} \right\} \\
&= (\sigma^2)^{-(a+1+n/2)} \exp \left\{ -\frac{1}{\sigma^2} \left[\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y})'(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y}) + b \right] \right\} \\
\therefore \sigma^2|\mathbf{y} &\sim IG \left(a + \frac{n}{2}, \frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y})'(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y}) + b \right).
\end{aligned} \tag{A.4}$$

Since equation A.5, we observe the heteroscedastic SAR model, where $\boldsymbol{\theta} = \{\rho, \boldsymbol{\beta}, \Sigma\}$.

$$\begin{aligned}
p(\rho, \boldsymbol{\beta}, \Sigma|\mathbf{y}) &\propto \pi(\boldsymbol{\beta}, \rho, \sigma_i^2)f(\mathbf{y}|\boldsymbol{\theta}) \\
&= \frac{1}{2} \frac{1}{(2\pi\eta)^{k/2}} \exp \left\{ -\frac{1}{2\eta}(\boldsymbol{\beta})'(\boldsymbol{\beta}) \right\} \prod_{i=1}^n \left(\frac{b_i^{a_i}}{\Gamma(a_i)} (\sigma_i^2)^{-(a_i+1)} \exp\{-b_i/\sigma_i^2\} \right) \\
&\quad (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y})'\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y}) \right\} \\
&\quad \propto \exp \left\{ -\frac{1}{2\eta}(\boldsymbol{\beta})'(\boldsymbol{\beta}) \right\} \prod_{i=1}^n ((\sigma_i^2)^{-(a_i+1)} \exp\{-b_i/\sigma_i^2\}) \\
&\quad |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y})'\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y}) \right\} \\
&\propto |\Sigma|^{-1/2} \prod_{i=1}^n ((\sigma_i^2)^{-(a_i+1)}) \exp \left\{ -\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y})'\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y}) - \frac{1}{2\eta}\boldsymbol{\beta}'\boldsymbol{\beta} - \sum_{i=1}^n \frac{b_i}{\sigma_i^2} \right\}.
\end{aligned} \tag{A.5}$$

$$\begin{aligned}
p(\boldsymbol{\beta}|\mathbf{y}) &= \int \int p(\boldsymbol{\beta}, \rho, \Sigma|\mathbf{y})d\Sigma d\rho \\
&\propto \exp \left\{ -\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y})'\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y}) - \frac{1}{2\eta}\boldsymbol{\beta}'\boldsymbol{\beta} \right\} \\
&\propto \exp \left\{ -\frac{1}{2} \left((X\boldsymbol{\beta})'\Sigma^{-1}X\boldsymbol{\beta} + \eta^{-1}\boldsymbol{\beta}'\boldsymbol{\beta} - 2\mathbf{y}'\Sigma^{-1}X\boldsymbol{\beta} - 2\rho(W\mathbf{y})'\Sigma^{-1}X\boldsymbol{\beta} \right) \right\} \\
&= \exp \left\{ -\frac{1}{2} \left(\boldsymbol{\beta}'[X'\Sigma^{-1}X + \eta^{-1}I_k]\boldsymbol{\beta} - 2[\mathbf{y}'\Sigma^{-1}X + \rho(W\mathbf{y})'\Sigma^{-1}X]\boldsymbol{\beta} \right) \right\} \\
\therefore \boldsymbol{\beta}|\mathbf{y} &\sim N_k \left([(\mathbf{y}'\Sigma^{-1}X + \rho(W\mathbf{y})'\Sigma^{-1}X)][X'\Sigma^{-1}X + \eta^{-1}I_k]^{-1}, [X'\Sigma^{-1}X + \eta^{-1}I_k]^{-1} \right).
\end{aligned} \tag{A.6}$$

