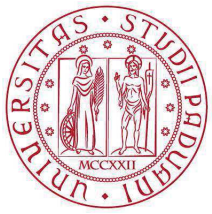


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXX

Developments in Bayesian Hierarchical Models and Prior Specification with Application to Analysis of Soccer Data

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Nicola Torelli

Co-supervisore: Prof. Francesco Pauli

Dottorando: Leonardo Egidi

31 Ottobre 2017

Abstract

In the recent years the challenge for new prior specifications and for complex hierarchical models became even more relevant in Bayesian inference. The advent of the Markov Chain Monte Carlo techniques, along with new probabilistic programming languages and new algorithms, extended the boundaries of the field, both in theoretical and applied directions. In the present thesis, we address theoretical and applied tasks. In the first part we propose a new class of prior distributions which might depend on the data and specified as a mixture between a noninformative and an informative prior. The generic prior belonging to this class provides less information than an informative prior and is more likely to not dominate the inference when the data size is small or moderate. Such a distribution is well suited for robustness tasks, especially in case of informative prior misspecification. Simulation studies within the conjugate models show that this proposal may be convenient for reducing the mean squared errors and improving the frequentist coverage. Furthermore, under mild conditions this class of distributions yields some other nice theoretical properties.

In the second part of the thesis we use hierarchical Bayesian models for predicting some soccer quantities and we extend the usual match goals' modeling strategy by including the bookmakers' information directly in the model. Posterior predictive checks on in-sample and out-of sample data show an excellent model fit, a good model calibration and, ultimately, the possibility for building efficient betting strategies.

Sommario

Negli ultimi anni la sfida per la specificazione di nuove distribuzioni a priori e per l'uso di complessi modelli gerarchici è diventata ancora più rilevante all'interno dell'inferenza Bayesiana. L'avvento delle tecniche Markov Chain Monte Carlo, insieme a nuovi linguaggi di programmazione probabilistici, ha esteso i confini del campo, sia in direzione teorica che applicata. Nella presente tesi ci dedichiamo a obiettivi teorici e applicati. Nella prima parte proponiamo una nuova classe di distribuzioni a priori che dipendono dai dati e che sono specificate tramite una mistura tra una a priori non informativa e una a priori informativa. La generica distribuzione appartenente a questa nuova classe fornisce meno informazione di una priori informativa e si candida a non dominare le conclusioni inferenziali quando la dimensione campionaria è piccola o moderata. Tale distribuzione è idonea per scopi di robustezza, specialmente in caso di scorretta specificazione della distribuzione a priori informativa. Alcuni studi di simulazione all'interno di modelli coniugati mostrano che questa proposta può essere conveniente per ridurre gli errori quadratici medi e per migliorare la copertura frequentista. Inoltre, sotto condizioni non restrittive, questa classe di distribuzioni dà luogo ad alcune altre interessanti proprietà teoriche.

Nella seconda parte della tesi usiamo la classe dei modelli gerarchici Bayesiani per prevedere alcune grandezze relative al gioco del calcio ed estendiamo l'usuale modellazione per i goal includendo nel modello un'ulteriore informazione proveniente dalle case di scommesse. Strumenti per sondare a posteriori la bontà di adattamento del modello ai dati mettono in luce un'ottima aderenza del modello ai dati in possesso, una buona calibrazione dello stesso e suggeriscono, infine, la costruzione di efficienti strategie di scommesse per dati futuri.

A chi crede nel progresso, ma non troppo.

Acknowledgements

Being a statistician, I need to summarize my three-years path through some quantities: with a lot of imagination, whoever is mentioned here may be seen as a *sufficient statistic* for my PhD project.

I want to thank my supervisors, Nicola and Francesco, simply for being as they are: professional, brilliant, familiar and funny. The reasons why I continued to study with my PhD project and I enjoy to explore statistical concepts are based on their motivations and on their daily support. *They gave me the Force*, and I will be always grateful to them for this.

I would like to thank all the professors and the researchers of the Department of Statistics, University of Padova, for being professional and helpful. Particularly, I thank prof. Nicola Sartori and prof. Monica Chiogna for being always available, brilliant and for their excellent role of guidance provided to the PhD students. A warm thank is also dedicated to the excellent work of Patrizia Piacentini, 'guardian angel' of each PhD student.

I will be always grateful to prof. Andrew Gelman and Jonah Sol Gabry for my amazing visiting period in New York, and for their priceless help. The latter is now also a good friend, and this is really important to me.

My XXX cycle colleagues deserve a deep hug for these years. In particular, without Umberto, Claudio, Ehsan and FIFA '14 I would be lost.

A special thank is dedicated to Susanna and Roberta for their human and statistical support across my period in DEAMS, for their trust on my skills and for belonging to my academic family.

Finally, I really want to thank all those guys which have improved my research with rich, wise and fruitful dialogues and suggestions: Stuart Coles, Gianluca Baio, Daniele Durante, Ben Goodrich and Bob Carpenter.

Of course, thank to my mother and my father: rather than a sufficient statistic, they represent my historical information. As archeologists, they already know how much important is the past.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The recent success of an algorithmic philosophy within Bayesian inference has had a lot of practical and theoretical efforts, which turn out to speed up the computational times, to improve the fit and the prediction power of the models and, in general, to broaden the set of the underlying assumptions. As the problem complexity grows, the need of new computational solutions is vital. The development of probabilistic programming languages as WinBUGS (Spiegelhalter et al., 2003) and Stan (Stan Development Team, 2016*b*,*a*) expanded and automatized some procedures for fitting the models, giving the possibility of manually setup the priors along with their hyperparameters and obtaining posterior estimates through automatic MCMC sampling.

The two topics that probably have been more deeply touched by the 'algorithmic revolution' are prior elicitation and hierarchical modeling, which are strongly connected. Eliciting the prior distribution is the milestone of Bayesian inference and is still one of the most debated tasks. Subjectivist Bayesians interpret the prior as an expert belief before observing the data (Garthwaite et al., 2005), while objective Bayesians (Berger et al., 2006) often push towards an automatic elicitation, without requiring external information, regardless of the noninformative nature of the prior. Furthermore, some authors have recently proposed approaches for eliciting the prior distribution using the data (Wasserman, 2000), while others tried to formalize data-dependent priors within Bayesian inference, either in terms of a new paradigm (Darnieder, 2011) or in terms of an approximation to a Bayesian model (Gelman, 2016*a*). With the advent of the MCMC techniques, the prior is not constrained to

be conjugate to the model any longer. As claimed by Gelman (2016*b*), we often need something in between a fully informative prior and a noninformative prior which is expected to give good results for any possible parameter value. In fact, priors should convey information, regularize and somehow restrict the parameter space. But in many instances some noninformative priors might have a big effect on the inferences: an example is given by the inverse gamma for the scale parameter in hierarchical models (Gelman et al., 2006). It is also not uncommon that an unrealistic informative prior would tend to convey too much information.

Hierarchical models often represent a tool flexible enough for describing complex problems. As suggested by Gelman et al. (2013), the common feature of such models is that the observed units y_{ij} are indexed by the statistical unit i in group j. In general, these observable outcomes are modeled conditionally on certain not observable parameters θ_j, viewed as drawn from a *population distribution*, which themselves are given a probabilistic (prior) distribution in terms of further parameters, known as *hyperparameters*. The data are then used for estimating relevant aspects of the population distribution. Hierarchical modeling is often appropriate for grasping complex data structures, as those provided by real problems. Furthermore, the population distribution may allow for different extents of dependence between the parameters and its use is often encouraged to avoid *overfitting*. Unlike for non-hierarchical models, it is in fact of common practice in the hierarchical framework using more parameters than data points.

Combining data and prior beliefs is one of the main task of a Bayesian statistician and the boundaries between the model and the data often turn out to overlap each other, as in hierarchical models. With a growing complexity carried by real problems and perhaps the need of more complicated models, the distinction in prior and likelihood could suffer from a rigidity extent. New models are required and new structures with hierarchical flavor are worth to be investigated through the development of new classes of algorithms.

In this thesis we deal with issues arising from prior elicitation and hierarchical modeling. Precisely, we focus on a particular class of prior distributions which might depend on the data, and we adopt hierarchical Bayesian models for predicting quantities related to soccer matches. This thesis may be then easily divided in two parts, one more theoretical and another one with more applied flavor.

In the first part we develop a new class of data-dependent priors. The idea behind our approach is that a Bayesian model is constituted by the pair prior-likelihood, and this view is coherent with those model checking approaches in the Bayesian setting in which the prior, as well as the likelihood, is seen as a potential source of model misspecification. In other words, we mantain with Gelman and Shalizi (2013) that a prior should be tested and is one of the assumption of the model.

We combine some existing insights and some separate theories about the use of the data in the prior: the adjusted-data-dependent paradigm (Darnieder, 2011), the approximation of a hierarchical model (Gelman, 2016*a*), and the specification of a model for the tuning parameter in the penalized likelihood approach. The new class consists of a mixture between a noninformative —baseline— prior distribution and an informative prior, where the mixture weights represent a sort of hyperparameter estimated through the data. This new formulation allows for assessing the robustness of a Bayesian model, usually achieved by the use of mixture priors, and the quantity of information provided by a prior distribution. The underlying idea is inspired by Gelman (2016*b*), who claims that we need something in between a '*wildly unrealistic in most settings prior informative distribution and a noninformative prior, feasible only in settings where data happen to be strongly informative about all parameters*". The amount of information carried by a prior distribution is relevant for us. In several frameworks the sample size is large enough to neutralize the impact of an informative prior. Conversely, when the sample size is 'small' it is not trivial to elicit an informative prior that does not dominate the inference. In our theoretical framework, we explicitely assume that the data size is likely to be not sufficiently large for neutralizing the impact of such a prior. Eliciting an informative prior distribution from historical data —as it is usual in medical studies, for instance— could result in a mismatch between the prior and the observed data, the so called *prior-data conflict* (Evans et al., 2006; Mutsvari et al., 2016).

Perhaps, our proposal merges together two separate approaches for eliciting a prior distribution: the mixture specification (Mutsvari et al., 2016; Berger and Berliner, 1986; Schmidli et al., 2014), and the use of the data in the prior formulation. Section 2.2 introduce different approaches for using data twice in Bayesian inference. In Section 2.3 we introduce the Mixture Data-dependent (MDD) prior class, and we describe the resampling algorithm and the natural procedure developed for estimating the mixture weights. Some theoretical results related to the hierarchical

approximation, the use of distribution-constant statistics along with the notion of effective sample size (Morita et al., 2008) are presented in Section 2.4. Simulation studies for the univariate conjugate models are performed in Section 2.5, whereas some nonstandard cases are briefly outlined in Section 2.6.

The second part of this thesis focuses on modeling some aspects of football (soccer) using Bayesian hierarchical models. In these last years the interest on sport modeling —both in terms of individual performance and team performance— is hugely growing. The call for prediction and description tools is urgent and several techniques have been developed with more or less success. In the first of our applied work on soccer, we develop three hierarchical Bayesian models for the player ratings provided by a popular Italian fantasy soccer game, used as proxies for the players' performance. Our central goals are to explore what can be accomplished with a simple freely available dataset (comprising only a few variables) for the 2015–2016 season in the top Italian league, Serie A, and to focus on a small number of interesting modeling and prediction questions that arise. Chapter 3 is devoted to the first applied work on soccer data. In Section 3.2 we take a brief overview on the fantasy game, describing the so called point scoring system; data and notation for the hierarchical models are presented in Section 3.3. The three models are proposed in Section 3.4, along with some strategies to deal with missing data which arise in this framework and with some considerations about the model identifiability. We present posterior estimates and a fake-data simulation in Section 3.5, while a graphical variety of posterior predictive checks along with some out-of sample predictions is proposed in Section 3.6.

Rather than modeling the individual performance in football, predicting the outcome of a football match has been of interest for many authors since the last decades of the XX century. According to the current literature, two choices are used for modeling the home and the away goals between two competing teams: two conditionally independent Poisson distributions (Maher, 1982; Baio and Blangiardo, 2010) or a bivariate Poisson distribution (Karlis and Ntzoufras, 2003; Dixon and Coles, 1997). Closely related to modeling the exact outcome, there is a huge literature regarding the bookmakers betting odds. It is empirically known that betting odds are the most accurate source of information for forecasting sports performances (Štrumbelj, 2014). As far as we know, no authors used the betting odds as a *part* of a statistical model for improving the predictive accuracy and the model fit. We try to fill the gap cre-

ating a bridge between the betting odds —and betting probabilities— on one hand and the statistical modeling on the other hand. Once we transform the betting odds into precise probabilities, we develop a procedure to (i) infer from these the implicit scoring intensities of the bookmakers (ii) use these implicit intensities directly in the conditionally independent Poisson model for the scores, according to a Bayesian perspective. In Chapter 4 we present the second applied work about soccer. The notion of betting odd and the transformation methods are presented in Section 4.2. In Section 4.3 we introduce the full model, along with the implicit scoring rates. The results and the predictive accuracy of the model on the top-four European leagues —Bundesliga, Premier League, La Liga and Serie A— are presented and discussed in Section 4.4 and summarized through posterior probabilities and graphical checks. Some preliminary betting strategies which reveal efficient profits are introduced in Section 4.5.

1.2 Main contributions of the thesis

- Development of a new class of data-dependent priors. Precisely:
  - the new class consists of a mixture between a noninformative —baseline— prior distribution and an informative prior, where the mixture weights represent a sort of hyperparameter estimated through the data. Rather than assigning the mixture weights a fixed value or an hyperprior distribution, we let them to incorporate data dependence and we treat the weight associated to the noninformative prior as a discrepancy measure between the data and the assumed informative prior;
  - we build our theoretical framework within the conjugate models and we formally prove that under some mild regularity conditions the information provided by a mixture distribution is never greater than the information of an informative prior distribution;
  - we perform some simulation studies which clearly show that this class may be well suited for reducing the mean squared errors and for improving the frequentist coverage;

  - we justify our class of priors as an approximation of a hierarchical model, as a prior conditioned in some particular cases on a distribution-constant statistic and as a sound alternative for specifying the tuning parameter in the penalized likelihood approach.

- Development of three hierarchical Bayesian models for predicting the player ratings provided by a popular Italian fantasy soccer game, used as proxies for the players' performance. We validate our models through graphical posterior predictive checks and we provide out-of-sample predictions for the second half of the season, using the first half as a training set. We use RStan to sample from the posterior distributions via Markov chain Monte Carlo.

- Development of a hierarchical Bayesian model for predicting the exact outcome of a football match using the past historical information and the weekly betting odds provided by the prominent bookmakers. We apply our procedure to the top-four European leagues —Bundesliga, Premier League, La Liga and Serie A— along with a variety of graphical posterior predictive checks for the model fit and a predictive accuracy analysis on hold-out data. Futhermore, we develop a betting strategy which is associated to profitable betting opportunities.

Chapter 2

Mixture Data-Dependent priors

2.1 Introduction

Prior elicitation is the core of every Bayesian analysis. In principle, the prior should represent the belief of the statistician before observing the data, but for several reasons in the last decades many attempts for including data information in the elicitation process have been proposed. Roughly speaking, the resulting data-dependent prior is just a prior that depends on the data and suffers from two main criticisms: data are used twice and the calculus of the Bayes' theorem may not be performed directly.

Despite this evident contravention of the Bayesian philosophy, many statisticians dealt with the double use of the data in Bayesian inference, and many others use data-dependent priors for complex models. However, as invoked by Wasserman (2000) almost twenty years ago, a theoretical justification for these distributions is missing and the need for data-dependent priors may become more common as the complexity for applied problems increases. Apparently, the call for the data-dependent Bayesians did not remain silent in these last years. As far as we can tell from reviewing the literature, we may recognize at least three frameworks for justifying the data-dependent approach within the Bayesian inference: interpreting these priors as an approximation of a hierarchical model through the estimation of some hyperparameters (Gelman, 2016*a*); the definition of an adjusted data-dependent paradigm allowing for the Bayes' Theorem computation (Darnieder, 2011); and the definition of a data-dependent prior as a measurable function from the data space $\mathcal{Y}^m$ to the set of priors $\mathcal{P}$ (Wasserman, 2000). In this chapter we propose a class of data-dependent prior distributions that

may be theoretically justified under all these frameworks. Moreover, the methodology presented in this chapter may be interpreted also in terms of a penalized likelihood framework (Cole et al., 2013) for regression models, where the penalty term is the kernel of a prior distribution and the weight of such penalization is not fixed in advance —as it happens for instance through cross-validation or empirical Bayes techniques.

The idea behind our approach is that a Bayesian model is constituted by the pair prior-likelihood. This view is coherent with those model checking approaches in the Bayesian setting in which the prior, as well as the likelihood, is seen as a potential source of model misspecification. In other words, unlike in the traditional paradigm of Bayesian inference where the only characteristic a prior must have to be justified is that it represents someone prior beliefs, we agree with Gelman and Shalizi (2013) that a prior should be tested and is one of the assumption of the model. If we take this kind of approach, that is, we admit that the prior can be (judged to be) misspecified, we imply that the prior is checked against the data and we may change it depending upon the results of this check (perhaps in an informal way: as it would occur if we visually inspect a PP plot and decide for a different prior not envisaged before). Thus, we are using the data for eliciting the prior, albeit possibly in an implicit and informal way. And this represents another way for saying that the prior can only be understood in the context of the likelihood (Gelman et al., 2017).

Why proposing a new data-dependent prior formulation? We acknowledge at least two reasons. From a Bayesian point of view, we want to investigate the information's extent of a prior distribution, and our proposal follows the words of Gelman (2016*b*), when he says that we need a compromise between the information carried by a "*wildly unrealistic in most settings prior informative distribution and a noninformative prior, feasible only in settings where data happen to be strongly informative about all parameters*". And from a broader statistical point of view, we are interested in the global quality of the model and on the assumptions we propose, and we believe our prior might be a good solution in case of model/prior misspecification.

According to the first argument, we are aware that the use of informative priors — or, at least, weakly informative priors (Gelman et al., 2008)— is strongly encouraged by subjectivist Bayesians, especially when a prior information for a specific application is actually available. However, even if the model is simple, when the sample size is 'small' it is not trivial to elicit an informative prior that does not dominate the inference. Using an informative prior distribution elicited from historical data —as it is

usual in medical studies, for instance— could result in a mismatch between the prior and the observed data, the so called *prior-data conflict* (Evans et al., 2006; Mutsvari et al., 2016). Thus, it emerges clearly that measuring the information contained in a prior distribution is not referred only as a mathematical exercise, but turns out to be helpful in terms of inference and prediction purposes. For instance, Morita et al. (2008) developed the so called prior effective sample size (ESS), an index which measures the amount of information contained in a proposed prior distribution π for the parameter θ, computed with respect to a posterior $q_m(\theta|y)$ resulting from a baseline prior π_b, with π_b less informative than π. When fitting a Bayesian model to a dataset consisting of 10 observations, an effective sample size of 1 is reasonable, whereas a value of 20 implies that the prior, rather than the data, dominates the inference: with a few data, there is the risk of being 'too much informative'.

Motivated by these considerations, our method uses data for dealing directly with the priors' construction. In what follows, we assume to be able to elicit a noninformative and an informative prior. Our procedure measures the discrepancy between the data and the informative prior. Depending on the sample size of the data at hand, we may need a resampling from the supposed true model, in order to neutralize the impact of the informative prior. The corresponding value of such a distance —bounded in the interval $[0, 1]$— is plugged into a two-components mixture of the two priors mentioned above. The greater is this value, the farther are the data (simulated and real) from the informative prior, and consequently the stronger is the influence of the diffuse prior in our specification. We prove that the so obtained class of mixture priors —hereafter MDD priors— satisfies some nice properties. Among these, the distributions of this class always have a closed form in conjugate models and preserve the conjugacy. Under mild conditions, they yield a lower effective sample size than that provided by the informative prior —substantially, they provide less information. Moreover, evidences from simulation studies show that they may also yield lower mean squared errors and improve the frequentist coverage.

