POLITECNICO
MILANO 1863

Bayesian Models for Prediction on Real-Estate Data

Bonetti, Lima, Panzeri, Remondina, Ruggieri, Zanin

February 14, 2023

Table of contents

Project goal and dataset

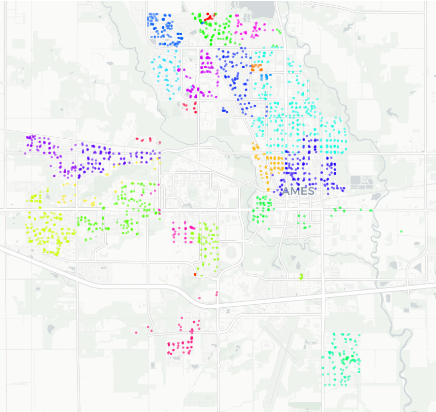


Figure: Houses in the dataset colored by neighbourhood

Model Averaging

Motivation

- Candidate models: CAR, Linear Mixed Effects and Gaussian Markov Random Fields
- We want to avoid overconfidence in the choice of the model structure
- We follow a principled approach for Model Averaging

Richard McElreath. *Statistical rethinking, a Bayesian course with examples in R and Stan*, pages 195–205.
2015

McElreath's WAIC Model Averaging

$$\forall i \in \{\mathsf{CAR}, \mathsf{LME}, \mathsf{GMRF}\}$$

$$\mathrm{dWAIC}_i = \mathrm{WAIC}_i - \min_{j \in \{\mathsf{CAR},\mathsf{LME},\mathsf{GMRF}\}} \mathrm{WAIC}_j$$

$$w_i = \frac{\exp\left(-\frac{1}{2}\mathrm{dWAIC}_i\right)}{\sum_{j=1}^{m} \exp\left(-\frac{1}{2}\mathrm{dWAIC}_j\right)}$$

$$Y_{\text{pred, ENSEMBLE}} = w_{\text{CAR}} Y_{\text{pred, CAR}} + w_{\text{LME}} Y_{\text{pred, LME}} + w_{\text{GMRF}} Y_{\text{pred, GMRF}}$$

Richard McElreath. *Statistical rethinking, a Bayesian course with examples in R and Stan*, pages 195–205. 2015

Conditionally Autoregressive Model

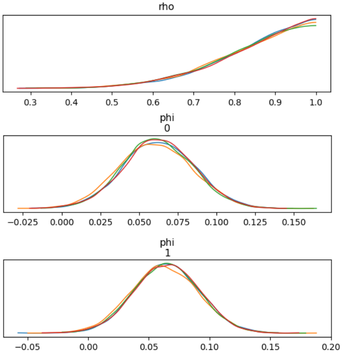
Model definition with multi-level Leroux prior

$$\forall i = 1, ..., p \qquad \forall j = 1, ..., N \qquad \forall k = 1, ..., K$$

$$\begin{aligned}
Y_j | b, \underline{\beta}, \text{Neigh}_j, \underline{x}_j, \underline{\phi}, \nu^2 &\sim_{\text{ind}} N(b + \underline{\beta}^T \underline{x}_j + \phi_{\text{Neigh}_j}, \nu^2) \\
\beta_i &\sim_{iid} N(0, 10) \\
b &\sim N(4.5, 0.01) \\
\nu^2 &\sim \text{Inv-Gamma}(0.1, 0.1) \\
\phi_k \mid \phi_{-k} &\sim \mathrm{N}\left(\frac{\rho \sum_{l=1}^K a_{kl} \phi_l}{\rho \sum_{l=1}^K a_{kl} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{l=1}^K a_{kl} + 1 - \rho} \right) \\
\tau^2 &\sim \text{Inv-Gamma}(0.1, 0.1) \\
\rho &\sim \text{Uniform}(0, 1) \\
[a_{kl}]_{k,l \in \{1,...,K\}} &\text{ is the neighbourhood adjacency matrix}
\end{aligned}$$

Model fitting and evaluation

	Mean	StdDev	5%	50%	95%	N_{eff}	$\hat{R}$
ρ	0.8483	0.1181	0.6164	0.8745	0.9889	16696	1.001
ϕ_0	0.0620	0.0233	0.0239	0.0617	0.1005	2000	1.002
ϕ_1	0.0642	0.0271	0.0199	0.0642	0.1089	3057	1.001