$$\begin{aligned}
p(\rho|\mathbf{y}) &= \int \int p(\boldsymbol{\beta}, \rho, \Sigma|\mathbf{y})d\boldsymbol{\beta}d\Sigma \\
&\propto \exp \left\{ -\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y})'\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y}) \right\} I_{\rho \in [-1,1]} \\
&\propto \exp \left\{ -\frac{1}{2} \left(\rho^2(W\mathbf{y})'\Sigma^{-1}W\mathbf{y} - 2\rho\mathbf{y}'\Sigma^{-1}W\mathbf{y} + \rho(W\mathbf{y})'\Sigma^{-1}X\boldsymbol{\beta} \right) \right\} I_{\rho \in [-1,1]} \\
&= \exp \left\{ -\frac{(W\mathbf{y})'\Sigma^{-1}W\mathbf{y}}{2} \left(\rho^2 - \frac{2\rho\mathbf{y}'\Sigma^{-1}W\mathbf{y}}{(W\mathbf{y})'\Sigma^{-1}W\mathbf{y}} + \frac{\rho(W\mathbf{y})'\Sigma^{-1}X\boldsymbol{\beta}}{(W\mathbf{y})'\Sigma^{-1}W\mathbf{y}} \right) \right\} I_{\rho \in [-1,1]} \\
&= \exp \left\{ -\frac{(W\mathbf{y})'\Sigma^{-1}W\mathbf{y}}{2} \left(\rho^2 - 2\rho \left[\frac{\mathbf{y}'\Sigma^{-1}W\mathbf{y}}{(W\mathbf{y})'\Sigma^{-1}W\mathbf{y}} - \frac{(W\mathbf{y})'\Sigma^{-1}X\boldsymbol{\beta}}{2(W\mathbf{y})'\Sigma^{-1}W\mathbf{y}} \right] \right) \right\} I_{\rho \in [-1,1]} \\
\therefore \rho|\mathbf{y} &\sim N \left(\frac{\mathbf{y}'\Sigma^{-1}W\mathbf{y}}{(W\mathbf{y})'\Sigma^{-1}W\mathbf{y}} - \frac{(W\mathbf{y})'\Sigma^{-1}X\boldsymbol{\beta}}{2(W\mathbf{y})'\Sigma^{-1}W\mathbf{y}}, ((W\mathbf{y})'\Sigma^{-1}W\mathbf{y})^{-1} \right) I_{\rho \in [-1,1]}.
\end{aligned} \tag{A.7}$$

$$\begin{aligned}
p(\sigma_v^2|\mathbf{y}) &= \int \int p(\boldsymbol{\beta}, \rho, \Sigma|\mathbf{y})d\boldsymbol{\beta}d\rho \\
&\propto |\Sigma|^{-1/2} \prod_{i=1}^n ((\sigma_i^2)^{-(a_i+1)}) \exp \left\{ -\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y})'\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y}) - \sum_{i=1}^n \frac{b_i}{\sigma_i^2} \right\} \\
&\propto (\sigma_v^2)^{-1/2} (\sigma_v^2)^{-(a_v+1)} \exp \left\{ -\frac{1}{2\sigma_v^2}(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y})'(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y}) - \frac{b_v}{\sigma_v^2} \right\} \\
&\propto (\sigma_v^2)^{-(a_v+3/2)} \exp \left\{ -\frac{1}{2\sigma_v^2}[(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y})'(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y}) + 2b_v] \right\} \\
\therefore \sigma_v^2|\mathbf{y} &\sim IG \left(a_v + \frac{1}{2}, \frac{1}{2}[(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y})'(\mathbf{y} - X\boldsymbol{\beta} - \rho W\mathbf{y}) + 2b_v] \right).
\end{aligned} \tag{A.8}$$

Since equation A.9, we observe the homoscedastic CAR model, where $\boldsymbol{\theta} = \{\rho, \boldsymbol{\beta}, \sigma^2\}$.

$$\begin{aligned}
p(\rho, \boldsymbol{\beta}, \sigma^2 | \mathbf{y}) &\propto \pi(\boldsymbol{\beta}, \rho, \sigma^2) f(\mathbf{y} | \boldsymbol{\theta}) \\
&= \frac{1}{2} \frac{1}{(2\pi\eta)^{k/2}} \exp \left\{ -\frac{1}{2\eta} (\boldsymbol{\beta})' (\boldsymbol{\beta}) \right\} \frac{b^a}{\Gamma(a)} (\sigma^2)^{-(a+1)} \exp \{ -b/\sigma^2 \} \\
&\quad (2\pi\sigma^2)^{-n/2} |I_n - \rho A|^{1/2} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta})' (I_n - \rho A) (\mathbf{y} - X\boldsymbol{\beta}) \right\} \\
&\propto \exp \left\{ -\frac{1}{2\eta} (\boldsymbol{\beta})' (\boldsymbol{\beta}) \right\} (\sigma^2)^{-(a+1)} \exp \{ -b/\sigma^2 \} \\
&\quad (\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta})' (I_n - \rho A) (\mathbf{y} - X\boldsymbol{\beta}) \right\} \\
&\propto (\sigma^2)^{-(a+1+n/2)} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta})' (I_n - \rho A) (\mathbf{y} - X\boldsymbol{\beta}) - \frac{1}{2\eta} \boldsymbol{\beta}' \boldsymbol{\beta} - \frac{2b}{\sigma^2} \right\}.
\end{aligned} \tag{A.9}$$