It is worth noting that the use of mixture priors —possibly with one relative precise component and the other more vague— is not a novelty in Bayesian statistic. They have been introduced for making the inference robust in terms of a Bayesian perspective (Berger and Berliner, 1986), and developed for assessing any prior-data conflict (Schmidli et al., 2014; Mutsvari et al., 2016). A mixture specification turns out to be useful also in Bayesian variable selection: a 'spike and slab' prior (Miller, 2002)

with fixed hyperparameters is assigned to the regression coefficients in the stochastic search variable selection approach —see O'Hara et al. (2009) for an overview on variable selection methods.

The chapter is organized as follows. Section 2.2 reviews the existing data-dependent approaches and presents in a few details the frameworks proposed by Darnieder (2011) and Gelman (2016*a*); moreover, this section puts also in evidence the connection between the double use of the data and the penalized likelihood methods under a Bayesian perspective. In Section 2.3 we introduce the MDD density class and describe the resampling algorithm and the natural procedure required for building these priors. After introducing the notion of effective sample size, in Section 2.4 we focus on some theoretical results for the MDD priors; still, in this section we put in evidence the distribution-constant behavior of the Hellinger distance in some special cases, if used as discrepancy measure. The information for the proposed class of priors is discussed in two examples for non standard models in Section 2.6: an exponential model with a Jeffreys prior and a logistic regression for determining the greatest amount of tolerable dose in phase I trial. Section 2.7 concludes.

2.2 Using data twice in Bayesian inference

The commonly used expression 'using data twice' in some Bayesian procedures does not mean nothing really precise, actually. However, it is not of interest for us taking an overview on all those tools which make use of the data twice for checking the fit of the model —posterior predictive checkings, posterior Bayes fators, etc.— or reviewing the empirical Bayes methods (Carlin and Louis, 2000). In this section we focus on those priors' procedures which explicitly consider data in the elicitation process.

As widely known, using data or the data mechanism process in the priors' elicitation is not properly Bayesian and suffers from two main criticisms: using data twice and not allowing for the direct computation of the Bayes' Theorem. However, some authors have attempted to circumvent these criticisms. In what follows, we take a brief overview on some existing data-dependent approaches. Firstly, we present the theoretical framework proposed by Darnieder (2011), who formalized the so called Adjusted Data-dependent Bayesian paradigm, a new approach which introduces an adjustment in order to obtain a proper Bayesian inference starting from a data-dependent prior. Then, we present and formalize the considerations presented

by Gelman (2016*a*), who proposed to approximate a hierarchical model by using a data-dependent prior. We refer at Wasserman (2000) and Richardson and Green (1997) for the formulation of data-dependent priors that yield proper posteriors for finite mixture-models.

Finally, we draw a parallel between data-dependent priors and the penalized likelihood methods commonly used in Bayesian variable selection. Although this chapter does not explicitly take in consideration regression models, it is of future interest for us to implement our procedure also for regression purposes, and we consider this subsection as a grounding motivation for future work.

2.2.1 Darnieder's approach

Let $\boldsymbol{y}$ denote the sample, $\boldsymbol{\theta}$ the vector of parameters and $T(\boldsymbol{y})$ a statistic. Let $\pi(\boldsymbol{\theta}|T(\boldsymbol{y}))$ denote a data-dependent prior which depends on the data through the statistic $T(\boldsymbol{y})$. Darnieder (2011) espresses the joint probability density of $(\boldsymbol{\theta}, \boldsymbol{y}, T(\boldsymbol{y}))$ as:

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{y}, T(\boldsymbol{y})) =& p(T(\boldsymbol{y})|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})m(\boldsymbol{y}) \\ =& f(\boldsymbol{y}|\boldsymbol{\theta}, T(\boldsymbol{y}))\pi(\boldsymbol{\theta}|T(\boldsymbol{y}))m(T(\boldsymbol{y})), \end{aligned}$$

where $m(\boldsymbol{y})$ is the marginal (or integrated) likelihood. By isolating the posterior distribution on the left side, we obtain

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta}, T(\boldsymbol{y}))\pi(\boldsymbol{\theta}|T(\boldsymbol{y}))m(T(\boldsymbol{y}))}{p(T(\boldsymbol{y})|\boldsymbol{\theta}, \boldsymbol{y})m(\boldsymbol{y})}. \tag{2.1}$$

Now, we observe that given $\boldsymbol{y}$, $T(\boldsymbol{y})|\boldsymbol{\theta}, \boldsymbol{y}$ is not random, and that the ratio $m(T(\boldsymbol{y}))/m(\boldsymbol{y})$ depends only on the observed data. Hence, we may write the above expression as

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta}, T(\boldsymbol{y}))\pi(\boldsymbol{\theta}|T(\boldsymbol{y})). \tag{2.2}$$

As stated by Darnieder (2011), the posterior in (2.2) is obtained through a *naive* approach. The equation is suggesting that, if a data-dependent prior is used, then, in order to derive a proper posterior, also the likelihood of the model should be conditioned on the statistic $T(\boldsymbol{y})$. This formula is mathematically appealing, but the

update of $\pi(\boldsymbol{\theta}|T(\boldsymbol{y}))$ is often not straightforward. Hence, after some simple algebra, the posterior may be expressed as

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|T(\boldsymbol{y}))}{g(T(\boldsymbol{y})|\boldsymbol{\theta})} = f(\boldsymbol{y}|\boldsymbol{\theta})\frac{\pi(\boldsymbol{\theta}|T(\boldsymbol{y}))}{g(T(\boldsymbol{y})|\boldsymbol{\theta})}, \tag{2.3}$$

where the ratio $\pi(\boldsymbol{\theta}|T(\boldsymbol{y})/g(T(\boldsymbol{y})|\boldsymbol{\theta})$ is the actual data-dependent prior, updated with the usual unconditioned likelihood $f(\boldsymbol{y}|\boldsymbol{\theta})$. Darnieder (2011) defines the posterior in (2.3) as an *adjusted* posterior, obtained through an adjusted procedure. He also shows a relationship between a genuine Bayesian approach and the data-dependent Bayesian approach, putting in evidence the following identity:

$$1 = \frac{\pi(\boldsymbol{\theta}|\boldsymbol{y})m(\boldsymbol{y})}{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} = \frac{\pi(\boldsymbol{\theta}|T(\boldsymbol{y}))m(T(\boldsymbol{y}))}{g(T(\boldsymbol{y})|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}. \tag{2.4}$$

By multiplying this expression by the genuine prior $\pi(\boldsymbol{\theta})$, we can state the following proportionality, the so called data-dependent Bayesian Principle:

$$\frac{\pi(\boldsymbol{\theta}|\boldsymbol{y})}{f(\boldsymbol{y}|\boldsymbol{\theta})} \propto \frac{\pi(\boldsymbol{\theta}|T(\boldsymbol{y}))}{g(T(\boldsymbol{y})|\boldsymbol{\theta})}, \tag{2.5}$$

which formally coincides with (2.3), but suggests something even stronger. In fact, this expression highlights that the principle is satisfied whether a genuine prior $\pi(\boldsymbol{\theta})$ exists or not. With the adjusted procedure we provide a posterior distribution which is directly implied by Bayes' Theorem, whatever is the choice for $\pi(\boldsymbol{\theta})$.

A natural question concerns the choice of the statistic $T(\boldsymbol{y})$. There are no particular guidelines for choosing $T(\boldsymbol{y})$, but Darnieder (2011) lists some theorems that are useful for this aim. For example, it is trivial to show that if $T(\boldsymbol{y})$ is sufficient for $\boldsymbol{y}$, then the data-dependent prior $\pi(\boldsymbol{\theta}|T(\boldsymbol{y}))$ coincides with the genuine posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. And the following theorem in case of a distribution-constant statistic $T(\boldsymbol{y})$ will be useful later.

Theorem 1. *Suppose $T(\boldsymbol{y})$ is distribution-constant for $\boldsymbol{\theta}$, then the naive expression* (2.2) *and the adjusted expression* (2.3) *coincide. Furthermore, the data-dependent prior $\pi(\boldsymbol{\theta}|T(\boldsymbol{y}))$ coincides with the genuine prior $\pi(\boldsymbol{\theta})$.*

For a quick proof see the Appendix A. As suggested by Darnieder (2011), it is hard to imagine a beneficial conditioning on a distribution-constant statistic, unless for those priors which depend only on the data sample size. However, in Section 2.4 we

will use this result for showing that, within some particular cases, our data-dependent prior procedure only depends on the sample size of our dataset.

2.2.2 Gelman's approach

Gelman (2016*a*) draws an appealing framework considering the data-dependent priors as an approximation of a hierarchical model. He moves from a concrete example of regression models with standardized predictors: rescaling a bunch of predictors based on the data and then putting informative priors on their coefficients means eliciting a prior that depends on the data. He does not go in depth with mathematical notation, but we consider relevant to formalize this setup.

As usual in hierarchical models (Gelman, Carlin, Stern and Rubin, 2014), let $\boldsymbol{y}$ represent the data, with y_{ij} the observed value for the units i in group j, $i = 1, \ldots, m$, $j = 1, \ldots, J$; let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$ denote the generic vector of parameters and $\boldsymbol{\phi}$ the vector of hyperparameters. The likelihood of the model is $p(\boldsymbol{y}|\boldsymbol{\theta})$. The joint prior distribution for $(\boldsymbol{\theta}, \boldsymbol{\phi})$ is

$$\pi(\boldsymbol{\theta}, \boldsymbol{\phi}) = \pi(\boldsymbol{\phi})\pi(\boldsymbol{\theta}|\boldsymbol{\phi}),$$

and the joint posterior distribution is

$$\pi(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta}, \boldsymbol{\phi})p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}) = p(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\phi})\pi(\boldsymbol{\phi}), \tag{2.6}$$

with the further assumption that the hyperparameter $\boldsymbol{\phi}$ affects $\boldsymbol{y}$ only through $\boldsymbol{\theta}$. In a full Bayesian model, $\boldsymbol{\phi}$ is not known and is assigned a prior distribution $p(\boldsymbol{\phi})$; however, in some circumstances it may be possible to consider $\boldsymbol{\phi}$ as known, or estimate it. As in the Gelman's example, if this hyperparameter, say a *population* parameter, is estimated from the data, then we denote this estimate with $\boldsymbol{\phi}(\boldsymbol{y})$ and the population distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\phi})$ reduces to $\pi(\boldsymbol{\theta}|\boldsymbol{\phi}(\boldsymbol{y}))$, which actually is a data-dependent prior. If we replace $\boldsymbol{\phi}$ with an estimate, $\boldsymbol{\theta}$ still preserves the dependence from $\boldsymbol{\phi}(\boldsymbol{y})$, but the joint posterior distribution in (2.6) reduces to the following approximate posterior,

$$\pi(\boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{y})|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta}|\boldsymbol{\phi}(\boldsymbol{y}), \boldsymbol{y})\pi(\boldsymbol{\phi}(\boldsymbol{y})|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta}|\boldsymbol{\phi}(\boldsymbol{y}), \boldsymbol{y}), \tag{2.7}$$

where $\pi(\boldsymbol{\theta}|\boldsymbol{\phi}(\boldsymbol{y}), \boldsymbol{y})$ may be interpreted as the marginal approximate posterior for $\boldsymbol{\theta}$ —analogous to the pseudo-posterior distribution in empirical Bayes methods

(Petrone et al., 2014), where $\boldsymbol{\phi}(\boldsymbol{y})$ is usually obtained through marginal maximum likelihood estimation. We may derive an explicit form for this quantity by applying the Bayes' Theorem and the assumption $p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{y})) = p(\boldsymbol{y}|\boldsymbol{\theta})$:

$$\pi(\boldsymbol{\theta}|\boldsymbol{\phi}(\boldsymbol{y}), \boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{y}))\pi(\boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{y})) \propto p(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\phi}(\boldsymbol{y})). \tag{2.8}$$

The comparison between this latter expression and (2.6), (2.7) highlights the relationship existing between a full Bayesian hierarchical model and an approximate hierarchical model, where $\boldsymbol{\phi}(\boldsymbol{y})$ naturally acts in place of $\boldsymbol{\phi}$ and Bayes' Theorem is guaranteed by the product between the usual likelihood and the data-dependent prior $\pi(\boldsymbol{\theta}|\boldsymbol{\phi}(\boldsymbol{y}))$. The framework above has the merit of interpreting a data-dependent prior as an approximation of a further level of hierarchy within hierarchical models, through the use of a data-statistic $\boldsymbol{\phi}(\boldsymbol{y})$ as a plug-in estimate for the hyperparameter $\boldsymbol{\phi}$; moreover, it introduces the definition of a pseudo-posterior $\pi(\boldsymbol{\theta}|\boldsymbol{\phi}(\boldsymbol{y}), \boldsymbol{y})$.

2.2.3 Wassermann's approach

The general theoretical framework of interest for Wasserman (2000) refers to finite mixture models. In this situation he justified a data-dependent approach from a practical point of view, proposing a prior distribution which yields a proper posterior with good frequentist properties. He shows in fact that the only priors that produce intervals with second-order correct coverage are data-dependent. However, his practical interest serves as a first step for deriving a general theory of data-dependent priors: despite their use in many contexts, he underlined that they miss a theoretical formalization.

For simplicity, here we take a simple mixture model with standard gaussian distributions and with two possible groups. Let $\boldsymbol{y}_m = (y_1, ..., y_m)$ be iid with density for the single value y_i

$$f_\mu(y_i) = \frac{1}{2}\phi(y_i) + \frac{1}{2}\phi(y_i - \mu), \tag{2.9}$$

where ϕ denotes here the usual standard normal density. The likelihood of the model is then

$$L_m(\mu; \boldsymbol{y}_m) = \prod_{i=1}^{m} f_\mu(y_i) = \prod_{i=1}^{m} \left[\frac{1}{2}\phi(y_i) + \frac{1}{2}\phi(y_i - \mu)\right] \tag{2.10}$$

Choosing an improper prior $\pi(\mu) \propto 1$ in (2.9) yields an improper posterior, i.e.

$$\int L_m(\mu; \boldsymbol{y}_m)\pi(\mu)d\mu = \infty. \tag{2.11}$$

Nevertheless, choosing a proper prior could dominate the inference creating an unacceptable bias. Wasserman (2000) introduced a data-dependent prior $\pi_m(\mu)$ as a measurable function from the sample space $\mathcal{Y}^m$ to the set of the priors $\mathcal{P}$, defined as the set of all the non-negative, measurable, bounded, twice differentiable functions on the real line. The idea of Wasserman is that of multiplying the usual Jeffreys (1998) prior by a factor depending on the data at hand. Before proceeding, let us review some basic theory about the Jeffreys prior formulation. Let $S(\mu, Y) = \partial[\log f_\mu(Y)]/\partial\mu$ be the score function for a random variable Y and let $I_\mu = E[S^2(\mu, Y)]$ be the Fisher information. The Jeffreys prior is defined as

$$j(\mu) = I_\mu^{1/2}. \tag{2.12}$$

Now we may introduce the Wasserman's prior. Let μ_0 denote the true value of μ, let

$$D(\mu_0, \mu) = \int f_{\mu_0} \log(f_{\mu_0}/f_\mu)d\mu \tag{2.13}$$

be the Kullback-Leibler distance between f_{μ_0} and f_μ, and let $a(\mu_0) = sup_{\mu_0}\{D(\mu_0, \mu)\}$. Then, define

$$D_m(\mu_0, \mu) = \frac{1}{m}\sum_{i=1}^{m} \log\{f_{\mu_0}(y_i)/f_\mu(y_i)\}. \tag{2.14}$$

Finally, we may introduce the Wasserman's prior as

$$\pi_m(\mu) = j(\mu)c_m(\mu), \tag{2.15}$$

with $c_m(\mu) = 1 - \exp[-m\{a(\mu_0) - D_m(\mu_0, \mu)\}]$. Since the quantity $a(\mu_0) - D_m(\mu_0, \mu)$ does not depend on μ_0, the prior (2.15) depends on the data but not on the true value of the parameter. Furthermore, Wasserman (2000) proved that the proposed prior generates a proper posterior, is second order correct and that the data dependence of $\pi_m(\mu)$ vanishes asymptotically, as $m \to \infty$.

2.2.4 Penalized likelihood

In the penalized likelihood approaches for regression models —Lasso (Tibshirani, 1996), Ridge regression, Bridge regression— it is usual to penalize some coefficients by inducing a certain amount of shrinkage in order to (i) overcome problems in the stability of parameter estimates due to a relatively flat likelihood and (ii) reduce the global mean squared error. A penalized log-likelihood with quadratic penalization is

$$\log L(\boldsymbol{\beta}, \boldsymbol{y}) - \frac{r}{2}(\boldsymbol{\beta} - \boldsymbol{g})^2, \tag{2.16}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_J)$ is the vector of regression parameters, $\boldsymbol{g} = (g_1, ..., g_J)$ is a vector of values which should be good guesses for the vector parameter $\boldsymbol{\beta}$ or correspond to a reference model (for inference in spline estimation they may correspond to a constant or a straight line), and r is usually called the *tuning* parameter. Perhaps, $(\boldsymbol{\beta} - \boldsymbol{g})^2 = \sum_{j=1}^{J}(\beta_j - g_j)^2$ is the quadratic penalty in the Ridge regression. The formula above may be easily interpreted in a Bayesian perspective. In fact, if we adopt the prior $\beta_j \sim \mathcal{N}(g_j, 1/r)$, then (2.16) represents a log-likelihood penalized by the log-density of the prior distribution for β_j, where r is the precision (the inverse of the prior variance). Thus, the quadratic log-likelihood penalization reduces to eliciting independent normal priors for the regression parameters with prior mean g_j and prior variance $1/r$. The ordinary Lasso of Tibshirani can be interpreted as a Bayesian Lasso (Park and Casella, 2008), i.e. as a Bayesian posterior mode estimate when regression parameters have Laplace independent priors. And more generally Bridge regression is a direct generalization for Lasso and Ridge regression, where the penalty is $(\boldsymbol{\beta} - \boldsymbol{g})^q$ for some $q \geq 0$ ($q = 1$ corresponds to the ordinary Lasso, $q = 2$ to the Ridge regression). Many approaches for estimating the tuning parameter r have been proposed: cross-validation, general cross-validation, empirical Bayes methods through marginal maximum likelihood estimation. But only assigning a diffuse hyperprior is purely Bayesian. Using data for estimating the tuning parameter makes in fact the Bayesian penalized log-likelihood approach affected by the data process and, more precisely, the prior on β affected by the data. In Section 2.4 we put in evidence that our methodology allows for a hierarchical approximation and may be also justified in terms of log-likelihood penalization. We will still interpret the penalized likelihood under a Bayesian point of view, but allowing the tuning parameter to depend on the data.