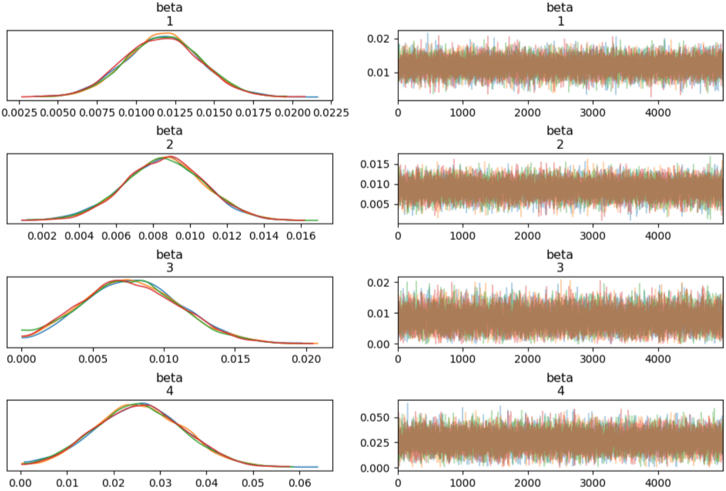
Linear Mixed Effects Model

Model definition

$$\forall i = 1, ..., p \qquad \forall j = 1, ..., N \qquad \forall k = 1, ..., K$$

$$\begin{aligned}
Y_j \,|\, \underline{X}_j, \text{Neigh}_j, \underline{\beta}, \underline{\gamma}, \sigma &\overset{\text{ind}}{\sim} \mathcal{N}(\underline{\beta}^T \underline{X}_j + \underline{\gamma}_{\text{Neigh}_j}, \sigma^2) \\
\beta_i \,|\, \tau &\overset{\text{iid}}{\sim} \text{half-}\mathcal{N}(0, \tau^2) \\
\gamma_k \,|\, \xi &\overset{\text{iid}}{\sim} \mathcal{N}(0, \xi^2) \\
\sigma^2 &\sim \text{InvGamma}(1, 30) \\
\tau^2 &\sim \text{InvGamma}(1, 0.05) \\
\xi^2 &\sim \text{InvGamma}(1, 0.05)
\end{aligned}$$

Fitting


beta
1
0.0025 0.0050 0.0075 0.0100 0.0125 0.0150 0.0175 0.0200 0.0225
beta
1
0.02
0.01
0
1000
2000
3000
4000
beta
2
0.002 0.004 0.006 0.008 0.010 0.012 0.014 0.016
beta
2
0.015
0.010
0.005
beta
3
0.000 0.005 0.010 0.015 0.020
beta
3
0.02
0.01
0.00
beta
4
0.00 0.01 0.02 0.03 0.04 0.05 0.06
beta
4
0.050
0.025
0.000

Gaussian Markov Random Fields Model

Model definition

$$
\begin{aligned}
Y_j \mid \underline{\beta}, w, \sigma_e^2 &\sim \mathcal{N}\left(\underline{X}_j^T \underline{\beta} + w(s), \sigma_e^2\right) \\
\beta_i &\sim_{\text{iid}} \mathcal{N}(0,1) \\
\sigma_e^2 &\sim \text{Log-Gamma}(0.1, 0.1) \\
w(s) \mid \theta &\sim \mathcal{N}\left(0, \Sigma(\theta)\right) \\
w(s) \text{ is the solution of } \left(\kappa^2 - \Delta\right)^{\alpha/2} (\tau w(s)) &= \mathcal{W}(s) \\
\log(\kappa) &= log(\kappa_0) - \theta_2 \\
\log(\tau) &= log(\tau_0) - \theta_1 + \nu\theta_2 \\
\theta_1 &\sim \mathcal{N}(0,1) \\
\theta_2 &\sim \mathcal{N}(0,10) \\
\Sigma(s;\theta) &= \tfrac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} (\kappa\|s\|)^{\nu} K_{\nu}(\kappa\|s\|)
\end{aligned}
$$

Fitting

- We fitted the model using the Integrated Nested Laplace Approximation.
- The calibration of hyperparameters of the covariance function has been done fixing $\alpha = 2$, $\nu = 1$, $\kappa_0 = \frac{\sqrt{8\nu}}{\rho}$ and $\tau_0 = \frac{1}{\sqrt{4\pi}\kappa_0\sigma}$ and choosing $\rho = 4.6$ and $\sigma = 1.5$ in order to minimize the WAIC.
- This method computes an approximation of the marginal posterior distributions for all the parameters and hyperparameters.

