Construction of a proper prior for a Bayesian Envelope Model

Costruzione di una prior propria per un modello Envelope bayesiano

Andrea Mascaretti and Antonio Canale

Abstract Envelope models are multivariate linear regression techniques that aims at reducing the variance of the estimator. Bayesian envelopes allow to quantify the uncertainty of inference by means of the posterior distribution. In this work, we construct a proper prior distribution and compare it to the existing literature. A prior sensitivity analysis is conducted, yielding similar results.

Abstract *I modelli envelope sono una particolare tipologia di regressione lineare multivariata finalizzata a ridurre la varianze degli stimatori. La formulazione bayesiana di questi modelli consente di quantificare direttamente l'incertezza degli stimatori mediante l'analisi della posterior. In questo lavoro, proponiamo una prior propria per il modello e valutiamo l'impatto della prior sull'inferenza, ottenendo risultati comparabili alle proposte presenti in letteratura.*

Key words: envelope models, bayesian statistics

1 Response Envelopes

Envelopes [2, 1] are a class of models aimed at increasing the efficiency of multivariate regression by exploiting the relations between response and predictors that affect the accuracy of the results and are not taken into account by standard methods. Within the usual multivariate regression setting, the expected value of a random variable $Y \in R^r$ is given a functional form such that we get

Andrea Mascaretti
University of Padova, Via Cesare Battisti, 241, 35121, Padova (PD), Italy, e-mail: mascaretti@stat.unipd.it

Antonio Canale
University of Padova, Via Cesare Battisti, 241, 35121, Padova (PD), Italy, e-mail: antonio.canale@unipd.it

$$Y_i = \mu + \beta X_i + \varepsilon_i, \;\; i = 1, \dots, n, \tag{1}$$

where $\{X_i\}_{i=1}^n$ is a sequence of non-stochastic vectors, with $X_i \in R^p$ for $i = 1, \dots, n$, the errors are independent and identically distributed multivariate normal vectors with zero mean and covariance Σ, $\mu \in R^r$ is an unknown vector of intercepts and $\beta \in M_{(r,p)}$ (where $M_{(a,b)}$ denotes the space of real matrices of dimensions (a,b)) is the unknown vector of regression coefficients. For simplicity (and without loss of generality), we assume that the predictors are centred, $\sum_{i=1}^n X_i = 0$. Moreover, let Y be the $(n \times r)$ matrix of rows $(Y_i - \bar{Y})^T$, where $\bar{Y}$ is the sample mean, and Y_0 be the non-centred matrix. In a similar fashion, let $X = \left\{X_i^T\right\}$ be the matrix of the predictors, $S_{Y,X} = n^{-1} Y^T X$ and $S_X = n^{-1} X^T X$. The maximum likelihood estimator,

$$\hat{\beta} = S_{Y,X} S_X^{-1}, \tag{2}$$

is incidentally equal the ordinary least squares estimator. From Eq. 2, we notice that this is akin to performing r separate univariate regressions: one for every element of Y on X. Inference on $\beta_{j,k}$, the (j,k)th element of β is the same we would obtain by constructing a univariate model. The model in Eq. 1 becomes operational when inference is conducted simultaneously on different rows of β or various elements of Y jointly.

The intuition behind envelope models is that there might be linear combinations of the response vectors whose distribution is invariant with respect to the non-stochastic predictors. Explicitly modelling for this property allows to obtain estimator whose variance is reduced. We call such linear combinations of Y X-invariant. Notice that for a linear transformation $G \in M_{(r,q)}$, with $q \leq r$, if $G^T Y$ is invariant, then also $A^T G^T Y$ has the same property for any non-stochastic matrix $A \in M_{(q,q)}$. In other words, only span(G) is identifiable.

From a mathematical point of view, this is equivalent to assuming the existence of two matrices Γ and Γ_0 such that $O = [\Gamma \; \Gamma_0]$ is orthogonal. We obtain

1. $\Gamma_0^T Y | X \sim \Gamma_0^T Y$
2. $\Gamma^T Y \perp \Gamma_0^T Y | X$

The conditions above entail that span$(\beta) \subseteq$ span(Γ) and $\Sigma = \Sigma_1 + \Sigma_2 = P_\Gamma \Sigma P_\Gamma + Q_\Gamma \Sigma Q_\Gamma$, where $P_{(\cdot)}$ is the orthogonal projector operation on a space and $Q_{(\cdot)} = I - P_{(\cdot)}$ is the projection on the orthogonal space. In this scenario, span(Γ) is a reducing subspace of Σ ([2]). The Σ-envelope of $\mathscr{B} =$ span(β), $\mathscr{E}_\Sigma(\mathscr{B})$, is the smallest reducing subspace of Σ that contains $\mathscr{B}$.