2.3 Mixture Data-dependent priors

Let $\boldsymbol{y}_m = (y_1, ..., y_m)$ be a data vector from a given sampling distribution $f(\boldsymbol{y}_m|\theta)$, with $\theta \in \mathbb{R}$. Let $\pi_b(\theta)$ denote a diffuse prior distribution for θ —hereafter called *baseline* prior— and suppose that, from a preliminary knowledge about the problem (for instance historical information), we are somehow able to assign a more informative prior distribution $\pi(\theta)$. When data consist of a relatively small number of observations, the choice between these two priors' options is not trivial, since the support and the shape of the posterior are sensitive to the choice of the prior distribution. Thus, the information contained in the prior could be dominant when the dataset is small. This is one of the reasons for combining our previous information about the problem with our data at hand —-or, with an augmented version of it, as will be clarified later— and proposing a data-dependent approach for eliciting a particular class of mixture prior distributions. We may then introduce the mixture data-dependent (MDD) prior $\varphi(\theta)$ with mixture weight ψ_{m^*} as

$$\varphi(\theta) = \psi_{m^*}\pi_b(\theta) + (1 - \psi_{m^*})\pi(\theta), \tag{2.17}$$

belonging to the corresponding MDD class

$$\Phi = \{\varphi : \varphi(\theta) = \psi_{m^*}\pi_b(\theta) + (1 - \psi_{m^*})\pi(\theta),\ \theta \in \Theta,\ 1 \geq \psi_{m^*} \geq 0,\ m^* \in \mathbb{N}\}.$$

The MDD prior (2.17) may then be viewed as a compromise between an informative prior and a noninformative one, with weight ψ_{m^*} which represents a discrepancy measure between the informative prior and the data with global length m^* —we refer at the next subsection for the meaning of the symbol m^*. As will be more clear in what follows, the data dependence of this class is represented by the mixture weight ψ_{m^*}. Note that mixture priors designed for overcoming the prior-data conflict and for robustness purposes have been already proposed by Mutsvari et al. (2016) and Schmidli et al. (2014): however, the authors do not propose any procedure for computing/assigning the mixture weights, and this is a crucial point for us, as explained in the next section.

2.3.1 The resampling algorithm for the mixture weigths

Assume to have observed the data vector $\boldsymbol{y}_m$ and let introduce here the symbol Ω_m for the Hellinger distance $\mathcal{H}$ —closely related to the Bhattacharyya distance

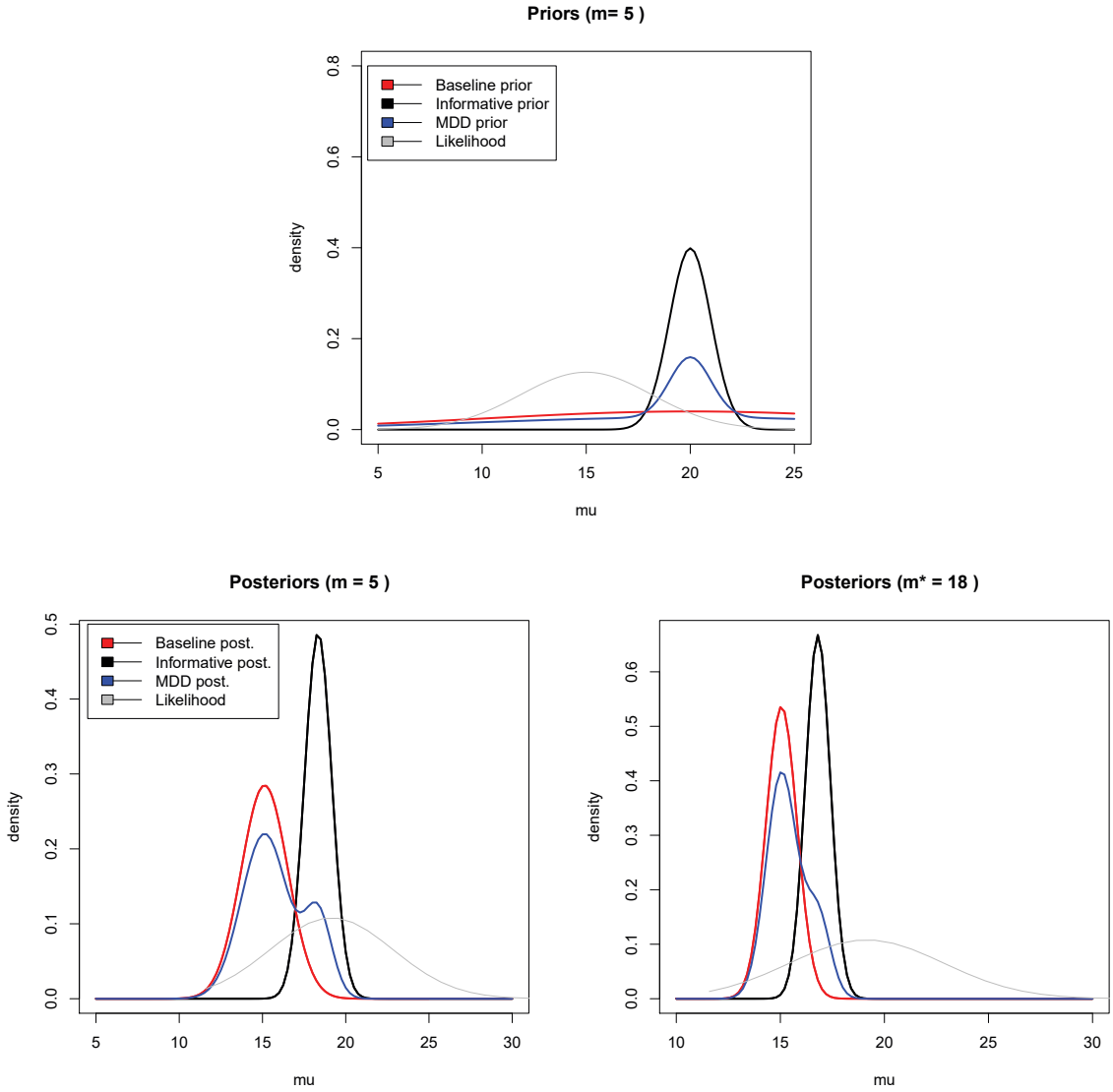


Figure 2.1: Normal-Normal model, resampling-algorithm, $\epsilon = 0.2$. (*Top*) $f(\boldsymbol{y}_m|\theta) = \mathcal{N}(15, 10)$ (grey line), $\pi_b(\theta) = \mathcal{N}(20, 100)$, $\pi(\theta) = \mathcal{N}(20, 1)$ and $\varphi(\theta) = \psi_{m^*}\mathcal{N}(20, 100) + (1 - \psi_{m^*})\mathcal{N}(20, 1)$. The initial sample is set to $m = 5$. (*Bottom row, left*) Baseline posterior $q_m(\theta|\boldsymbol{y}_m)$, informative posterior $\pi_m(\theta|\boldsymbol{y}_m)$, MDD posterior $\varphi_m(\theta|\boldsymbol{y}_m)$ for the initial sample size m. The gray line is the density for the new values $\boldsymbol{y}_\varkappa$ generated under $f(\boldsymbol{y}_m|\theta^*)$. (*Bottom row, right*) Baseline posterior $q_{m^*}(\theta|\boldsymbol{y}_{m^*})$, posterior $\pi_{m^*}(\theta|\boldsymbol{y}_{m^*})$, MDD posterior $\varphi_{m^*}(\theta|\boldsymbol{y}_{m^*})$, for the sample size $m^* = m + \varkappa$, here 18.

(Bhattacharyya, 1946)— between the baseline posterior $q_m(\theta|\boldsymbol{y}_m)$ and the informative posterior $\pi_m(\theta|\boldsymbol{y}_m)$:

$$\Omega_m \equiv \mathcal{H}(q_m(\theta|\boldsymbol{y}_m), \pi_m(\theta|\boldsymbol{y}_m)) = \frac{1}{\sqrt{2}} \left(\int (q_m^{1/2} - \pi_m^{1/2})^2 d\boldsymbol{y}_m \right)^{1/2}. \tag{2.18}$$

For any couple of density functions g, h, the Hellinger distance satisfies the property: $0 \leq \mathcal{H}(g,h) \leq 1$. The basic idea of our procedure consists of weighting a pair of priors $\pi(\theta), \pi_b(\theta)$ through a discrepancy measure between the proposed informative prior and the data at hand. This measure of data-prior compatibility may be formulated in several ways, as will be explained later. According to this task, we could rely on the available data, *conditioning* on them for checking a misfit between the prior and the likelihood. But for reasons that will be clarified later, we may need a sequential generation of further $\varkappa$ draws from the *true* model $f(\boldsymbol{y}_m|\theta_0)$: the so obtained augmented sample size $m^* = m + \varkappa$ should be then large enough for neutralizing the impact of the informative prior π, as is clarified below. The generic $\varkappa$ may be seen as a *tuning* parameter which needs to be computed for the specification of the mixture weight. Since the true model is unknown, we decide to generate from an approximation of the true model. Then, we may have the following situations:

(1) $y_1, \ldots, y_m \sim f(\boldsymbol{y}_m|\theta_0)$ (no resampling),

(2) $y_1, \ldots, y_m \sim f(\boldsymbol{y}_m|\theta_0)$; generate $y_{m+1}, \ldots, y_{m+\varkappa} \sim f(\boldsymbol{y}_m|\hat{\theta}_0)$,

where $f(\boldsymbol{y}_m|\theta^*)$ is the likelihood evaluated for $\theta^* \sim \pi(\theta)$, $\hat{\theta}_0$ is an estimate for the true parameter θ_0, and $\varkappa$ is the dimension of the augmented dataset for **(2)**. In most of the statistical applications the true parameter value is unknown and needs to be estimated. Among the others, one possibility is that of using the maximum likelihood (ML) estimate $\hat{\theta}_0$, obtained equating at zero the log-derivative of the sampling distribution. Thus, for **(2)** we re-compute (2.18) for each new draw and we stop the procedure when a certain condition of similarity between the posterior distributions π_m and q_m is satisfied. Precisely, the stop condition is expressed by

$$\varkappa = \inf \{k \in \mathbb{N} \, | \Omega_{m+k} < \epsilon, \ \epsilon > 0\} \tag{2.19}$$

for a fixed tolerance ϵ, with $\epsilon > 0$. This posterior similarity may be seen as an approximate matching between the proposed posterior distributions [1]. The use of the Hellinger distance is appropriate for some nice theoretical properties, as will be clarified in Section 2.4, and for being defined in $[0, 1]$. However, other measures of discrepancy as the Kullback-Leibler divergence or the Bhattacharyya distance could be adopted. As suggested above, the factor ψ_{m^*} computed for the augmented sample with length m^* in (2.17) measures the observed *discrepancy* between the informative prior π and the data. According to situations **(1)** and **(2)**, we propose two possible discrepancy measures:

(1*) $\Psi_{m^*} \equiv \mathcal{H}(\pi(\theta), \pi_m(\theta|\boldsymbol{y}_m)),\ m^* = m,$

(2*) $\Psi_{m^*} \equiv \mathcal{H}(f(\boldsymbol{y}_m|\hat{\theta}_0^{(m^*)}), f(\boldsymbol{y}_m|\theta^*)),\ m^* = m + \varkappa.$

The first equation is the Hellinger distance between the informative prior and the informative posterior for the data at hand. The second equation measures the discrepancy between the sampling distribution for the augmented set of data, where $\hat{\theta}_0^{(m^*)}$ is the ML estimate for θ_0 based on $y_1, \ldots, y_m,\ y_{m+1}, \ldots, y_{m^*}$, and the same sampling distribution evaluated in terms of the informative prior, since $\theta^* \sim \pi(\theta)$. As mentioned above, resampling may result to be beneficial in some applications, as will be shown later. But it could be also demanding in terms of computational times whenever the required sample size for matching the posterior distribution —perhaps, for neutralizing the impact of a *misspecified* informative prior— results extremely large. The MDD priors built respectively under procedures **(1)**-**(1*)** and **(2)**-**(2*)** are called *natural* —hereafter MDD-natural— and with *resampling-algorithm* —hereafter MDD-res.

Resampling-algorithm

Given $y_1, ..., y_m \sim f(\boldsymbol{y}_m|\theta)$, generate $\theta^* \sim \pi(\theta)$.

Fix the tolerance ϵ, with $\epsilon > 0$.

Given $\Psi_m \equiv \mathcal{H}(f(\boldsymbol{y}_m|\theta_0), f(\boldsymbol{y}_m|\theta^*))$, $\Omega_m \equiv \mathcal{H}(q_m(\theta|\boldsymbol{y}_m), \pi_m(\theta|\boldsymbol{y}_m))$, compute the observed value ω_m. If the true value θ_0 is unknown, provide an estimate for it.

If $\omega_m > \epsilon$ set $k = 1$.

[1] Note that the idea of matching the posterior uncertainty carried by two different posteriors does not represent a novelty, and a procedure based on the average posterior uncertainty is proposed by Reimherr et al. (2014).

$\triangle$ generate y_{m+k} from $f(\boldsymbol{y}_m|\hat{\theta}_0^{(k)})$. Given

$$\Psi_{m+k} \equiv \mathcal{H}(f(\boldsymbol{y}_m|\hat{\theta}_0^{(k)}), f(\boldsymbol{y}_m|\theta^*))$$
$$\Omega_{m+k} \equiv \mathcal{H}(q_{m+k}(\theta|\boldsymbol{y}_{m+k}), \pi_{m+k}(\theta|\boldsymbol{y}_{m+k}))$$

with $\hat{\theta}_0^{(k)}$ the ML estimate for θ_0 at step k.

$\triangle\triangle$ Compute the observed values ψ_{m+k}, ω_{m+k}.

while $\{\omega_{m+k} > \epsilon\}$ set $k = k + 1$ and go back to $\triangle$.

Save $\psi_{m+\varkappa}$, $\omega_{m+\varkappa}$ and the new sample size $m^* = m + \varkappa$. Set the prior (2.17) with ψ_{m^*}.

For illustration purposes only, Figure 2.1 displays a graphical example for the MDD prior and posterior (blue lines) obtained through the resampling-algorithm for a simple Normal-Normal model, where the baseline variance is set to 100 and the informative variance is set to 1, $\epsilon = 0.2$. The computation of the Hellinger distance has been obtained through the `R` function `HellingerDist` of the `distrEx` package (Kohl et al., 2007).

2.3.2 Justification for resampling

Given any pair of density functions f, g, let consider the following equivalent formulation for the Hellinger distance:

$$\begin{aligned}\mathcal{H}(f,g) &= \frac{1}{\sqrt{2}}\left(\int (f^{1/2} - g^{1/2})^2 d\boldsymbol{y_m}\right)^{1/2} = \\ &= \frac{1}{\sqrt{2}}\left(\int f d\boldsymbol{y_m} + \int g d\boldsymbol{y_m} - 2\int [fg]^{1/2} d\boldsymbol{y_m}\right)^{1/2} \\ &= \left(1 - \int [fg]^{1/2} d\boldsymbol{y_m}\right)^{1/2},\end{aligned} \tag{2.20}$$

where the last integral in (2.20) is also called *affinity* (Van der Vaart, 1998). Now, suppose that the data vector $y_1, \ldots, y_m$ may derive from one among the above densities, where $f = \mathcal{N}(\theta_0, 1)$ and $g = \mathcal{N}(\theta^*, 1)$, with $\theta^* \sim \pi(\theta)$. For simplicity, let consider $m^* = m$. Then, the squared Hellinger distance is:

$$
\begin{aligned}
\mathcal{H}^2(f,g) \equiv \psi^2_{m^*} &= 1 - \int [fg]^{1/2} d\boldsymbol{y_m} = \\
&= 1 - \int \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{4}(\sum_i y_i^2 + m\theta_0^2 - 2m\bar{y}\theta_0 + \sum_i y_i^2 + m\theta^{*2} - 2\bar{y}m\theta^*\})d\boldsymbol{y_m} = \\
&= 1 - \int \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(\sum_i y_i^2 + \frac{1}{4}m(\theta^{*2} + \theta_0^2 + 2\theta^*\theta_0) - m\bar{y}(\theta^* + \theta_0))\}\cdot \\
&\qquad \cdot \exp\{-\frac{1}{2}(\frac{1}{4}m(\theta^{*2} + \theta_0^2 + 2\theta^*\theta_0)\} \exp\{\frac{1}{2}m\theta^*\theta_0\} d\boldsymbol{y_m} = \\
&= 1 - \exp\{-\frac{1}{2}(\frac{1}{4}m(\theta^{*2} + \theta_0^2 + 2\theta^*\theta_0)\} \exp\{\frac{1}{2}m\theta^*\theta_0\} = \\
&= 1 - \exp\{-\frac{1}{2}(\frac{1}{4}m\theta_0^2 + \frac{1}{4}m\theta^{*2} - \frac{1}{2}m\theta^*\theta_0)\} = \\
&= 1 - \exp\{-\frac{m}{8}(\theta^* - \theta_0)^2\}.
\end{aligned}
$$

As $m \to \infty$, the quantity above approximates one, and then plugging $\psi_{m^*} = 1$ in (2.17) means eliciting the noninformative prior. From this toy example, it emerges that the more data at hand one has, the more reliable is the noninformative choice, and this is intuitive. *But*, in this chapter we consider cases where we deal with small samples: if we do not have enough data, *we should generate them from the supposed true model*.

A rigorous argument for justifying our proposed algorithm moves from the theory of the rescaling rates developed in Van der Vaart (1998). Given a sequence of models $\{P_{\theta_m}, \theta_m \in \Theta,\ m \in \mathbb{N}\}$, for each m we may be interested in testing the null hypothesis $H_0 : \theta = \theta_0$ versus the alternatives $\theta = \theta_m$. The hypothesis testing approach is of course not relevant in our procedure, but it is useful as a theoretical tool. The L_1-distance between two distributions $P_{\theta_m}, P_{\theta_0}$ with densities $p_{\theta_m} = dP_{\theta_m}/d\mu$ and $p_{\theta_0} = dP_{\theta_0}/d\mu$, for a given measure μ, is defined as

$$
||P_{\theta_m} - P_{\theta_0}|| = \int |p_{\theta_m} - p_{\theta_0}| d\mu.
$$

It is worth noting that $||P_{\theta_m} - P_{\theta_0}|| \to 0$ if and only if $\mathcal{H}(p_{\theta_m}, p_{\theta_0}) \to 0$. The Lemma 14.31 in Van der Vaart (1998) states that if $\mathcal{H}^2(p_\theta, p_{\theta_0}) = O(|\theta - \theta_0|^\alpha)$ as $\theta \to \theta_0$, then, for any sequence of alternatives θ_m, $||P_{\theta_m} - P_{\theta_0}||$ is bounded away from 0 and ∞ when $m^{1/\alpha}|\theta_m - \theta_0|$ is bounded away from 0 and ∞. In the exponential family models we have $\mathcal{H}^2(p_\theta, p_{\theta_0}) = O(|\theta - \theta_0|^2)$ and hence the rate of convergence is $\sqrt{m}$. This last condition is another way for stating that $\mathcal{H}^2(p_{\theta_m}, p_{\theta_0}) = O(m^{-1})$,

or, equivalently, $\mathcal{H}(p_{\theta_m}, p_{\theta_0}) = O(m^{-1/2})$. *Within exponential family models, for any sequence of alternatives* θ_m *that does not converge at zero or diverge at* ∞ *as* m *grows, the Hellinger distance is bounded between 0 and 1 and has a rate of convergence equals* $\sqrt{m}$.