Fitting

Given the gaussian vector $\underline{\xi} = (\underline{\beta}, \underline{w})$ we can approximate the posterior distribution of $\underline{\theta} = (\theta_1, \theta_2)$:

$$\pi(\underline{\theta}|\underline{y}) \propto \left.\frac{\pi(\underline{\theta},\underline{\xi},\underline{y})}{\pi(\underline{\xi}|\underline{\theta},\underline{y})}\right|_{\underline{\xi}=\underline{\xi}^*(\underline{\theta})} \approx \left.\frac{\pi(\underline{\theta},\underline{\xi},\underline{y})}{\tilde{\pi}(\underline{\xi}|\underline{\theta},\underline{y})}\right|_{\underline{\xi}=\underline{\xi}^*(\underline{\theta})}$$

where $\tilde{\pi}(\xi|\underline{\theta}, \underline{y})$ is a gaussian approximation of $\pi(\underline{\xi}|\underline{\theta}, \underline{y})$ and $\xi^*(\underline{\theta}) = (\pi(\underline{\xi}|\underline{\theta}, \underline{y})$.

Using this method we can compute the approximation of the posterior of the parameters:

$$\pi(\theta_k|\underline{y}) \approx \int \tilde{\pi}(\underline{\theta}|\underline{y})d\underline{\theta}_{-k}$$

$$\pi(\xi_j|\underline{y}) \approx \int \tilde{\pi}(\xi_j|\underline{\theta}, \underline{y})\tilde{\pi}(\underline{\theta}|\underline{y})d\underline{\theta}$$

Model Assessment

We computed the ensemble model from the three fitted models

$$\text{WAIC}_{\text{CAR}} = -5717 \quad \text{WAIC}_{\text{LME}} = -3929 \quad \text{WAIC}_{\text{GMRF}} = -5526$$
$$w_{\text{CAR}} \simeq 1 \qquad w_{\text{LME}} \simeq 0 \qquad w_{\text{GMRF}} \simeq 0$$

We conclude that the ensemble model coincides with the CAR model; its performance on the test set is

$$\sqrt{\text{MSE}_{\text{ensemble}}} \simeq 27.000 \ \$$$

Results

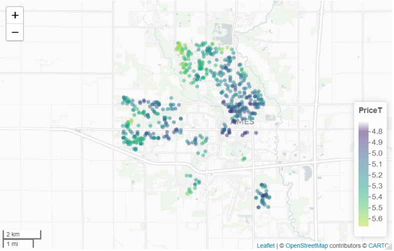


Figure: Prediction on the test set

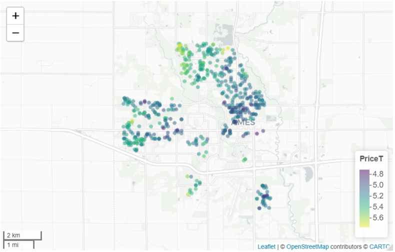


Figure: True log(price) of the test set

Bibliography

References I

[1] Dean De Cock. Ames, iowa: Alternative to the boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3), 2011.

[2] Alessandra Guglielmi and Mario Beraha. Lecture notes in Bayesian Statistics, 2022.

[3] Duncan Lee. CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, 55(13):1–24, 2013.

References II

[4] Brian G. Leroux, Xingye Lei, and Norman Breslow. Estimation of disease rates in small areas: A new mixed model for spatial dependence. In M. Elizabeth Halloran and Donald Berry, editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 179–191, New York, NY, 2000. Springer New York.

[5] Sara Martino and Andrea Riebler. Integrated Nested Laplace Approximations (INLA), 2019.

[6] Richard McElreath. *Statistical rethinking, a Bayesian course with examples in R and Stan*, pages 195–205. 2015.

[7] Håvard Rue, Finn Lindgren, Janet van Niekerk, and Elias Krainski. R-INLA Project - Documentation.

Model evaluation