Model in Eq. 1 can be rewritten as

$$Y_i = \mu + \Gamma \eta X_i + \varepsilon, \tag{3}$$

where $\beta = \Gamma\eta, \Gamma \in M_{(r,u)}$ is an orthogonal basis of $\mathscr{E}_\Sigma(\mathscr{B})$ and u is the dimension of the envelope $\mathscr{E}_\Sigma(\mathscr{B})$. Moreover, the variance is $\Sigma = \Sigma_1 + \Sigma_2 = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T$, where $\Omega \in M_{(u,u)}$ and $\Omega_0 \in M_{(r-u,r-u)}$ are two diagonal matrices carrying the coordinate information with respect to the basis Γ and Γ_0.

1.1 Bayesian Envelopes

The only contribution, to the best of our knowledge, on Bayesian envelopes models is [3]. The rationale behind Bayesian envelopes is that it allows to quantify the uncertainty of the predictions by computing the posterior distribution (as opposed to bootstrap or asymptotic considerations), as well as extending the model to the cases where $n < r$. Moreover, prior information can be incorporated into the learning process, be it on the values of the parameters or to induce sparsity or other desirable properties. As for the selection of u, the dimension of the envelope, [3] adopt a Deviance Information Criterion to obtain the best value, in lieu of the Likelihood Ratio Tests used within the frequentist framework. The interest in obtaining a proper prior distribution for a Bayesian envelope stems from the fact that this is a prerequisite to extend it to more complex scenarios, such as mixtures or nonparametric formulations.

The prior distribution is defined on the parameters $(\mu, \eta, (\Gamma, \Gamma_0), \Omega, \Omega_0)$. Notice that, for identifiability, we constrain Ω and Ω_0 to be diagonal matrices with entries disposed in decreasing order. This is equivalent to post-multiplying Γ and Γ_0 by the matrices of eigenvectors of the original Ω and Ω_0. From a mathematical point of view, this is equivalent to fix Γ and Γ_0 to be bases of the envelope and, thus, as elements of a subset of a Stiefel manifold restricted to have that the maximum element for each column as positive sign, denoted by $S^+_{(\cdot,\cdot)}$. In this respect, we notice that the Stiefel manifold of arbitrary finite dimensions (a,a) is a compact unimodular group with a unique Haar measure, which induces a measure on $S_{(a,b)}$ and $S^+_{(a,b)}$.

The parameter space is then given by $M_{(r,1)} \times M_{(u,p)} \times S^+_{(r,r)} \times O_u \times O_{r-u}$, where O_a is the set of diagonal matrices of dimension a with entries disposed in decreasing order.

We define the prior on the parameters as follows:

1. μ is set to be independent from the other parameters. We endow it with a multivariate normal distribution, so that $\pi(\mu) = \mathscr{N}_r(\mu_0, \Sigma_0)$,
2. The conditional prior on η is a matrix normal:

$$\pi(\eta | (\Gamma, \Gamma_0, \Omega, \Omega_0)) = \mathscr{N}_{(u,p)}\left(\Gamma^T, \Omega, C^{-1}\right),$$

   where C^{-1} is a positive definite matrix in $M_{(p,p)}$.
3. The prior on $O = (\Gamma, \Gamma_0)$ is a matrix Bingham distribution with parameters G and D, where G is a positive semi-definite matrix in $M_{(r,r)}$ and D is in O_r with positive entries. Thus, $\pi(O) = \mathscr{B}_{(r,r)}\left(G, D^{-1}\right)$. The density is proportional to $\exp\left\{(-1/2)\operatorname{tr}\left(D^{-1}O^T GO\right)\right\}$
4. Denoting by ω and ω_0 the diagonal vectors of, respectively, Ω and Ω_0, we assume that, a priori, they are distributed as order statistics of u and $r-u$ independent and identically distributed observations from Inverse-Gamma distributions of shape and rate parameters α, ψ and α_0, ψ_0.