Our procedure sequentially generates new data $y_{m+1}, \ldots, y_{m+\varkappa}$, and at each k, $k = 1, \ldots, \varkappa$ we measure a distance between $f(\boldsymbol{y}|\theta^*)$ and $f(\boldsymbol{y}|\hat{\theta}^{(k)})$, where $\theta^* \sim \pi(\theta)$. Translated in the Van der Vaart hypothesis testing framework, at each k our method may be seen as a rude test of the null hypothesis $H_0 : \theta = \theta_0$ versus an alternative $H_1 : \theta = \theta_k$, where $\theta_k = \theta_0 + h_k$ depends on the k-th draw through a certain constant h_k. In the resampling-algorithm data are generated from $f(\boldsymbol{y}_m|\hat{\theta}_0^{(k)})$, where $\hat{\theta}_0^{(k)}$ is an estimate for the true-value parameter θ_0, usually unknown. The discrepancy measure is defined as $\Psi_k \equiv \mathcal{H}(f(\boldsymbol{y}_m|\hat{\theta}_0^{(k)}), f(\boldsymbol{y}_m|\theta^*)),\ k = 1, \ldots, \varkappa$, and this is simply the Hellinger distance between two absolute continuous distributions, where the true parameter is estimated at each k. Then, this algorithm consists in testing for each k the null hypothesis $H_0 : \theta = \hat{\theta}_0^{(k)}$ versus the alternative $H_1 : \theta = \theta^*$, where $\theta^* \sim \pi(\theta)$.

Unlike for the natural procedure, the resampling method proposes a discrepancy measure based on a data augmentation. Hence, $\Psi_{m^*} = O((m + \varkappa)^{-1/2})$, where m is fixed and $\varkappa$ may vary, and the rate of convergence is $\sqrt{m + \varkappa}$. Thus, sampling further data, rather than using only the original sample with length m, guarantees an asymptotic rate of convergence for our discrepancy measure. Intuitively, the so obtained observed discrepancy should really assess the reliability of the prior π, checking its similarity with the data through resampled data generated from the supposed true model.

2.4 Theoretical results for the MDD class: conjugate models

In this section we present some theoretical results for the MDD class presented in Section 2.3 within the univariate conjugate models. Precisely, we introduce here the notion of effective sample size proposed by Morita et al. (2008), showing that the information of the MDD prior is always lower than the information of any informative prior. Moreover, we frame the MDD prior class in the theoretical approaches of Darnieder (2011) and Gelman (2016*a*), summarized in Section 2.2. According to the

first reference, we review the notion of distribution-constant statistics and we put in evidence that in some special cases —e.g. the Normal-Normal model, but generally all the statistical models for which the Fisher information does not depend on the parameter— the Hellinger distance is a distribution-constant statistic. This property implies that in these special models our proposed methodology substantially reduces to choosing a genuine prior.

Before proceeding, we introduce here a general vector notation that turns out to be helpful in the following sections. Without any loss of generality, let $\boldsymbol{\theta} \in \mathbb{R}^d$, denote the parameters' vector, with $d \geq 1$. Let the symbols $\pi_b(\boldsymbol{\theta})$, $\pi(\boldsymbol{\theta})$ denote as before respectively a baseline prior and an informative prior for $\boldsymbol{\theta}$. Let m denote the generic sample size and $f(\boldsymbol{y}_m|\boldsymbol{\theta})$ the likelihood for our sample $\boldsymbol{y}_m = (y_1, ..., y_m)$. Finally, let $q_m(\boldsymbol{\theta}|\boldsymbol{y}_m)$ denote the baseline posterior for our parameter $\boldsymbol{\theta}$. In Section 2.3 we used the symbols m for the initial sample size, $\varkappa$ for the sample size of the generated sample of data and, consequently, $m^* = m + \varkappa$ for the global dimension of the data vector, comprising both the data at hand and those generated via resampling-algorithm. The further technical assumptions are

$$\begin{aligned} &E_{\pi_b}(\boldsymbol{\theta}) = E_{\pi}(\boldsymbol{\theta}) \\ &\text{Corr}_{\pi}(\theta_i, \theta_j) = \text{Corr}_{\pi_b}(\theta_i, \theta_j),\ i \neq j \\ &\text{Var}_{\pi_b}(\theta_j) >> \text{Var}_{\pi}(\theta_j),\ j = 1, ..., d. \end{aligned} \tag{2.21}$$

2.4.1 Effective sample size (ESS)

The idea of measuring and quantifying the amount of information contained in a prior distribution is of a great theoretical appeal. Nevertheless, it has not yet been studied by many authors and many technical difficulties arise, including the impossibility of encompassing in a unique philosophical and mathematical framework the task of assessing the impact of a prior distribution: several distance measures and many definitions of prior sample size may be in fact adopted. In what follows we will refer to the work of Morita et al. (2008), who defined the prior effective sample size (ESS) of $\pi(\boldsymbol{\theta})$, with respect to the likelihood $f(\boldsymbol{y}_m|\boldsymbol{\theta})$ as that integer m which minimizes the distance between $\pi(\boldsymbol{\theta})$ and the baseline posterior $q_m(\boldsymbol{\theta}|\boldsymbol{y}_m)$. To define this distance, they used the negative second partial derivatives of the log-densities (the observed

Table 2.1: $\theta \in \mathbb{R}$, $c \geq 1$. Suppose $\boldsymbol{y}_m = (y_1, ..., y_m) \sim f(\boldsymbol{y}_m|\theta)$. Prior $\pi(\theta)$, baseline prior $\pi_b(\theta)$, MDD prior $\varphi(\theta)$, likelihood $f(\boldsymbol{y}_m|\theta)$, baseline posterior $q_m(\theta|\boldsymbol{y}_m)$ and MDD posterior $\varphi_m(\theta|\boldsymbol{y}_m)$ for the univariate conjugate models: Normal-Normal (NN), Gamma-Poisson (GP), Gamma-Exponential (GExp) and Beta-Binomial (BB). Following Gelman, Carlin, Stern and Rubin (2014), we denote $\mathcal{N}(\mu, \sigma^2)$, $\mathcal{G}\text{a}(\alpha, \beta)$, $\mathcal{B}\text{e}(\alpha, \beta)$, $\mathcal{B}\text{in}(n, \theta)$, $\mathcal{P}\text{ois}(\theta)$ and $\mathcal{E}\text{xp}(\theta)$ for the normal, gamma, beta, binomial, Poisson and exponential distributions. For the Normal-Normal model let $\bar{\mu}(\tau^2) = (\frac{\mu}{\tau^2} + \frac{m}{\sigma^2}\bar{y})/(\frac{1}{\tau^2} + \frac{m}{\sigma^2})$ denote the posterior mean in function of the prior variance τ^2, and $\bar{\tau}^2(\tau^2) = (\frac{1}{\tau^2} + \frac{m}{\sigma^2})^{-1}$ the posterior variance in function of the prior variance τ^2.

	NN	*GP*
$\pi_b(\theta)$	$\mathcal{N}(\mu, c\tau^2)$	$\mathcal{G}\text{a}(\frac{\alpha}{c}, \frac{\beta}{c})$
$\pi(\theta)$	$\mathcal{N}(\mu, \tau^2)$	$\mathcal{G}\text{a}(\alpha, \beta)$
$\varphi(\theta)$	$\psi_{m^*}\mathcal{N}(\mu, c\tau^2)+$	$\psi_{m^*}\mathcal{G}\text{a}(\frac{\alpha}{c}, \frac{\beta}{c})+$
	$(1-\psi_{m^*})\mathcal{N}(\mu, \tau^2)$	$(1-\psi_{m^*})\mathcal{G}\text{a}(\alpha, \beta)$
$f(\boldsymbol{y}_m\|\theta)$	$\mathcal{N}(\theta, \sigma^2)$	$\mathcal{P}\text{ois}(\theta)$
$q_m(\theta\|\boldsymbol{y}_m)$	$\mathcal{N}(\bar{\mu}(c\tau^2), \bar{\tau}^2(c\tau^2))$	$\mathcal{G}\text{a}(\frac{\alpha}{c} + \sum y_i, \frac{\beta}{c} + m)$
$\varphi_m(\theta\|\boldsymbol{y}_m)$	$\psi_{m^*}\mathcal{N}(\bar{\mu}(c\tau^2), \bar{\tau}^2(c\tau^2))+$	$\psi_{m^*}\mathcal{G}\text{a}(\frac{\alpha}{c} + \sum y_i, \frac{\beta}{c} + m)+$
	$(1-\psi_{m^*})\mathcal{N}(\bar{\mu}(\tau^2), \bar{\tau}^2(\tau^2))$	$(1-\psi_{m^*})\mathcal{G}\text{a}(\alpha + \sum y_i, \beta + m)$

	GExp	*BB*
$\pi_b(\theta)$	$\mathcal{G}\text{a}(\frac{\alpha}{c}, \frac{\beta}{c})$	$\mathcal{B}\text{e}(\frac{\alpha}{c}, \frac{\beta}{c})$
$\pi(\theta)$	$\mathcal{G}\text{a}(\alpha, \beta)$	$\mathcal{B}\text{e}(\alpha, \beta)$
$\varphi(\theta)$	$\psi_{m^*}\mathcal{G}\text{a}(\frac{\alpha}{c}, \frac{\beta}{c})+$	$\psi_{m^*}\mathcal{B}\text{e}(\frac{\alpha}{c}, \frac{\beta}{c})+$
	$(1-\psi_{m^*})\mathcal{G}\text{a}(\alpha, \beta)$	$(1-\psi_{m^*})\mathcal{B}\text{e}(\alpha, \beta)$
$f(\boldsymbol{y}_m\|\theta)$	$\mathcal{E}\text{xp}(\theta)$	$\mathcal{B}\text{in}(m, \theta)$
$q_m(\theta\|\boldsymbol{y}_m)$	$\mathcal{G}\text{a}(\frac{\alpha}{c} + m, \frac{\beta}{c} + m\bar{y})$	$\mathcal{B}\text{e}(\frac{\alpha}{c} + m\bar{y}, \frac{\beta}{c} + m - m\bar{y})$
$\varphi_m(\theta\|\boldsymbol{y}_m)$	$\psi_{m^*}\mathcal{G}\text{a}(\frac{\alpha}{c} + m, \frac{\beta}{c} + m\bar{y})+$	$\psi_{m^*}\mathcal{B}\text{e}(\frac{\alpha}{c} + m\bar{y}, \frac{\beta}{c} + (m - m\bar{y}))+$
	$(1-\psi_{m^*})\mathcal{G}\text{a}(\alpha + m, \beta + m\bar{y})$	$(1-\psi_{m^*})\mathcal{B}\text{e}(\alpha + m\bar{y}, \beta + m - m\bar{y})$

Table 2.2: $\theta \in \mathbb{R},\ c \geq 1,\ m$ is the generic sample size. Negative second derivatives of the log-densities and effective sample sizes for the baseline prior $\pi_b(\theta)$, the informative prior $\pi(\theta)$ and the MDD prior $\varphi(\theta)$, for the four univariate conjugate models. Let $\bar{\theta} = E_\pi(\theta)$ denote the plug-in estimate. See Table 2.1 for the priors' specification.

	NN	*GP*	*GExp*	*BB*
$D_{\pi_b}(\theta)$	$1/c\tau^2$	$\frac{(\alpha/c-1)}{\bar{\theta}^2}$	$\frac{(\alpha/c-1)}{\bar{\theta}^2}$	$(\frac{\alpha}{c}-1)\frac{1}{\bar{\theta}^2} + (\frac{\beta}{c}-1)\frac{1}{(1-\bar{\theta})^2}$
$D_{\pi}(\theta)$	$1/\tau^2$	$(\alpha-1)\bar{\theta}^{-2}$	$(\alpha-1)\bar{\theta}^{-2}$	$\frac{(\alpha-1)}{\bar{\theta}^2} + \frac{(\beta-1)}{(1-\bar{\theta})^2}$
$D_q(m,\theta,\boldsymbol{y}_m)$	m/σ^2	$\frac{(\alpha/c+\sum y_i-1)}{\bar{\theta}^2}$	$\frac{(\alpha/c+m-1)}{\bar{\theta}^2}$	$\frac{(\frac{\alpha}{c}+\sum_i y_i-1)}{\bar{\theta}^2} + \frac{(\frac{\beta}{c}+m-\sum_i y_i-1)}{(1-\bar{\theta})^2}$
$ESS(\pi_b(\theta))$	$\sigma^2/c\tau^2$	0	0	0
$ESS(\pi(\theta))$	σ^2/τ^2	$\frac{\alpha-\alpha/c}{\bar{y}}$	$\alpha-\alpha/c$	$\alpha+\beta$

informations):

$$D_{\pi,j}(\boldsymbol{\theta}) = -\frac{\partial^2 \log(\pi(\boldsymbol{\theta}))}{\partial\theta_j^2},\ \ D_{q,j}(m,\boldsymbol{\theta},\boldsymbol{y}_m) = -\frac{\partial^2 \log(q_m(\boldsymbol{\theta}|\boldsymbol{y}_m))}{\partial\theta_j^2},\ j=1,...,d. \quad (2.22)$$

In what follows, we will sometimes use the simplified notations π, q_m in place of $\pi(\boldsymbol{\theta}), q_m(\boldsymbol{\theta}|\boldsymbol{y}_m)$ and $D_{\pi,j}, D_{q_m,j}$ in place of $D_{\pi,j}(\boldsymbol{\theta}), D_{q,j}(m,\boldsymbol{\theta},\boldsymbol{y}_m)$. Let $D_{\pi,+} = \sum_{j=1}^d D_{\pi,j}$ and $D_{q_m,+} = \sum_{j=1}^d \int D_{q_m,j} f(\boldsymbol{y}_m) d\boldsymbol{y}_m$ denote the global information for the prior π and the posterior q_m, respectively. The distance between the prior and the posterior for the sample size m is then defined as

$$\delta(m,\bar{\boldsymbol{\theta}},\pi,q_m) = |D_{\pi,+}(\bar{\boldsymbol{\theta}}) - D_{q_m,+}(\bar{\boldsymbol{\theta}})|, \quad (2.23)$$

evaluated in $\bar{\boldsymbol{\theta}} = E_\pi(\boldsymbol{\theta})$, the prior informative mean. The ESS for π is defined as

$$ESS(\pi(\boldsymbol{\theta})) = \underset{m\in\mathbb{N}}{\text{Argmin}}\{\delta(m,\bar{\boldsymbol{\theta}},\pi,q_m)\}. \quad (2.24)$$

When $d=1$, we will simply write D_π, D_{q_m}, suppressing the subscript '+'. Table 2.1 shows an example of the priors and the posteriors for four univariate conjugate models: Normal-Normal, Gamma-Poisson, Gamma-Exponential and Beta-Binomial. Note that, under the assumptions in (2.21), the baseline prior mean corresponds to the informative prior mean, and the hyperparameter c is a large constant chosen to inflate the variance of the informative prior. Table 2.2 reports the distances and the effective sample sizes for these univariate conjugate models. Similarly to the gen-

eral expression in (2.23), the distance between the MDD prior $\varphi(\theta)$ and the baseline posterior $q_m(\theta|\boldsymbol{y}_m)$ evaluated in $\bar{\theta} = E_\pi(\theta)$ is defined as

$$\delta(m, \bar{\theta}, \varphi, q_m) = |D_\varphi(\bar{\theta}) - D_{q_m}(\bar{\theta})|, \tag{2.25}$$

where D_φ has not in general a closed form and may be computed through an R routine. The effective sample size $ESS(\varphi(\theta))$ is computed for the MDD prior analogously as in (2.24). For the univariate conjugate models the following theorem holds.

Theorem 2. *Given $\theta \in \mathbb{R}$, the likelihood $f(\boldsymbol{y}_m|\theta)$, an informative prior $\pi(\theta)$, a baseline prior $\pi_b(\theta)$, the baseline posterior $q_m(\theta|\boldsymbol{y}_m)$ and the MDD prior $\varphi(\theta)$ defined in* (2.17)*, assume to be in a conjugate case and that the technical conditions in* (2.21) *hold. Then*

$$ESS(\varphi(\theta)) \leq ESS(\pi(\theta)) \tag{2.26}$$

For a formal proof, see Appendix A. Formula (2.26) provides an upper bound for the effective sample size of the MDD prior class, and yields an intuitive result. Although an analytic form of the ESS for this class of priors is not available, the interpretation is that whatever are the observed weights and the priors π_b, π used in the formulation, the information contained in the MDD prior is never greater than the information contained in π. From a practical point of view, this prior distribution provides a lower information than that contained in the prior π, and is then more likely to not dominate the likelihood.

As will be more clear from the simulation studies in Section 2.5, the ESS is a powerful tool for deciding whether the resampling procedure in Section 2.3.1 could provide any benefit in terms of posterior estimates. A natural suggestion could be using the ESS of the informative prior as a *threshold* for the resampling. $ESS(\pi(\theta)) >> m$ means that the informative prior is wildly informative and that generating further data from the supposed true model could neutralize its impact. Conversely, $ESS(\pi(\theta)) \leq m$ suggests that resampling should not yield any benefit, since the prior π does not provide an extra amount of information compared to the current sample size.

2.4.2 Distribution-constant statistics

In this section we frame the MDD priors approach within the general theoretical framework for the data-dependent priors proposed by Darnieder (2011) —and summarized in Section 2.2— and we draw an appealing theoretical comparison between the MDD priors and the Bayesian approach, under certain technical conditions.

As alluded in Section 2.2, one of the key-points of the Darnieder's approach concerns the choice of the statistic $T(\boldsymbol{y})$ on which conditioning the prior distribution. As widely explained in Section 2.3, the MDD prior depends on the data only through the Hellinger distance, defined in 2.3.1. For illustration purposes only and without loss of generality —the theorems listed below preserve their validity in a multidimensional case— let θ be a scalar parameter, $\theta \in \mathbb{R}$, and put $r^{(m)}(\theta, \theta + \triangle) \equiv \mathcal{H}(f(\boldsymbol{y}_m|\theta), f(\boldsymbol{y}_m|\theta + \triangle))$, where the parameters' difference $\triangle$ is not a parameter, but just a known quantity which may be computed for each m. Let $I_m(\theta; f) = mI(\theta)$ denote the Fisher information for the parametric family $\{f(\boldsymbol{y}_m; \theta) : \theta \in \Theta\}$ in case of independent observations. Borovkov and Moullagaliev (1998) state the following theorem. In what follows, let $r^{(m)}(\triangle)$ denote $r^{(m)}(\theta, \theta + \triangle)$ for simplicity purposes.

Theorem 3. *If the function $\sqrt{f(\boldsymbol{y}_m|\theta)}$ is differentiable with respect to $\boldsymbol{\theta}$, and $I_m(\theta; f)$ is continuous, then there exists the limit:*

$$\lim_{\triangle \to 0} \frac{r^{(m)}(\triangle)}{\triangle^2} = I_m(\theta; f) \tag{2.27}$$

This Theorem provides a limiting behaviour for the Hellinger distance in a neighborhood of 0 of the parameters' difference $\triangle$. Furthermore, they also provide some uniform bounds for $r^{(m)}(\triangle)/\triangle^2$:

Theorem 4. *If*

(i) *the parameters set Θ is compact;*

(ii) *$f(\boldsymbol{y}_m|\theta) \neq f(\boldsymbol{y}_m|\theta + \triangle)$ whenever $\triangle \neq 0$. Under this condition we have $r^{(m)}(\theta, \theta + \triangle) > 0$ for $\triangle \neq 0$;*

(iii) *$0 < I(\theta) \leq h < \infty$ for a given constant h.*

Then there exists a constant $g > 0$ *such that the following relation holds:*

$$g \leq \frac{r^{(m)}(\triangle)}{\triangle^2} \leq h, \quad \forall \theta \in \Theta. \tag{2.28}$$

Theorem (4) is stating that, for every choice of θ, $r^{(m)}(\triangle)$ is bounded between $g\triangle^2$ and $h\triangle^2$.