$$\begin{aligned}
p(\boldsymbol{\beta} | \mathbf{y}) &= \int \int p(\boldsymbol{\beta}, \rho, \sigma^2 | \mathbf{y}) d\sigma^2 d\rho \\
&\propto \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta})' (I_n - \rho A) (\mathbf{y} - X\boldsymbol{\beta}) - \frac{1}{2\eta} \boldsymbol{\beta}' \boldsymbol{\beta} \right\} \\
&\propto \exp \left\{ -\frac{\eta}{2\eta\sigma^2} (X\boldsymbol{\beta})' (I_n - \rho A) (X\boldsymbol{\beta}) - \frac{\sigma^2}{2\eta\sigma^2} \boldsymbol{\beta}' \boldsymbol{\beta} \right\} \\
&= \exp \left\{ -\frac{1}{2\eta\sigma^2} (\boldsymbol{\beta}' [\eta X'(I_n - \rho A) X + \sigma^2 I_n] \boldsymbol{\beta}) \right\} \\
&\therefore \boldsymbol{\beta} | \mathbf{y} \sim N_k(0, \eta\sigma^2 [\eta X'(I_n - \rho A) X + \sigma^2 I_n]^{-1}).
\end{aligned} \tag{A.10}$$

$$\begin{aligned}
p(\rho | \mathbf{y}) &= \int \int p(\boldsymbol{\beta}, \rho, \sigma^2 | \mathbf{y}) d\boldsymbol{\beta} d\sigma^2 \\
&\propto \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta})' (I_n - \rho A) (\mathbf{y} - X\boldsymbol{\beta}) \right\} I_{\rho \in [-1, 1]} \\
&\propto \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y}' (I_n - \rho A) \mathbf{y} + (X\boldsymbol{\beta})' (I_n - \rho A) (X\boldsymbol{\beta})) \right\} I_{\rho \in [-1, 1]} \\
&\propto \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y}' (-\rho A) \mathbf{y} - (X\boldsymbol{\beta})' (\rho A) (X\boldsymbol{\beta})) \right\} I_{\rho \in [-1, 1]} \\
&= \exp \left\{ \frac{\rho}{2\sigma^2} (\mathbf{y}' A \mathbf{y} + (X\boldsymbol{\beta})' A X \boldsymbol{\beta}) \right\} I_{\rho \in [-1, 1]} \\
&\therefore \rho | \mathbf{y} \sim \text{Exp} \left(\frac{1}{2\sigma^2} (\mathbf{y}' A \mathbf{y} + (X\boldsymbol{\beta})' A X \boldsymbol{\beta}) \right) I_{\rho \in [0, 1]}.
\end{aligned} \tag{A.11}$$

$$\begin{aligned}
p(\sigma^2|\mathbf{y}) &= \int \int p(\boldsymbol{\beta}, \rho, \sigma^2|\mathbf{y})d\boldsymbol{\beta}d\rho \\
&\propto (\sigma^2)^{-(a+1+n/2)} \exp \left\{ -\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta})'(I_n - \rho A)(\mathbf{y} - X\boldsymbol{\beta}) - \frac{2b}{2\sigma^2} \right\} \\
&= (\sigma^2)^{-(a+1+n/2)} \exp \left\{ -\frac{1}{\sigma^2} \left[\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta})'(I_n - \rho A)(\mathbf{y} - X\boldsymbol{\beta}) + b \right] \right\} \\
&\therefore \sigma^2|\mathbf{y} \sim IG \left(a + \frac{n}{2}, \frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta})'(I_n - \rho A)(\mathbf{y} - X\boldsymbol{\beta}) + b \right).
\end{aligned} \tag{A.12}$$

Since equation A.13, we observe the heteroscedastic CAR model, where $\boldsymbol{\theta} = \{\rho, \boldsymbol{\beta}, \Sigma\}$.

$$\begin{aligned}
p(\rho, \boldsymbol{\beta}, \Sigma|\mathbf{y}) &\propto \pi(\boldsymbol{\beta}, \rho, \sigma_i^2)f(\mathbf{y}|\boldsymbol{\theta}) \\
&= \frac{1}{2} \frac{1}{(2\pi\eta)^{k/2}} \exp \left\{ -\frac{1}{2\eta}(\boldsymbol{\beta})'(\boldsymbol{\beta}) \right\} \prod_{i=1}^n \left(\frac{b_i^{a_i}}{\Gamma(a_i)} (\sigma_i^2)^{-(a_i+1)} \exp\{-b_i/\sigma_i^2\} \right) \\
&(2\pi)^{-n/2} |I_n - \rho A|^{1/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta})'(I_n - \rho A)\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta}) \right\} \\
&\propto \exp \left\{ -\frac{1}{2\eta}(\boldsymbol{\beta})'(\boldsymbol{\beta}) \right\} \prod_{i=1}^n ((\sigma_i^2)^{-(a_i+1)} \exp\{-b_i/\sigma_i^2\}) \\
&|I_n - \rho A|^{1/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta})'(I_n - \rho A)\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta}) \right\} \\
&\propto |\Sigma|^{-1/2} \prod_{i=1}^n ((\sigma_i^2)^{-(a_i+1)}) \exp \left\{ -\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta})'(I_n - \rho A)\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta}) - \frac{1}{2\eta}\boldsymbol{\beta}'\boldsymbol{\beta} - \sum_{i=1}^n \frac{b_i}{\sigma_i^2} \right\}.
\end{aligned} \tag{A.13}$$