Notice that the main difference between our work and [3] is the prior on μ. From a computational point of view, this means that the structure of the Gibbs sampler is similar, the only difference being the structure of the full-conditional for μ, which can be easily computed to be of the form

$$\pi\left(\mu|\eta,\left(\Gamma,\Gamma_0\right),\omega,\omega_0,Y\right)=\mathscr{N}_r\left(\mu_c,\Sigma_c\right), \tag{4}$$

where

$$\Sigma_c=\left(\Sigma_0^{-1}+\left(\frac{\Sigma}{n}\right)^{-1}\right)^{-1},$$

and

$$\mu_c=\Sigma_c\left(\Sigma_0^{-1}\mu_0+\left(\frac{\Sigma}{n}\right)^{-1}\bar{Y}\right).$$

Notice that the Harris ergodicity of the chain is also a straightforward extension of [3].

2 Simulation and Data Analysis

We now perform a test for different values of the prior distribution on a synthetic dataset. The aim is to assess the sensitivity with respect to the choice of the hyper-parameters. We generated $n=100$ data points from a normal distribution with zero mean and identity matrix as covariance. We set $u=1$, $p=2$, $r=3$. The parameters are defined as follows:

1. $\mu=(12,12,12)$
2. $\omega=6.2$
3. $\omega_0=(3.2,1.4)$
4. $O=I_r$

and Y_i are randomly drawn a multivariate normal with mean $\mu+\Gamma\eta X_i$ and covariance $\Gamma\Omega\Gamma^T+\Gamma_0\Omega_0\Gamma_0^T$. As for the hyperparamaters, we distinguish between three cases. We focus on μ as it is the most relevant change we make. In the first case, we use a weakly informative proper prior with $\mu_0=(0,0,0)$ and $\Sigma_0=\kappa I_r$, with $\kappa=10$. In the second test case, we set $\mu_0=\bar{Y}$ and $\Sigma=I_r$. Finally, we consider the improper prior as in [3]. The other parameters are set as follows: $C=I_p$, $D=I_r$, $G=I_r$, $\alpha=3$, $\psi=3$, $\alpha_0=3$ and $\psi_0=3$.

For each case study, we run a Gibbs sampler for 1000 iterations, with a burn in of 300. The initialisation for each chain was from the same random point in the parameter space.

Results for the three components of μ are reported, respectively in Tables 1, 2, and 3.

We see that even though the empirical and the noninformative priors lead to somewhat closer posterior estimates, the effect of placing a weakly informative

prior also yields posterior higher density intervals that are in line with the other two classes of prior distributions. However, the true advantage of a proper prior is that it allows for extending the model to more complex settings. The fact that it yields similar results notwithstanding different hyperparameters is certainly encouraging, although, as always, some care should be put in their refinement.

Table 1 Posterior inference for $\mu = (\mu_1, \mu_2, \mu_3)$ with a weakly informative prior: posterior higher density interval (HDI) are reported.

Parameter	Mean	3% HDI	97% HDI
μ_1	12.67	12.005	13.359
μ_2	12.043	11.715	12.324
μ_3	12.097	11.882	12.313

Table 2 Posterior inference for $\mu = (\mu_1, \mu_2, \mu_3)$ with a empirical prior: posterior higher density interval (HDI) are reported.

Parameter	Mean	3% HDI	97% HDI
μ_1	12.832	12.1	13.49
μ_2	12.071	11.746	12.361
μ_3	12.117	11.903	12.307

Table 3 Posterior inference for $\mu = (\mu_1, \mu_2, \mu_3)$ with a noninformative prior: posterior higher density interval (HDI) are reported.

Parameter	Mean	3% HDI	97% HDI
μ_1	12.85	12.246	13.425
μ_2	12.07	11.74	12.364
μ_3	12.121	11.916	12.286

3 Conclusions

In this work, we have constructed a proper prior distribution for a Bayesian envelope model. We carried out an assessment of the prior sensitivity on a simple test case, obtaining that the choice of the hyperparameters for the parameter μ yield similar results in the three cases studied: a weakly informative, an empirical one and a non-informative prior. As such, this model endowed with a proper prior can be extended

to more complex scenarios. For instance, it can be used as a building block for a mixture model, as opposed to the a model with an improper prior. This is especially true given the stability of posterior HDIs in three cases.

References

1. R. D. Cook. *An Introduction to Envelopes: Dimension Reduction for Efficient Estimation in Multivariate Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, 1st edition edition, 2018.
2. R. D. Cook, B. Li, and F. Chiaromonte. Envelope Models for Parsimonious and Efficient Multivariate Linear Regression. *Statistica Sinica*, 20(3):927–960, 2010.
3. K. Khare, S. Pal, and Z. Su. A Bayesian approach for envelope models. *The Annals of Statistics*, 45(1):196–222, Feb. 2017.