In our framework, the dependence of the MDD class to the data is represented by the observed Hellinger distance ψ_{m^*}. Then, a natural choice is to set $T(\boldsymbol{y}_{m^*}) = r^{(m^*)}(\triangle)$. If $I_{m^*}(\theta; f)$ does not depend on the parameter $\theta^{(m^*)}$ —this happens for instance for the Normal, LogNormal, Cauchy and Logistic distributions— from Theorem 3 the distribution of $T(\boldsymbol{y}_{m^*})$ does not depend on θ, but only on the parameters' difference $\triangle$, as $\triangle \to 0$. In other words, $T(\boldsymbol{y}_{m^*})$ is a distribution-constant statistic and Theorem 1 in Section 2.2.1 holds. We may summarize these results and state the following resuming theorem.

Theorem 5. *Given a parametric family of continuous distributions* $\{f(\boldsymbol{y}_{m^*}|\theta), \theta \in \Theta\}$ *for which the Fisher information* $I_{m^*}(\theta; f)$ *does not depend on* θ*, then the Hellinger distance* $r^{(m^*)}(\triangle)$ *does not depend on* $\theta^{(m^*)}$ *as* $\triangle \to 0$*, but only on the difference* $\triangle$*. This means that the statistic* $T(\boldsymbol{y}_{m^*}) = r^{(m^*)}(\triangle)$ *is distribution-constant and the MDD prior* (2.17) $\pi(\theta|T(\boldsymbol{y}_{m^*}))$ *reduces to a genuine prior* $\pi(\theta)$*.*

It is straightforward to show that, in this particular case, the MDD prior still depends on the data, but exhibits its dependence on the data only through conditioning on the sample size m, plus an augmented sample size $\varkappa$. And in such a case, as Darnieder (2011) suggests, there is no need of doing any adjustment, since the sample size m is intrinsic in the likelihood and does not convey any information about θ.

By concluding, we found some special cases in which conditioning the prior on a data statistic may be reduced to choosing a genuine Bayesian approach.

2.4.3 Approximation of a hierarchical model

As suggested by Gelman (2016*a*), data-dependent priors may sometimes be interpreted as an approximation of a hierarchical model, and in Section 2.2.2 we provided

a brief formalization of this intuition. Using again the Normal-Normal model as a toy example, let consider the following hierarchical model:

$$y_{ij} \sim \mathcal{N}(\theta_{j[i]}, \sigma^2),\ i = 1 \dots m,\ j = 1, \dots, J \tag{2.29}$$

$$\theta_j \sim \mathcal{N}(0, \tau_j^2) \tag{2.30}$$

$$\tau_j^2 = \zeta_j^2 \ \text{ with probabilities } \psi_j,\ \sum_{j=1}^{J} \psi_j = 1, \tag{2.31}$$

where the nested index $j[i]$ codes as usual in the hierarchical models (Gelman and Hill, 2006) the group membership for the statistical unit i; the group-level parameter θ_j is assigned a normal prior distribution; the prior variance τ_j^2 may assume in our framework $J = 2$ different values, $\tau_1^2 = c\zeta^2, \tau_2^2 = \zeta^2$, with probabilities ψ and $1 - \psi$; c, ζ^2 are for simplicity fixed hyperparameters. If we fit this model according to the Bayesian paradigm, we should also assign a prior distribution to the probability ψ, for instance $\psi \sim \mathcal{B}\text{e}(a, b)$, depending on some hyperparameters a, b. The MDD prior for θ, $\theta \sim \psi\mathcal{N}(0, c\zeta^2) + (1 - \psi)\mathcal{N}(0, \zeta^2)$, is another way for expressing equations (2.30), (2.31). We may then argue that the MDD class is a natural approximation of the model above, with the parameter ψ that is not assigned a prior but estimated from the data through the procedures described in Section 2.3.1. For illustration purposes only, Figure 2.2 displays the mean squared errors obtained from the posterior estimates of the hierarchical model (2.29), (2.30), (2.31) above and the MDD-res prior. We performed the computations using RStan (Stan Development Team, 2016*a*), the R (R Core Team, 2016) interface to the Stan C++ library (Stan Development Team, 2016*b*), with $c = 100,\ \zeta^2 = 1,\ \sigma^2 = 5,\ m = 5$ and for different values of the Beta hyperparameters $a,\ b$; the MDD prior globally shows lower MSEs as the true value θ_0 moves away from zero, the prior mean.

2.4.4 Model for the tuning parameter

As mentioned in Section 2.2.4, the relationship between the penalized likelihood approach and the Bayesian theory is related to interpreting the penalty as the kernel of a prior log-density. However, the estimation of the penalty weight related to the prior variance remains open. Hastie et al. (2002) suggest to use cross-validation, whereas

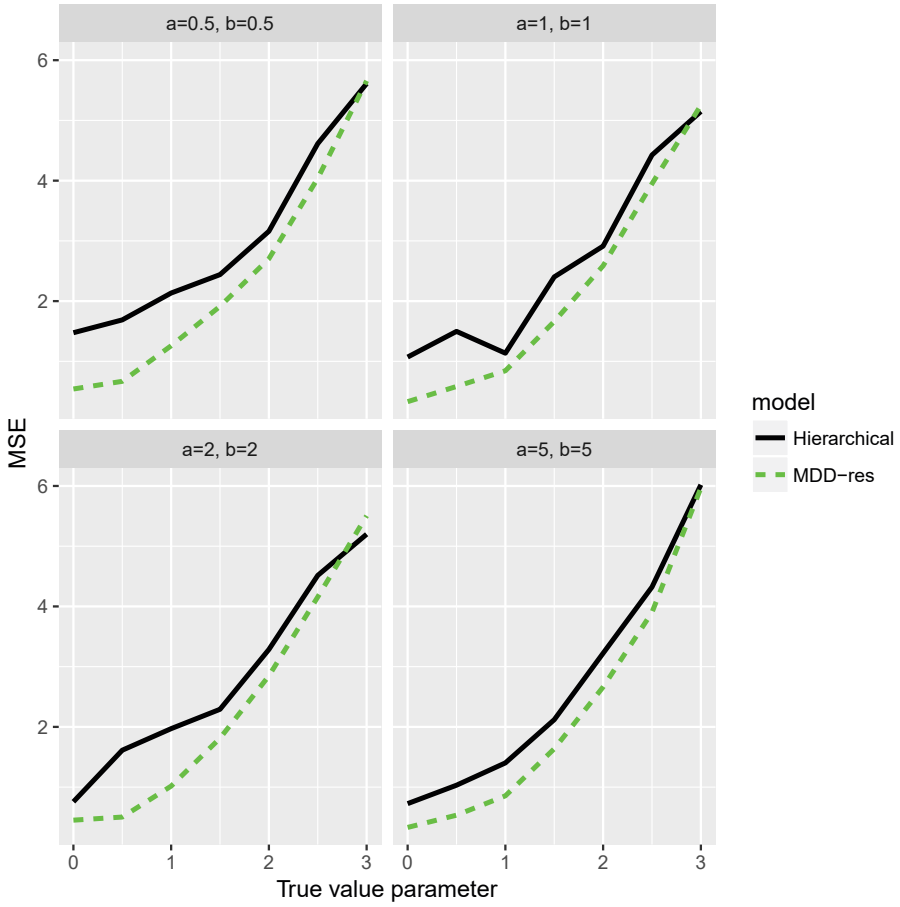


Figure 2.2: Comparison between the MSE of the hierarchical model (dashed black line) for different values of the Beta hyperparameters a, b and of the MDD-res prior (green line), with the weight ψ_{m^*} estimated from data. On the x-axis the true parameter value that generated the data. $c = 100$, $\zeta^2 = 1$, $\sigma^2 = 5$, $m = 5$. MSEs computed over 200 replications. The hierarchical model has been fitted using RStan (Stan Development Team, 2016*a*), the R (R Core Team, 2016) interface to the Stan C++ library (Stan Development Team, 2016*b*).

Efron (2012) propose empirical Bayes methods. Otherwise, Cole et al. (2013) set different values and examine the results for these different inputs. The MDD prior specification may be seen as a natural alternative for estimating the tuning parameter in the penalized likelihood approach. For illustration purposes only, let consider the regression model

$$y_i = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. And consider now the penalized log-likelihood with quadratic penalty for this model

$$l(\boldsymbol{\beta}; \boldsymbol{y}) - \frac{1}{2\tau^2}\boldsymbol{\beta}^2, \tag{2.32}$$

where $\beta_j \sim \mathcal{N}(0, \tau^2)$ according to the Bayesian interpretation of the Ridge regression. The penalty weight/tuning parameter is then $r = 1/\tau^2$, the inverse of the prior variance. Instead of estimating directly this factor, specifying a MDD prior for β_j is an automatic tool for introducing an auxiliary level for the variance, as in (2.31):

$$l(\boldsymbol{\beta}; \boldsymbol{y}) - \frac{1}{2\tau^2}\boldsymbol{\beta}^2, \tag{2.33}$$

$$\tau^2 = \begin{cases} \zeta^2 \ \text{ with } \psi \\ c\zeta^2 \text{ with } 1 - \psi. \end{cases} \tag{2.34}$$

Although we use the Normal-Normal model, this approach allows flexibility also for other types of prior distributions (Wood, 2017).

The penalized methods —Lasso, Ridge regression, etc.— are designed for reducing the mean squared errors, and the MDD class of priors represents a built-in method for addressing the same objective. Further work should be developed in order to implement the MDD priors for regression models and within the Bayesian variable selection framework.

2.5 Simulation studies

In this section we provide some numerical and graphical examples which (a) introduce and assess the frequentist coverage of the posterior credible intervals (Carlin and

Louis, 2000) and the mean squared errors obtained from the MDD, noninformative and informative priors and (b) clarify some theoretical results related to the notion of effective sample size (ESS) presented in Section 2.4 . Moreover, these practical examples contribute to understand the theory behind the MDD priors' formulation described in Section 2.3.

2.5.1 Mean squared errors and frequentist coverage

Given a generic prior $\pi(\theta)$, let $\pi(\theta|\boldsymbol{y}_m)$ be the corresponding posterior. Define $a_m(\alpha)$ by the relationship

$$\int_{-\infty}^{a_m(\alpha)} \pi(\theta|\boldsymbol{y}_m)d\theta = \alpha.$$

Let $A_m = (-\infty, a_m(\alpha)]$, then $Pr(A_m|y_1, \ldots, y_m) = \alpha$. Perhaps, let

$$\text{coverage}_\theta(A_m, \alpha) = \Pr(\theta \in A_m|\theta) \tag{2.35}$$

be the frequentist coverage of the Bayesian interval estimate A_m. According to this definition of Wasserman (2000), we say that the prior π generates *second-order correct intervals* if

$$\int L(\theta; \boldsymbol{y}_m)\pi(\theta)d\theta < \infty$$

almost surely for all m greater than some m_0 and

$$\text{coverage}_\theta(A_m, \alpha) = \alpha + O(1/m)$$

for every θ, where $L(\theta; \boldsymbol{y}_m)$ is the likelihood.

In this section we replicated $\boldsymbol{y}_m^{(b)} \sim f(\boldsymbol{y}_m|\theta_0),\ b = 1, ..., B$, under different choices for θ_0 and m —see Table 2.3 for details— and we counted how many times the true value parameter θ_0 is contained in the credible intervals obtained from the posterior distributions. In this manner we obtained the *actual* coverage $\hat{\alpha}$ and we compared it to the *nominal* coverage α through the coverage difference

$$\triangle\alpha = |\alpha - \hat{\alpha}|. \tag{2.36}$$

The smaller is this quantity, the more reliable is the credible interval according to frequentist criteria.

The frequentist coverage is often used as a performance tool of some Bayesian procedures, and it represents a powerful tool for assessing the goodness of the posterior estimates. However, we would need also an error measure for quantifying whether our posteriors yield wrong results. For this aim, the empirical mean squared error (MSE) for θ over the B samples is computed as

$$\text{MSE}_\theta = B^{-1} \sum_{b=1}^{B} (\hat{\theta}^{(b)} - \theta_0)^2, \tag{2.37}$$

where $\hat{\theta}^{(b)}$ is the posterior median for θ for the b-th sample.

Simulation scheme

(i) for $b = 1, \ldots, B$:

- replicate $\boldsymbol{y}_m^{(b)} \sim f(\boldsymbol{y}_m|\theta_0)$;
- derive the credible interval at level $\alpha = 0.95$, $A_m^{(b)} = (-\infty, a_m^{(b)}(\alpha)]$ for the posterior distribution arising from $\boldsymbol{y}_m^{(b)}$, $\pi(\theta|\boldsymbol{y}_m^{(b)})$;
- derive the posterior median $\hat{\theta}^{(b)}$.

(ii) compute the actual coverage $\hat{\alpha} = B^{-1} \sum_{b=1}^{B} |\theta_0 \in A_m^{(b)}|$ and then the coverage difference (2.36).

(iii) compute the empirical mean squared error (2.37).

(NN) Figure 2.3 displays the empirical MSEs for the Normal-Normal model computed for the informative posterior (black line), noninformative posterior (red line), MDD-natural posterior (blue line), and MDD-res posterior (green line), plotted against the true parameter value θ_0 that generated the data at hand, in correspondence of different sample sizes m and likelihood's variances σ^2. As a general consideration, the MSEs appear to be jointly sensitive to the sample size and the likelihood's variance: in fact, they rapidly grow for $m = 5$ and $\sigma^2 = 20$, as the true parameter value moves away from zero (the informative/noninformative prior mean). Conversely, they are uniformly flat for $m = 25$ and $\sigma^2 = 1$, even when the true value parameter is far from the prior means. This is intuitive, since a richer dataset tends to adjust the posteriors,

Table 2.3: True value parameter which generated data; sample size; baseline and informative priors for the four univariate conjugate models used in the simulation study and ESS for the informative prior. MSEs and frequentist coverages are computed over 200 samples.

	NN	*GP*	*GExp*	*BB*
True θ_0	$0, 0.5, 1, \ldots, 3$	$1, 1.5, \ldots, 4$	$1, 1.5, \ldots, 4$	$[0, 1]$
m	5,10,15,25	1,2,3,4	1,2,3,4	1,2,3,4
$\pi_b(\theta)$	$\mathcal{N}(0, 1000)$	$\mathcal{G}a(0.4, 0.2)$	$\mathcal{G}a(0.4, 0.2)$	$\mathcal{B}e(0.02, 0.18)$
$\pi(\theta)$	$\mathcal{N}(0, 1)$	$\mathcal{G}a(400, 200)$	$\mathcal{G}a(400, 200)$	$\mathcal{B}e(20, 180)$
$E_b(\theta) = E(\theta)$	0	2	2	0.1
$ESS(\pi(\theta))$	$1, 5, 10, 15$	$399.6/\bar{y}$	399.6	200

which appear then to be less distinguished. In terms of performance, the MDD-natural and the MDD-res register similar MSEs: in this case, resampling does not seem to be beneficial. We may derive similar conclusions from the frequentist coverage displayed in Figure 2.4, where the credible intervals provided by the informative posterior fail in covering the true parameter value as this moves away from zero, whereas the MDD-natural behaves better than the MDD-res and often approximates the noninformative posterior.

(GExp) In Figure 2.5 the MSEs for the Gamma-Exponential model for the informative, MDD-natural and MDD-res posteriors behave quite analogously: they decrease in correspondence of the prior mean $\alpha/\beta = 2$, and they rapidly grow as the true parameter moves away from this value. Conversely, the MSEs registered by the noninformative posterior tend to be generally higher than those registered by the other posteriors. Here, the chosen sample sizes are really small. One could be tempted to conclude that also in this case the resampling does not provide any benefit. But the frequentist coverage plotted in Figure 2.6 suggests something different. Here, the noninformative posterior registers the lowest coverages' differences for each sample size, but the behavior of the MDD-res is different from the MDD-natural. The latter seems to strictly follow the informative posterior, which yields a good coverage difference only in correspondence of the prior mean, but is extremely high elsewhere. Whereas the MDD-res yields lower values and appear to be closer to the noninformative

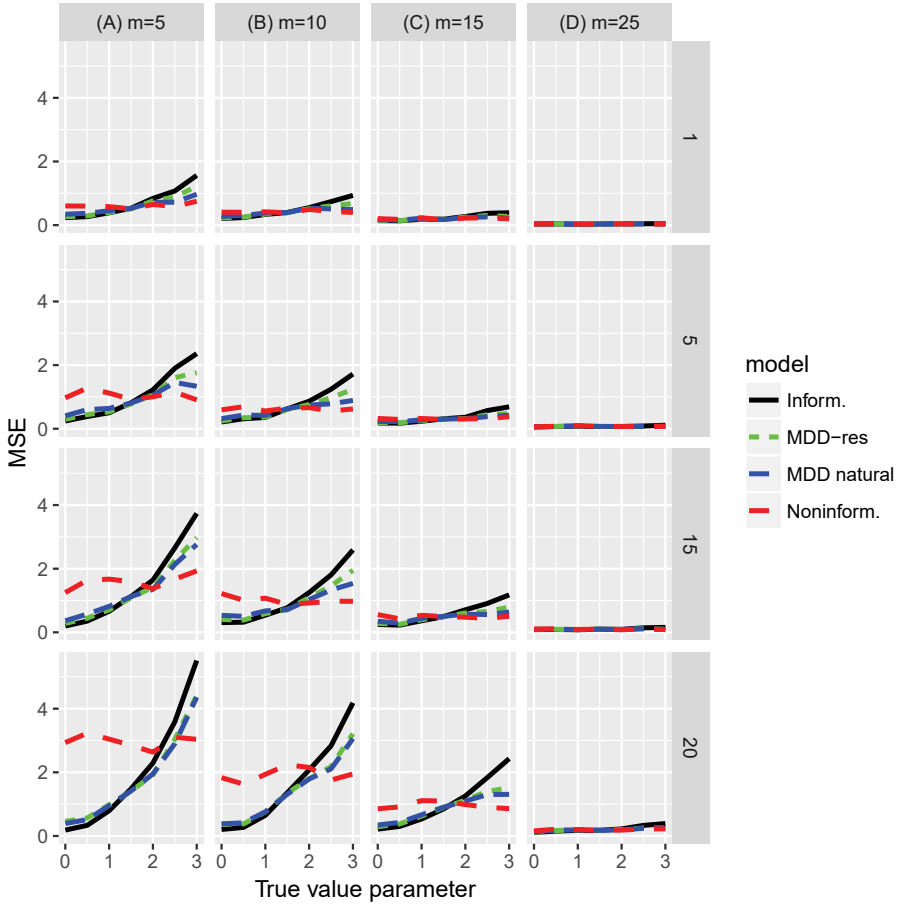


Figure 2.3: Normal-Normal model: $\boldsymbol{y}_m \sim \mathcal{N}(\theta_0, \sigma^2)$, $\pi_b(\theta) = \mathcal{N}(0, 1000)$, $\pi(\theta) = \mathcal{N}(0, 1)$. True value parameter θ_0 (x-axis) and MSEs obtained from π, π_b, MDD-natural φ, MDD-res φ in correspondence of $\sigma^2 = (1, 5, 10, 15)$.