$$\begin{aligned}
p(\boldsymbol{\beta}|\mathbf{y}) &= \int \int p(\boldsymbol{\beta}, \rho, \Sigma|\mathbf{y})d\Sigma d\rho \\
&\propto \exp \left\{ -\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta})'(I_n - \rho A)\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta}) - \frac{1}{2\eta}\boldsymbol{\beta}'\boldsymbol{\beta} \right\} \\
&\propto \exp \left\{ -\frac{1}{2}(X\boldsymbol{\beta})'(I_n - \rho A)\Sigma^{-1}(X\boldsymbol{\beta}) - \frac{1}{2\eta}\boldsymbol{\beta}'\boldsymbol{\beta} \right\} \\
&= \exp \left\{ -\frac{1}{2\eta}(\boldsymbol{\beta}' [\eta X'(I_n - \rho A)\Sigma^{-1}X + I_n] \boldsymbol{\beta}) \right\} \\
&\therefore \boldsymbol{\beta}|\mathbf{y} \sim N_k (0, \eta[\eta X'(I_n - \rho A)\Sigma^{-1}X + I_n]^{-1}).
\end{aligned} \tag{A.14}$$

$$\begin{aligned}
p(\rho|\mathbf{y}) &= \int \int p(\boldsymbol{\beta}, \rho, \Sigma|\mathbf{y}) d\boldsymbol{\beta} d\Sigma \\
&\propto \exp \left\{ -\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta})'(I_n - \rho A)\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta}) \right\} I_{\rho \in [-1,1]} \\
&\propto \exp \left\{ -\frac{1}{2}(\mathbf{y}'(I_n - \rho A)\Sigma^{-1}\mathbf{y} + (X\boldsymbol{\beta})'(I_n - \rho A)\Sigma^{-1}(X\boldsymbol{\beta})) \right\} I_{\rho \in [-1,1]} \quad (\text{A.15}) \\
&\propto \exp \left\{ -\frac{\rho}{2}(-\mathbf{y}'A\Sigma^{-1}\mathbf{y} - (X\boldsymbol{\beta})'A\Sigma^{-1}(X\boldsymbol{\beta})) \right\} I_{\rho \in [-1,1]} \\
&\therefore \rho|\mathbf{y} \sim \text{Exp} \left(\frac{1}{2}(\mathbf{y}'A\Sigma^{-1}\mathbf{y} + (X\boldsymbol{\beta})'A\Sigma^{-1}(X\boldsymbol{\beta})) \right) I_{\rho \in [0,1]}.
\end{aligned}$$

$$\begin{aligned}
p(\sigma_v^2|\mathbf{y}) &= \int \int p(\boldsymbol{\beta}, \rho, \Sigma|\mathbf{y}) d\boldsymbol{\beta} d\rho \\
&\propto |\Sigma|^{-1/2} \prod_{i=1}^n ((\sigma_i^2)^{-(a_i+1)}) \exp \left\{ -\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta})'(I_n - \rho A)\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta}) - \sum_{i=1}^n \frac{b_i}{\sigma_i^2} \right\} \\
&\propto (\sigma_v^2)^{-1/2} (\sigma_v^2)^{-(a_v+1)} \exp \left\{ -\frac{1}{2\sigma_v^2}(\mathbf{y} - X\boldsymbol{\beta})'(I_n - \rho A)(\mathbf{y} - X\boldsymbol{\beta}) - \frac{b_v}{\sigma_v^2} \right\} \\
&\propto (\sigma_v^2)^{-(a_v+3/2)} \exp \left\{ -\frac{1}{2\sigma_v^2}[(\mathbf{y} - X\boldsymbol{\beta})'(I_n - \rho A)(\mathbf{y} - X\boldsymbol{\beta}) + 2b_v] \right\} \\
&\therefore \sigma_v^2|\mathbf{y} \sim IG \left(a_v + 1/2, \frac{1}{2}[(\mathbf{y} - X\boldsymbol{\beta})'(I_n - \rho A)(\mathbf{y} - X\boldsymbol{\beta}) + 2b_v] \right) \quad (\text{A.16})
\end{aligned}$$