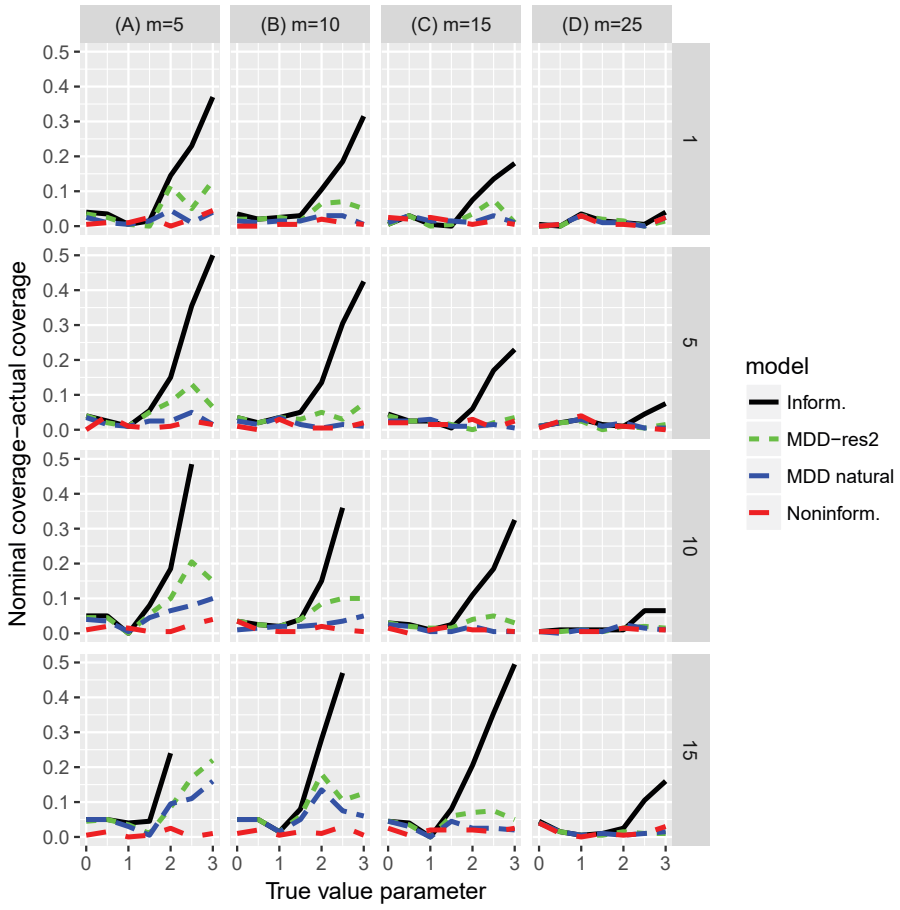


Figure 2.4: Normal-Normal model: $\boldsymbol{y}_m \sim \mathcal{N}(\theta_0, \sigma^2)$, $\pi_b(\theta) = \mathcal{N}(0, 1000)$, $\pi(\theta) = \mathcal{N}(0, 1)$. True value parameter θ_0 (x-axis) and coverage difference obtained from π, π_b, MDD-natural φ, and MDD-res φ in correspondence of $\sigma^2 = (1, 5, 10, 15)$. (Values exceeding 0.5 are removed from the plot).

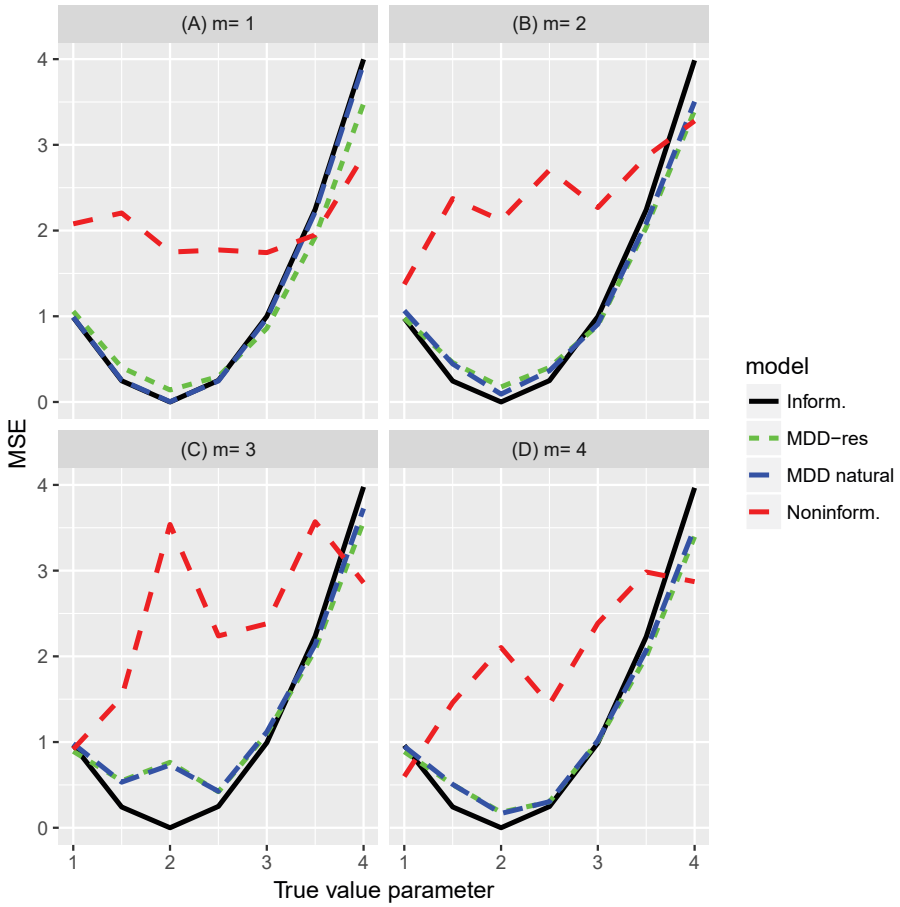


Figure 2.5: Gamma-Exponential model: $\boldsymbol{y}_m \sim \mathcal{E}xp(\theta)$, $\pi_b(\theta) = \mathcal{G}a(0.4, 0.2)$, $\pi(\theta) = \mathcal{G}a(400, 200)$. True value parameter θ_0 (x-axis) and MSE obtained from π, π_b, MDD-natural φ, MDD-res φ.

posterior.

(GP) The considerations made for the Gamma-Exponential model are almost identical for the Poisson-Gamma model, whose MSEs and frequentist coverages are displayed in Figure 2.7 and Figure 2.8, respectively.

(BB) The performances of the MSEs for the Beta-Binomial model in Figure 2.9 show a different trend if compared with the MSEs of the previous models. The MDD-natural tends to overlap the informative posterior for each sample size, whereas the beneficial of the resampling appears evident in the trend for the MDD-res, which approximates the noninformative posterior and sometimes is even preferable. Also the coverage differences displayed in Figure 2.10 highlight

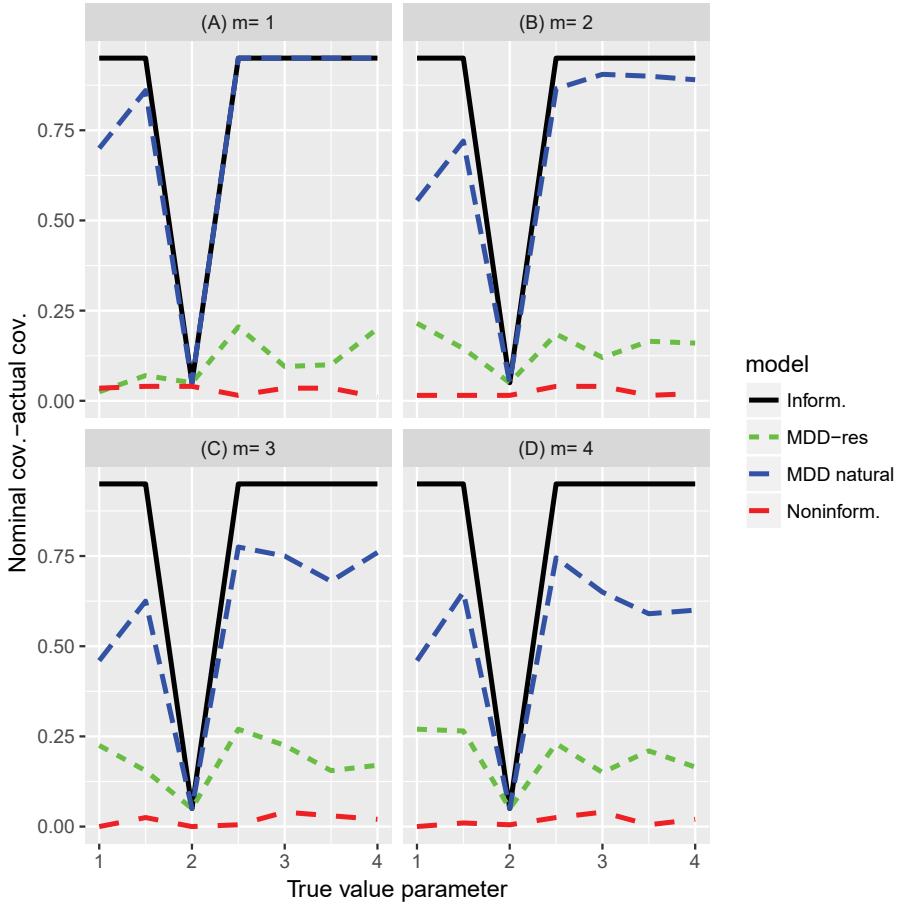


Figure 2.6: Gamma-Exponential model: $\boldsymbol{y}_m \sim \mathcal{E}xp(\theta)$, $\pi_b(\theta) = \mathcal{G}a(0.4, 0.2)$, $\pi(\theta) = \mathcal{G}a(400, 200)$. True value parameter θ_0 (x-axis) and coverage difference obtained from π, π_b, MDD-natural φ, MDD-res φ.

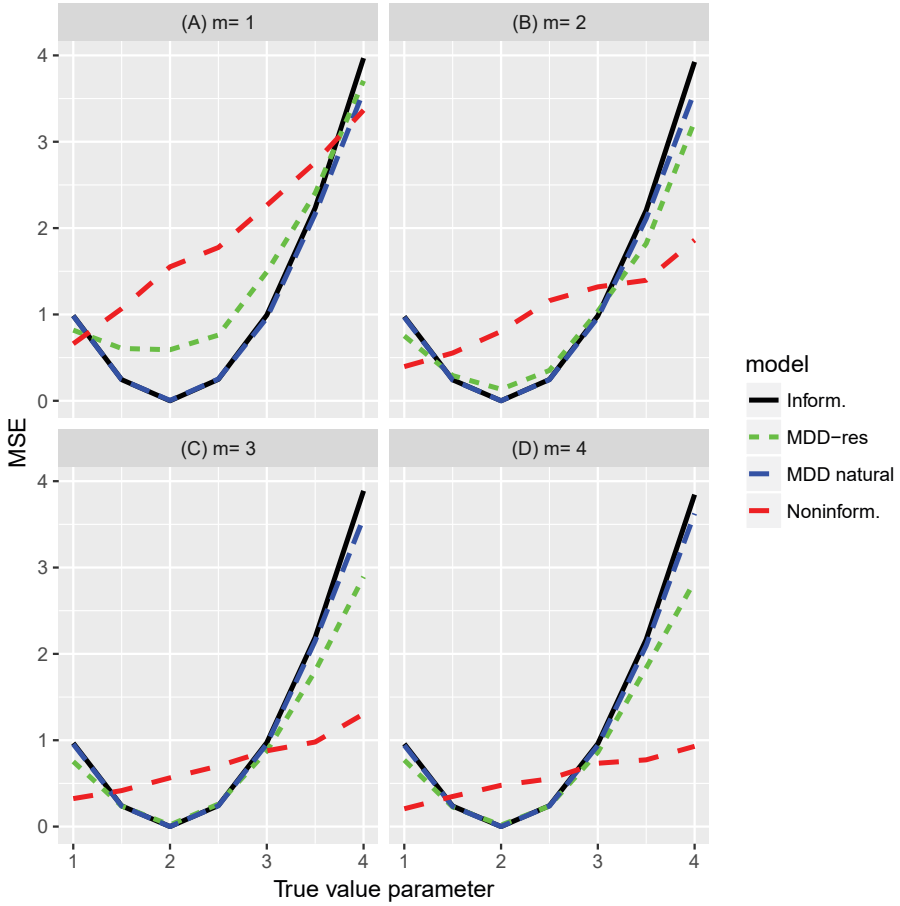


Figure 2.7: Gamma-Poisson model: $\boldsymbol{y}_m \sim \mathcal{P}ois(\theta)$, $\pi_b(\theta) = \mathcal{G}a(0.4, 0.2)$, $\pi(\theta) = \mathcal{G}a(400, 200)$. True value parameter θ_0 (x-axis) and MSE obtained from π, π_b, MDD-natural φ, MDD-res φ.

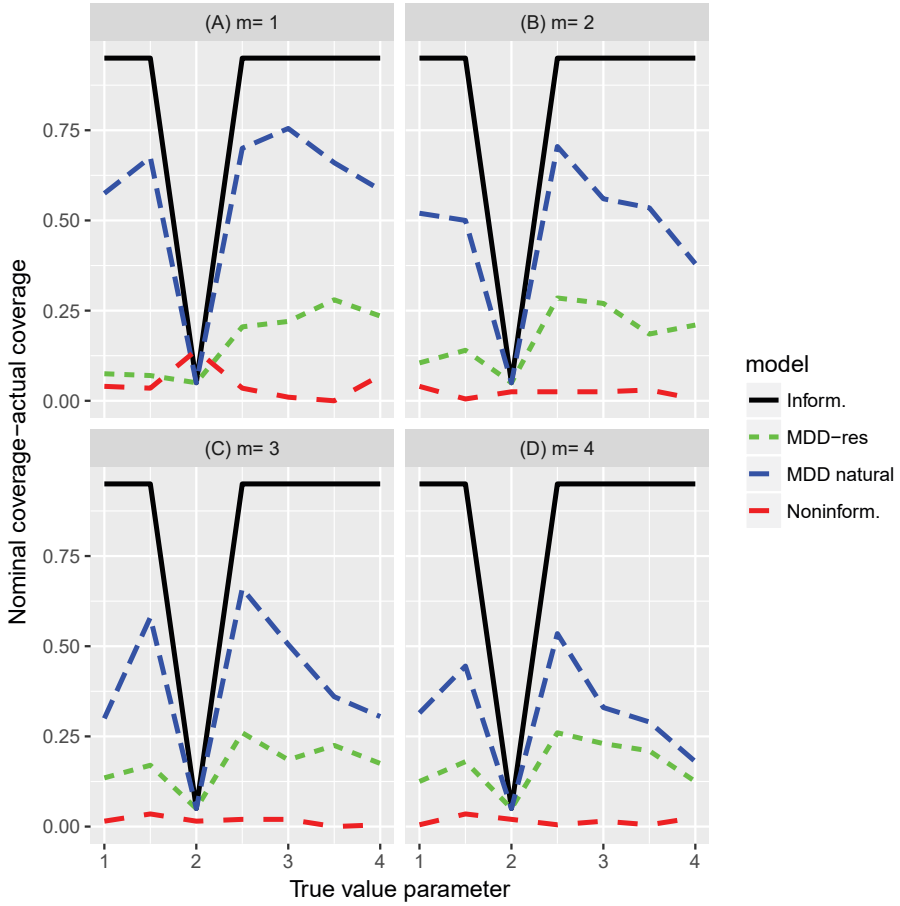


Figure 2.8: Gamma-Poisson model: $\boldsymbol{y}_m \sim \mathcal{P}ois(\theta)$, $\pi_b(\theta) = \mathcal{G}a(0.4, 0.2)$, $\pi(\theta) = \mathcal{G}a(400, 200)$. True value parameter θ_0 (x-axis) and coverage difference obtained from π, π_b, MDD-natural φ, MDD-res φ.

a global improvement for the MDD-res, which often registers the lowest values.

As a general comment of these simulation studies, we may argue that, apart for the Normal-Normal model, the MDD-res is well suited for reducing the MSEs and the coverage differences in correspondence of small sample sizes. In fact, by taking a joint look at these two quantities for each of the three other models, it clearly appears that the MDD-res, along with the informative and the MDD-natural, overcomes the noninformative prior in terms of MSEs, at least in a region of the true parameter value close enough to the informative/nonnformative prior mean; at the same time, his performance in terms of frequentist coverage tends to be much better than those provided by the informative and by the MDD-natural.

An empirical justification of these results is provided by the effective sample sizes provided by the different models. In the Normal-Normal model the effective sample size amounts to $\sigma^2/\tau^2 = 1, 5, 10, 15$ according to the different input values for the variances. The comparison with the input sample sizes $m = 5, 10, 15, 25$ highlights the usefulness of the resampling: the current sample size is in fact already able to absorb and neutralize the impact of the informative prior. The informative effective sample sizes for the Gamma-Exponential, Gamma-Poisson and Beta-Binomial model amount respectively to $\alpha - \alpha/c = 399.6$, $(\alpha - \alpha/c)/\bar{y}$ and $\alpha + \beta = 200$. Compared to the current sample sizes $m = 1, 2, 3, 4$, they are sensitively huge: here the resampling is beneficial for neutralizing the impact of a wildly informative prior distribution, and the MDD-res is preferable to the MDD-natural but, more generally, overcomes in our opinion the noninformative choice as well.

2.5.2 Effective sample size

In this simulation framework, let $\theta \in \mathbb{R}$ denote the generic parameter of interest, and $\pi_b(\theta), \pi(\theta)$ the baseline and the informative priors. The MDD prior is

$$\varphi(\theta) = \psi_{m^*}\pi_b(\theta) + (1 - \psi_{m^*})\pi(\theta),$$

with mixture weight ψ_{m^*}. For illustration purposes only, the mixture weights in this section are fixed in advance and no procedure —natural or resampling— is applied in this case for specifying them. We may compute the distance (2.25) between the MDD prior $\varphi(\theta)$ and the baseline posterior $q_m(\theta|\boldsymbol{y}_m)$ evaluated in $\bar{\theta} = E_\pi(\theta)$. The

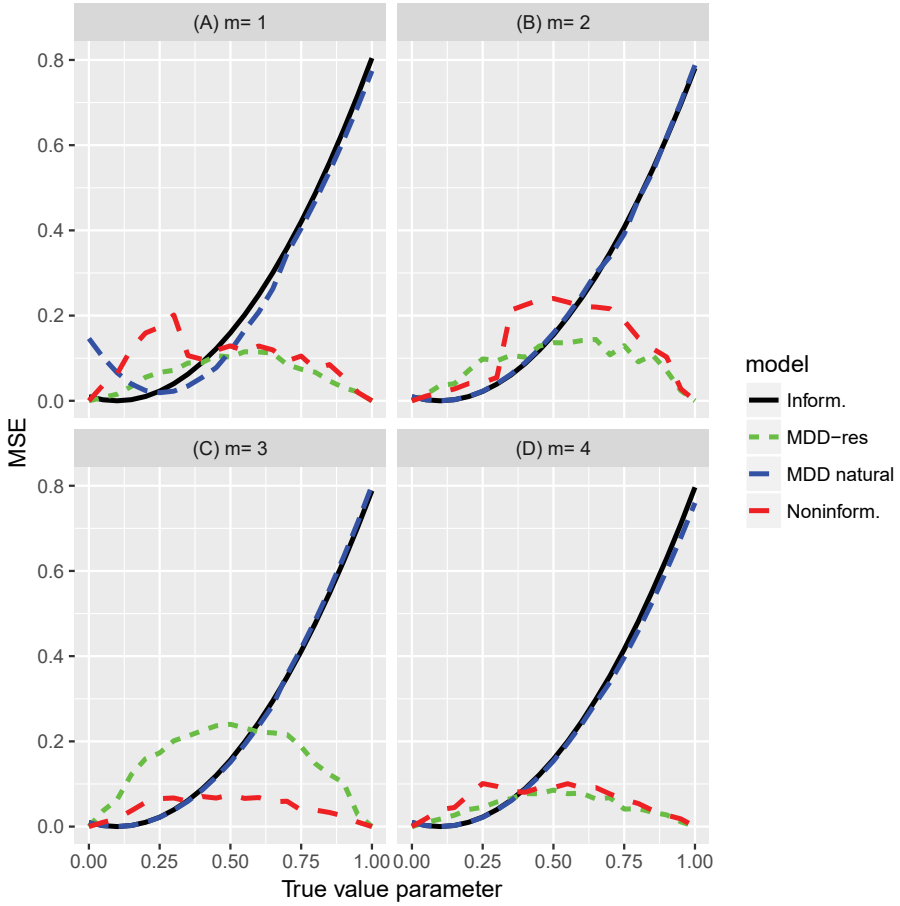


Figure 2.9: Beta-Binomial model: $\boldsymbol{y}_m \sim \mathcal{B}in(m, \theta)$, $\pi_b(\theta) = \mathcal{B}e(0.02, 0.18)$, $\pi(\theta) = \mathcal{B}e(20, 180)$. True value parameter θ_0 (x-axis) and MSE obtained from π, π_b, MDD-natural φ, MDD-res φ.

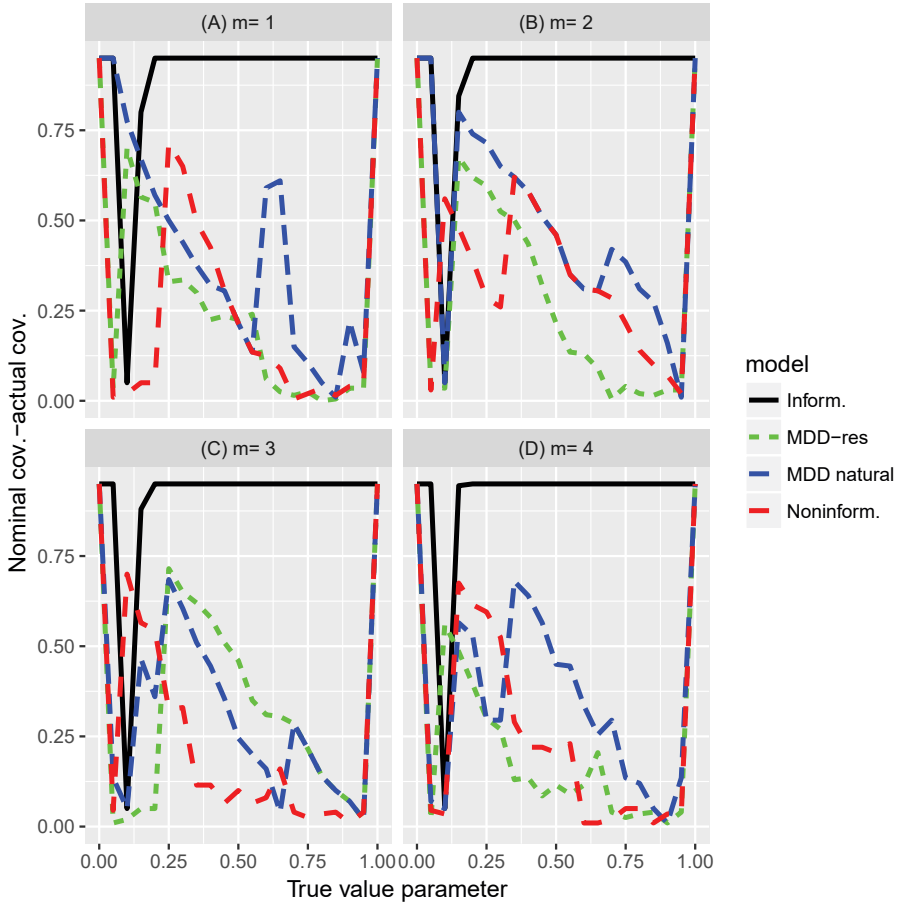


Figure 2.10: Beta-Binomial model: $\boldsymbol{y}_m \sim \mathcal{B}in(m, \theta)$, $\pi_b(\theta) = \mathcal{B}e(0.02, 0.18)$, $\pi(\theta) = \mathcal{B}e(20, 180)$. True value parameter θ_0 (x-axis) and coverage difference obtained from π, π_b, MDD-natural φ, MDD-res φ.

negative second derivative of the MDD log-density D_φ has not in general a closed form and it is computed through an R routine. We simulate $m = 5$ initial data from $f(\boldsymbol{y}_m|\theta_0) = \mathcal{N}(\theta_0, \sigma^2)$, with the true mean $\theta_0 = 0$ and variance $\sigma^2 = 15$. We choose $\pi_b(\theta) = \mathcal{N}(0, c)$, $\pi(\theta) = \mathcal{N}(0, 1)$: according to the technical conditions in (2.21), both the priors π_b, π are centered at the same mean, here $\mu = 0$, with informative variance equals 1 and baseline prior variance set to c.

From Section 2.4 we know that the ESS carried by the MDD prior class is always lower than the ESS for the informative prior. Now it is of interest for us assessing *how much less information* is provided by the MDD prior in correspondence of specific values of the mixture weights and of the hyperparameter c. Figure 2.11 shows the effective sample sizes for the informative, noninformative and the MDD prior with different mixture weights against the hyperparameter c, which inflates the informative variance. As a first consideration, this plot is a confirm of Theorem 2: the effective sample size for the MDD prior φ is never greater than the effective sample size for π. $ESS(\pi(\theta)) = \sigma^2/\tau^2 = 15/1 = 15$, while the effective sample size for the baseline prior clearly depends on the value of c, being σ^2/c. As is intuitive, the MDD class is sensitive to the choice of ψ_{m^*} and c . As the mixture weight increases and the baseline prior is then favored, the information of the MDD prior decreases. Less intuitive is the behavior of the MDD class in function of the hyperparameter c. Each of the MDD in the plot is stepwise increasing with c and this deserves a quick technical consideration. By definition (2.25), we are using the second derivatives of the log-densities, which means differentiate twice the function $\log(\varphi(\theta)) = \log(\psi_{m*}\pi_b(\theta)+(1-\psi_{m^*})\pi(\theta))$: the greater is c, the flatter becomes π_b, and consequently the smaller is the contribute in terms of mass probability carried by the baseline prior. Furthermore, the logarithm contributes to shrink the values of φ. Hence, the curvature of $\log(\varphi(\theta))$ will approximate the curvature of $\log(\pi(\theta))$ as c increases, and the information carried by the MDD and the informative prior will tend to be closer. This counterintuitive fact may be read in two directions: we could need another notion of distance, possibly less sensitive to the values of the hyperparameter c —taking, for instance, the second derivatives of the densities instead of the second derivative of the log densities— or we could use another kind of baseline prior, which does not depend on c (improper prior, Jeffreys prior,...). We address this second issue in the next section.

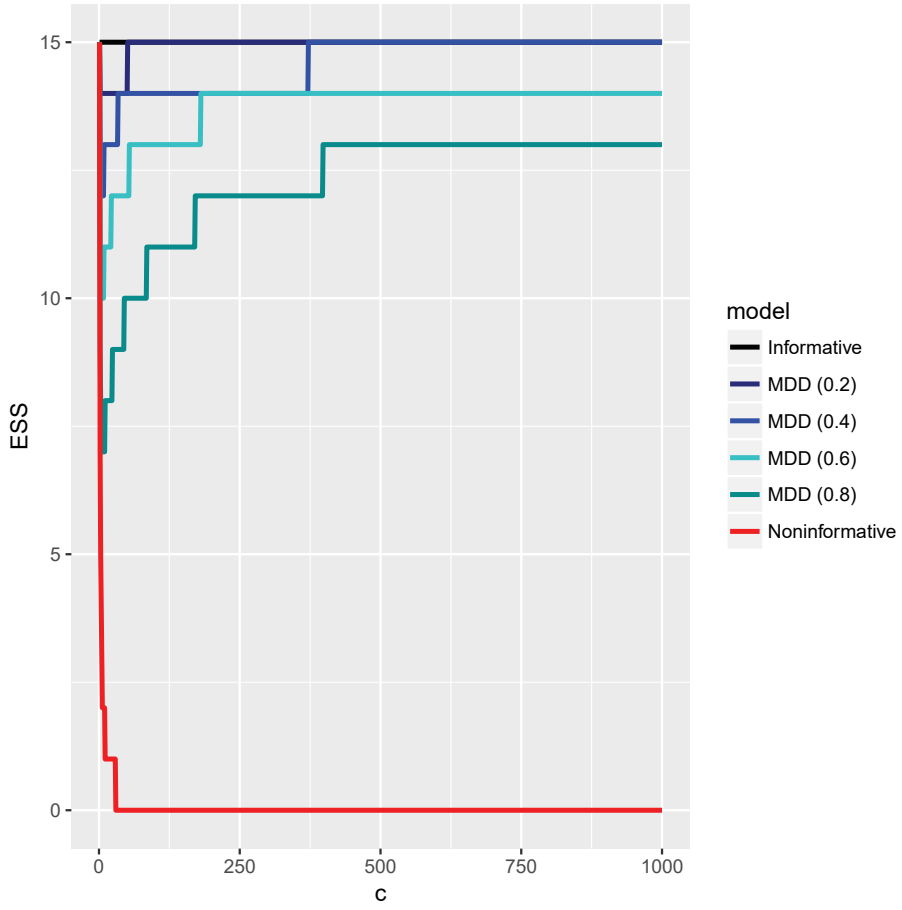


Figure 2.11: Normal-Normal model: $\boldsymbol{y}_m \sim \mathcal{N}(\theta_0, 15)$, $\pi_b(\theta) = \mathcal{N}(0, c)$, $\pi(\theta) = \mathcal{N}(0, 1)$. Effective sample sizes plotted against the hyperparameter c. $ESS(\pi(\theta)) = 15$.

2.6 Examples to Some Nonstandard Models

In the previous sections we dealt with two priors π and π_b belonging to the same family, under the technical condition (2.21). This is a common choice, adopted for instance in Morita et al. (2008), that allows for simply raising the noninformative variance by a factor c and falling into the conjugate models. However, one may be interested in exploring other prior choices for π_b, possibly automatic priors, and attempt to measure the information carried by the MDD prior (2.17), by taking unchanged the informative prior π. Or one may be interested in exploring non-conjugate models, setting an MCMC sampler. In this section we explore the first possibility and we focus on the corresponding amount of priors' information through a toy example and through a real case from a phase I trial study. But it is of future interest explore the non-conjugate models as well.

2.6.1 Jeffreys prior for an exponential model

Let $\boldsymbol{y}_m = (y_1, ..., y_m) \underset{iid}{\sim} \mathcal{E}xp(\theta)$, with $\pi(\theta) = \mathcal{G}a(\alpha, \beta)$. The likelihood is then

$$L_m(\theta; \boldsymbol{y}_m) = \prod_{i=1}^{m} f(y_i) = \theta^m \exp(-\theta \sum_{i=1}^{m} y_i). \tag{2.38}$$

We introduce the Fisher information for the exponential model computed for a single observation:

$$I_\theta = E\left[-\frac{d^2 \log f(y;\theta)}{d\theta^2}\right] =$$

$$= -E\left[\frac{d^2}{d\theta^2}\left[\log(\theta) - \theta y\right]\right] = -E\left[\frac{d}{d\theta}[1/\theta - y]\right] = E\left[\frac{1}{\theta^2}\right] = \frac{1}{\theta^2}.$$

Let $\pi_b(\theta) = j(\theta)$, where $j(\theta) = \sqrt{I_\theta}$ is the Jeffreys prior. For the exponential model, the Jeffreys prior for θ is

$$j(\theta) = \sqrt{I_\theta} = 1/\theta. \tag{2.39}$$

Now we compute the Jeffreys posterior $q_m(\theta|y_1, ..., y_m) = j_m(\theta|y_1, ..., y_m)$:

$$j_m(\theta|\boldsymbol{y}_m) \propto j(\theta) L_m(\theta; \boldsymbol{y}_m) = \theta^{-1} \prod_{i=1}^{m} \theta \exp\{-\theta y_i\} = \theta^{m-1} \exp\{-\theta \sum_{i=1}^{m} y_i\}. \tag{2.40}$$

We immediately realize that this is the kernel of a Gamma distribution, $\mathcal{G}a(m, \sum_i y_i)$:

$$j_m(\theta|\boldsymbol{y}_m) = \frac{(\sum_i y_i)^m}{\Gamma(m)}\theta^{m-1}\exp\{-\theta\sum_{i=1}^{m} y_i\}.$$

We compute the negative second log derivative of $j_m(\theta|\boldsymbol{y}_m)$ and we find the familiar result for a Gamma distribution

$$D_{j_m} = -\frac{d^2}{d\theta^2}\left[j_m(\theta|\boldsymbol{y}_m)\right] = \frac{m-1}{\theta^2}. \tag{2.41}$$

We code with the symbol $\varphi^j(\theta)$ the MDD prior built using the Jeffreys prior as baseline; the informative prior is set to $\pi(\theta) = \mathcal{G}a(4, 1)$. Finally, by using the plug-in estimate $\bar{\theta} = \alpha/\beta$, we may compute:

- the distance (2.23) between the informative prior π and the Jeffreys posterior j_m, with a corresponding sample size $ESS(\pi(\theta)) = \alpha = 4$;
- the distance (2.23) between the informative prior π and the baseline Gamma posterior q_m, resulting from the baseline prior $\mathcal{G}a(\alpha/c, \beta/c)$, with a corresponding sample size $ESS(\pi(\theta)) = \alpha - \alpha/c$;
- the distance between the Jeffreys prior j and the Jeffreys posterior j_m;
- the distance (2.25) between the MDD prior φ^j and the Jeffreys posterior j_m.

Figure 2.12 shows the effective sample sizes associated to the list above, plotted against the hyperparameter c; the mixture weights in the MDD class are fixed in advance for a sensitivity analysis. $ESS(\varphi^j(\theta))$ is always bounded between the effective sample sizes respectively for j and π, and results to be obviously constant for each value of the hyperparameter c. As is intuitive, as the mixture weight increases, $ESS(\varphi^j(\theta))$ gets closer to $ESS(j(\theta))$, the effective sample size for the baseline Jeffreys prior. Whereas, analogously to what happened in Section 2.5.2 for the Normal-Normal model, the MDD prior $\varphi(\theta)$ is increasing with c. In some sense, as already mentioned, we would like to observe the inverse relation: the vaguer the baseline, the lower should be the information of the MDD prior. We have already observed and discussed this issue in Section 2.5.2: we may now conclude that the use of automatic/improper priors —when this use is possible— in the MDD formulation keeps the baseline information constant and avoids an information growth depending

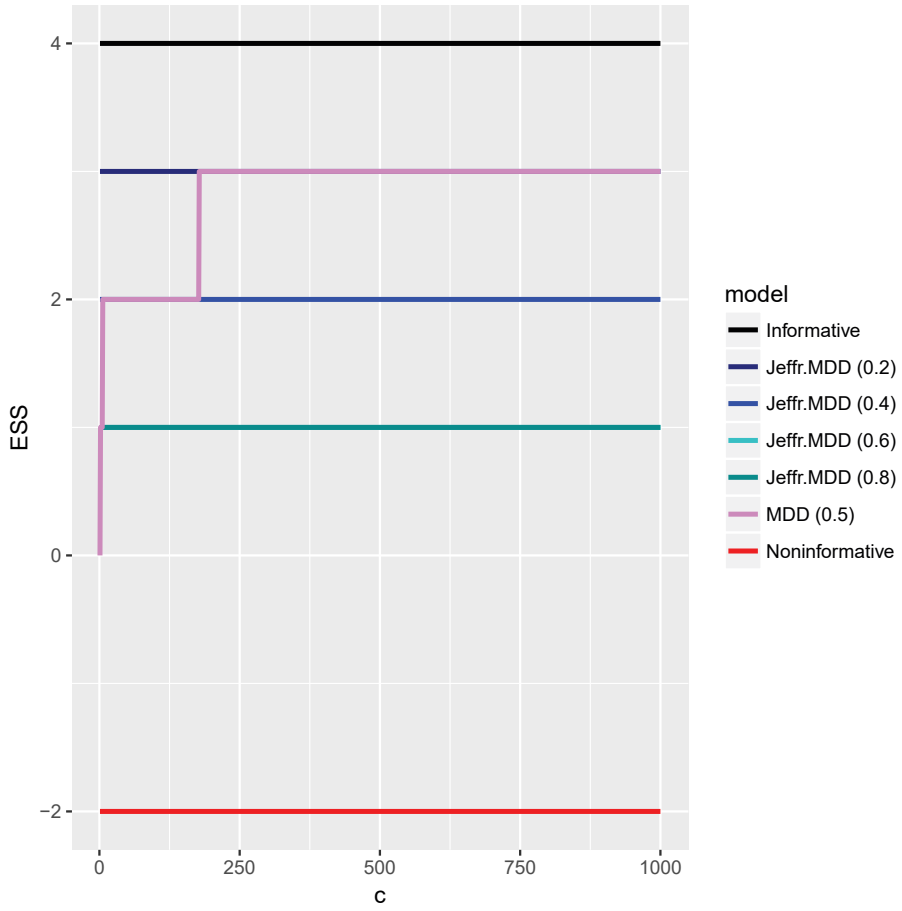


Figure 2.12: Gamma-Exponential model: $\boldsymbol{y}_m \sim \mathcal{E}xp(\theta)$, $j(\theta) = \theta^{-1}$, $\pi(\theta) = \mathcal{G}a(4,1)$, $\pi_b(\theta) = \mathcal{G}a(\alpha/c, \beta/c)$. Effective sample sizes plotted against the hyperparameter c. $ESS(\pi(\theta)) = \alpha = 4$.

on a further inflating hyperparameter c. Probably, using a baseline depending on a factor c results in a *partially* noninformative prior, rather than eliciting a Jeffreys —or another improper— prior. We obtain a further confirm of this idea in the next subsection: improper priors in the MDD formulation yield lower values for the global amount of information.

2.6.2 Logistic regression for phase I trial

Thall and Lee (2003) proposed a logistic regression to determine the greatest amount of tolerable dose in a phase I trial. In this section we follow the approach of Morita et al. (2008), who used the same example for studying the properties of the effective

sample size for different values of the hyperparameters.

The level of dose which each patient may receive is one among 100, 200, 300, 400, 500, 600 mg/m^2, denoted by $x_1, \ldots, x_6$. These values are then standardized on the log scale and denoted with $X_1, ..., X_6$. The response variable is $y_i = 1$ if patient i suffers toxicity, $y_i = 0$ if not. They assume the following logistic model:

$$P(y_i = 1) \equiv \pi(X_i, \boldsymbol{\theta}) = logit^{-1}(\mu + \beta X_i),\ i = 1, ..., m, \tag{2.42}$$

where $logit^{-1}(x) = e^x/(1 + e^x)$. Unlike for the conjugate models considered in Section 2.4.1, here the dimension of the parameters' space is $d = 2$, $\boldsymbol{\theta} = (\mu, \beta)$, where μ is the intercept of the linear predictor and β is the coefficient associated to the different levels of the doses. In order to compute the effective sample size, we need the extension to the multivariate case outlined by Morita et al. (2008). The likelihood for a sample of m patients $\boldsymbol{y}_m = (y_1, ..., y_m)$ is

$$f(\boldsymbol{y}_m|X, \boldsymbol{\theta}) = \prod_{i=1}^{m} \pi(X_i, \boldsymbol{\theta})^{y_i}(1 - \pi(X_i, \boldsymbol{\theta}))^{1-y_i}. \tag{2.43}$$

Thall and Lee (2003) elicited two independent informative priors for μ and β based on preliminary sensitivity analysis:

$$\begin{aligned} \mu &\sim \pi(\mu) = \mathcal{N}(\tilde{\mu}_\mu, \tilde{\sigma}^2_\mu) = \mathcal{N}(-0.11313, 2^2) \\ \beta &\sim \pi(\beta) = \mathcal{N}(\tilde{\mu}_\beta, \tilde{\sigma}^2_\beta) = \mathcal{N}(2.3980, 2^2). \end{aligned} \tag{2.44}$$

Hence, the baseline joint prior for $\boldsymbol{\theta}$ is $\pi(\boldsymbol{\theta}) = \mathcal{N}(\tilde{\mu}_\mu, c\tilde{\sigma}^2_\mu)\mathcal{N}(\tilde{\mu}_\beta, c\tilde{\sigma}^2_\beta)$, where the hyperparameter c is fixed at 10000. We follow the steps of the algorithm formulated by Morita et al. (2008) for determining (i) the effective sample size of each subvector and (ii) the global effective sample size of the parameter vector $\boldsymbol{\theta}$ as those values which respectively minimize the distances $\delta_1(m_\mu, \bar{\boldsymbol{\theta}}, \pi_\mu, q_{m_\mu}), \delta_2(m_\beta, \bar{\boldsymbol{\theta}}, \pi_\beta, q_{m_\beta})$ and $\delta(m, \bar{\boldsymbol{\theta}}, \pi, q_m)$, by using the plug-in vector $\bar{\boldsymbol{\theta}} = (\tilde{\mu}_\mu, \tilde{\mu}_\beta)$. See the Appendix A for a deep illustration of the algorithm. In this way, we compute the effective sample size of each parameter's subvector and then the global effective sample size of the logistic model. Given the two priors π_μ, π_β in (2.44), we will denote the first two quantities with $ESS(\pi(\mu)), ESS(\pi(\beta))$, and the third one simply with $ESS(\pi(\boldsymbol{\theta}))$. Table 2.4 reports these effective sample sizes, obtained replicating the experiment of Morita et al. (2008) and evaluated with respect to different values of the priors variances

$\sigma^2_\mu, \sigma^2_\beta$. As intuitive, the information contained in the prior distributions decreases as the variances increase. In any case, the parameter β, associated to the effect of the doses, yields a greater knowledge than the parameter μ, which represents the average response. We repeat the same steps above, but eliciting two MDD priors for the vector parameter $\boldsymbol{\theta}$:

$$\begin{aligned} \mu \sim \varphi(\mu) &= \psi\mathcal{N}(\tilde{\mu}_\mu, c\tilde{\sigma}^2_\mu) + (1-\psi)\mathcal{N}(\tilde{\mu}_\mu, \tilde{\sigma}^2_\mu) \\ \beta \sim \varphi(\beta) &= \psi\mathcal{N}(\tilde{\beta}_\beta, c\tilde{\sigma}^2_\beta) + (1-\psi)\mathcal{N}(\tilde{\mu}_\beta, \tilde{\sigma}^2_\beta), \end{aligned} \tag{2.45}$$

where the hyperparameter c is fixed at 10000 as before and ψ is the mixture weight. Being in absence of actual data at hand, here we do not assume a MDD-natural or a MDD-res obtained through the procedures described in Section 2.3.1. We limit our attention to the global information of the mixture formulation for generic values of the mixture weights, fixed in advance. Thus, for illustration purposes only, we drop the subscript m^* and we consider three different values for ψ, $\psi = \{0.2, 0.5, 0.8\}$. Then, we compare the so obtained results with those obtained with the above mentioned prior distributions. As may be noticed from Table 2.5, as ψ increases the effective sample sizes for the MDD priors (2.45) slightly decrease, as expected. However, the values obtained under these mixture priors are quite close to those obtained under the above priors $\pi(\mu),\ \pi(\beta)$ originally chosen by Thall and Lee (2003). It would be worth assessing how much varies the information of the mixture priors φ by choosing other baseline priors instead of flat normal distributions. Let us consider now two improper priors, $\pi_b(\mu) \propto 1,\ \pi_b(\beta) \propto 1$. The resulting MDD priors $\varphi^j(\mu),\ \varphi^j(\beta)$ are then defined as

$$\begin{aligned} \mu \sim \varphi^j(\mu) &= \psi + (1-\psi)\mathcal{N}(\tilde{\mu}_\mu, \tilde{\sigma}^2_\mu) \\ \beta \sim \varphi^j(\beta) &= \psi + (1-\psi)\mathcal{N}(\tilde{\mu}_\beta, \tilde{\sigma}^2_\beta). \end{aligned} \tag{2.46}$$

Table 2.6 reports the effective sample sizes for the priors in (2.46). In this case, there is an evident decrease of the information associated to the MDD priors φ: as ψ increases and the improper priors are then preferred, the effective sample size rapidly decreases. This is intuitive, since the improper priors which appear in (2.46) provide less information than the two flat normal priors in (2.45).

The example suggests that even inflating the informative variances by a great factor c does not affect in a sensible way the amount of information contained in the

Table 2.4: Effective sample sizes $ESS(\pi(\boldsymbol{\theta})), ESS(\pi(\mu)), ESS(\pi(\beta))$ for the tolerable dose in a phase I trial.

	$\pi(\boldsymbol{\theta})$	$\pi(\mu)$	$\pi(\beta)$
$\sigma_\mu^2 = \sigma_\beta^2 = 0.5^2$	37.00	22.73	98.11
$\sigma_\mu^2 = \sigma_\beta^2 = 1^2$	10.00	5.75	25.56
$\sigma_\mu^2 = \sigma_\beta^2 = 2^2$	3.00	1.37	6.53
$\sigma_\mu^2 = \sigma_\beta^2 = 3^2$	2.00	1.03	3.06
$\sigma_\mu^2 = \sigma_\beta^2 = 5^2$	1.00	1.00	1.38

Table 2.5: Effective sample sizes $ESS(\varphi(\boldsymbol{\theta})), ESS(\varphi(\mu)),\ ESS(\varphi(\beta))$ for the MDD priors $\varphi(\mu) = \psi\mathcal{N}(\tilde{\mu}_\mu, c\tilde{\sigma}_\mu^2) + (1-\psi)\mathcal{N}(\tilde{\mu}_\mu, \tilde{\sigma}_\mu^2),\ \varphi(\beta) = \psi\mathcal{N}(\tilde{\mu}_\beta, c\tilde{\sigma}_\beta^2) + (1-\psi)\mathcal{N}(\tilde{\mu}_\beta, \tilde{\sigma}_\beta^2)$ according to different values of the mixture weight ψ, $c = 10000$.

	$\psi = 0.2$			$\psi = 0.5$			$\psi = 0.8$		
	$\varphi(\boldsymbol{\theta})$	$\varphi(\mu)$	$\varphi(\beta)$	$\varphi(\boldsymbol{\theta})$	$\varphi(\mu)$	$\varphi(\beta)$	$\varphi(\boldsymbol{\theta})$	$\varphi(\mu)$	$\varphi(\beta)$
$\sigma_\mu^2 = \sigma_\beta^2 = 0.5^2$	37.00	22.70	98.06	37.00	22.62	97.90	37.00	22.30	97.18
$\sigma_\mu^2 = \sigma_\beta^2 = 1^2$	10.00	5.73	25.50	10.00	5.69	25.31	9.00	5.52	24.58
$\sigma_\mu^2 = \sigma_\beta^2 = 2^2$	3.00	1.37	6.49	3.00	1.37	6.42	3.00	1.31	6.06
$\sigma_\mu^2 = \sigma_\beta^2 = 3^2$	2.00	1.03	3.03	2.00	1.03	3.01	2.00	1.03	2.68
$\sigma_\mu^2 = \sigma_\beta^2 = 5^2$	1.00	1.00	1.38	1.00	1.00	1.37	1.00	1.00	1.26

Table 2.6: Effective sample sizes $ESS(\varphi^j(\boldsymbol{\theta})), ESS(\varphi^j(\mu)),\ ESS(\varphi^j(\beta))$ for the MDD priors $\varphi^j(\mu) = \psi + (1-\psi)\pi_\mu,\ \varphi^j(\beta) = \psi + (1-\psi)\pi_\beta$ according to different values of the mixture weight ψ.

	$\psi = 0.2$			$\psi = 0.5$			$\psi = 0.8$		
	$\varphi^j(\boldsymbol{\theta})$	$\varphi^j(\mu)$	$\varphi^j(\beta)$	$\varphi^j(\boldsymbol{\theta})$	$\varphi^j(\mu)$	$\varphi^j(\beta)$	$\varphi^j(\boldsymbol{\theta})$	$\varphi^j(\mu)$	$\varphi^j(\beta)$
$\sigma_\mu^2 = \sigma_\beta^2 = 0.5^2$	32.00	19.71	87.65	23.00	14.03	62.43	11.00	6.55	29.06
$\sigma_\mu^2 = \sigma_\beta^2 = 1^2$	6.00	3.58	15.78	3.00	1.68	7.42	1.00	1.03	2.48
$\sigma_\mu^2 = \sigma_\beta^2 = 2^2$	1.00	1.00	1.99	1.00	1.00	1.14	1.00	1.00	1.03
$\sigma_\mu^2 = \sigma_\beta^2 = 3^2$	1.00	1.00	1.10	1.00	1.00	1.03	1.00	1.00	1.03
$\sigma_\mu^2 = \sigma_\beta^2 = 3^2$	1.00	1.00	1.03	1.00	1.00	1.03	1.00	1.00	1.03

mixture prior. We may conclude that the best way for reducing an extra amount of information is combining an informative prior with an improper or —when possible— with a Jeffreys prior, as suggested in Section 2.6.1 as well.

2.7 Discussion and further work

In this chapter a new class of data-dependent prior distributions is proposed. This class consists of a two-component mixture of a baseline (flat) prior π_b and an informative prior π, weighted through a discrepancy measure between the informative prior and the data generating model: π_b is favored if π appears to be far from the data at hand (natural procedure) or from an augmented sample size (resampling procedure). This mixture formulation is a good proposal in terms of robustness and is designed for avoiding prior-data conflict in presence of small sample sizes. First evidences from simulation studies suggest also good performances for reducing the mean squared errors and for improving the frequentist coverage.

The notion of effective sample size and, more generally, the amount of information provided by a prior distribution are central in our work. We proved that the MDD prior always provides a lower information than an informative prior within conjugate models. Furthermore, we suggest using the effective sample size as an effective threshold for choosing between one among the two described procedures for the MDD class.

Furthermore, different solutions for eliciting the baseline prior π_b are explored: flat prior belonging to the same family of π, Jeffreys prior, improper prior. As is just partially intuitive, different strategies for the noninformative prior yield different extents of information for the MDD prior.

Further work should be done in many directions. We should in fact explore more complex models, whose a brief sketch is only outlined in this chapter. Performing a proper sensitivity test for the selected priors π_b, π is also a task of future interest. Finally, we strongly believe that extending the proposed methodology for regression models in terms of Bayesian variable selection is one crucial point in future research.

Chapter 3

Hierarchical Bayesian models for individual performance in soccer

3.1 Introduction

Compared to the volumes statisticians (professional and amateur) have written about baseball, and to the growing statistical literature on sports like basketball and American football, there has been relatively little published by statisticians about soccer. A few highlights from the limited statistical literature include: Baio and Blangiardo (2010), who use a Bayesian hierarchical model to predict the outcome of individual matches throughout a season in the top Italian league, Serie A; Karlis and Ntzoufras (2000), in which the authors take a frequentist approach to estimating parameters related to the number of goals scored by specific teams; Dixon and Coles (1997), who use a familiar Poisson model for the number of goals between two teams and also consider suitable betting strategies based on their model; and Karlis and Ntzoufras (2009), which is a Bayesian model for the goal differential between two teams using a Skellam (Poisson difference) distribution.

In most of the published statistical research on soccer, including the papers mentioned above, the authors do not focus on modeling the performance of individual players over the course of a season but rather on some aspect of the global result of a match between opposing teams (e.g., goal differential), or on predicting the order of the league table at the end of a season. Relative to sports like baseball (Albert, 1992) or American football (Becker and Sun, 2016), the performance of individual soccer players is noisy and hard to predict. The dimensions of the soccer field combined

with the number of players, the difficulty of controlling the ball without the use of hands, and many other factors all contribute to the predictive challenge.

More primitive than the question of how to model player performance is how to *measure* it. Although there is no consensus on how to quantify individual performance in any sport, there has been less development in this area for soccer than for other major sports. And only after measurement is defined does modeling make sense. The oldest procedure for measuring the individual performance in the so called *goal-based* team sports —hockey, soccer and basketball, among others— is the so called *plus/minus* approach (see Thomas et al. (2013) for some references and recent improvements). This method tracks the number of goals scored, both for and against, for each player on the ice (hockey) or on the field (soccer); in such a way, the more/less goals are scored/conceded when a player is on the ice/field, the better is rated that player. But measuring the individual abilities of some players who share the ice/field for much of their time may be hard. Moreover, the rarity of the goals is often another problematic issue.

Although we are interested in modeling the overall performance of individual players, we are not yet convinced that there is an available holistic measure of individual performance worth modeling. In fact, even as the amount and variety of publicly available soccer data grows —particularly data at the individual player/match level— the interpretability and predictive relevance of that data will remain a question. However, we do suspect that the *fantasy* soccer framework (Bonomo et al., 2014; Lomax, 2006) may provide a valid measure of individual performance: in such a way, a prediction task for individual performance/ratings could be well posed and also serve as an example of a possible approach to use in the future when better measures of individual performance in soccer matches become available. The outcome of interest is the fantasy rating of each player in Italy's top league, Serie A, for each match of the 2015–2016 season. We strongly believe that these fantasy ratings may be seen as a proxy for the quality of a player's performance; in fact, they combine a subjective evaluation with an objective factor accounting for specific in-game events. Moreover, given the popularity of such fantasy games, these ratings are themselves an interesting variable to model. In this chapter we present and critique several Bayesian hierarchical models (Gelman et al., 2013; Gelman and Hill, 2006) designed to predict the results of the Italian fantasy game Fantacalcio. We use RStan (Stan Development Team, 2016*a*), the R (R Core Team, 2016) interface to the Stan C++ library (Stan

Development Team, 2016*b*), to sample from the posterior distributions via Markov chain Monte Carlo. As far as we can tell from reviewing the literature, there have been no published attempts to use a hierarchical Bayesian framework to address the challenges of modeling this kind of data.

Our central goals are to explore what can be accomplished with a very simple dataset comprising only a few variables (that are freely and easily available), and to focus on a small number of interesting modeling and prediction questions that arise (for instance, those due to the missingness of certain values). For this reason we also gloss over many issues that we believe should be of interest in subsequent research, for instance variable selection, additional temporal correlation structures, and the possibility of constructing more informative prior distributions. Although we restrict our focus to Fantacalcio, the process of developing these models and comparing them on predictive performance does not entirely depend on the idiosyncrasies of this particular fantasy system and is applicable more broadly.

The rest of the chapter is structured as follows. In Section 3.2 we briefly introduce the Italian fantasy soccer game Fantacalcio. We then describe our dataset in Section 4.4. The models we fit to the data are presented in Section 3.4 with results in Section 3.5. In Section 3.6 we carry out a variety of posterior predictive checks as well as out-of-sample prediction tasks. Section 4.6 concludes.

3.2 Overview of the game

Fantasy sports games typically involve roster selection and match-by-match challenges against other participants with the results determined by the collective performance of the players on the fantasy rosters. In Italy, fantasy soccer was popularized by the brand Fantacalcio edited by Riccardo Albini in the 1990s (`http://www.fantacalcio.it`). At the beginning of the season, Fantacalcio managers are allocated a limited amount of virtual money with which to buy the players that will comprise their roster. We will refer to the athletes, the soccer players, as *players*, and use *manager* for the Fantacalcio participants. Each player in the Italian Serie A league has an associated price determined by various factors including past performance and forecasts for the upcoming season.

After every match in Serie A, the prominent Italian sports periodicals assign each player a rating, a so-called *raw score*, on a scale from one to ten. In practice there

Table 3.1: Bonus/Malus points in Fantacalcio. The events marked with a * symbol are only applicable to goalkeepers.

Event	Points
Goal	+3
Assist	+1
Penality saved*	+3
Yellow card	−0.5
Red Card	−1
Goal conceded*	−1
Own Goal	−2
Missed penality	−3

is not much variability in these scores; they typically range from four to eight, with the majority between five and seven. These raw scores are very general and largely subjective performance ratings that do not account for significant individual events (goals, assists, yellow and red cards, etc.) in a consistent way.

As a means of systematically including specific in-game events in the ratings, Fantacalcio provides the so-called *point scoring* system. Points are added or deducted from a player's initial raw score for specific positive or negative events during the match. The point scores are more variable than the raw scores, especially across positions (e.g., when comparing defending and attacking players). Goalkeepers suffer the most from the point scoring system, as they are deducted a point for every goal conceded. On the other extreme, forwards (attacking players) typically receive the highest point scores because every goal scored is worth three points.

For player i in match t the total rating y_{it} is

$$y_{it} = \mathsf{R}_{it} + \mathsf{P}_{it}, \tag{3.1}$$

where R is the raw score and P is the point score. Table 3.1 lists the game features that contribute to a player's point score P_{it} for a given match. It is worth noting that negative ratings are also possible, although not very common. For instance, a goalkeeper with a raw score of three who also allows four goals would have a rating $y_{it} = -1$.

Since it is very rare for a player to participate in all matches, some y_{it} are *missing*,

and this may be due to different reasons. First, player i's rating for match t will be missing if the player does not play in the match because of injury, disqualification, coach's decision, or some other reason. In addition, this can occur when a player does not participate in the match for long enough for their impact to be judged by those tasked with assigning the subjective raw score ($\mathsf{R}_{it} = 0$) or for the player to accumulate or lose any objective points ($\mathsf{P}_{it} = 0$).

Modeling the missingness is one of the main concerns of this chapter. We return to this issue later in Sections 3.4.2 (mixture models) and 3.4.3 (missing data models) when we confront the challenge it poses for our modeling and prediction tasks and consider methods for modeling the missingness that naturally arises in our dataset.

3.3 Data

All data for this chapter are from the 2015–2016 season of the Italian Serie A and were collected from the Italian publication La Gazzetta dello Sport (`http://www.gazzetta.it`). We use all of the ratings for every player satisfying the following two criteria:

- The player participated in at least a third of matches during the *andata* (the first half of the season). This amounts to dropping players who played in fewer than seven matches in the first half.

- The player participated in the *final* match of the *andata*.

This results in a dataset containing ratings for 237 players (18 goalkeepers, 90 defenders, 78 midfielders, and 51 forwards). Figure 3.1 displays the distributions of average ratings by position, while Figure 3.2 shows the bivariate relationship between average rating and the initial standardized price for each player.

Although the full season comprises 38 matches for each team, as alluded to in Section 3.2 rarely does a player participate in all matches. For the 237 players in our data that meet the two criteria above, the mean number of matches played is 27.5 with a standard deviation of about 7, and 75% of these players missed at least 5 matches.

Note that the professional European soccer leagues may allow for a players' transfer market occurring approximately at the midpoint of the season. According to this

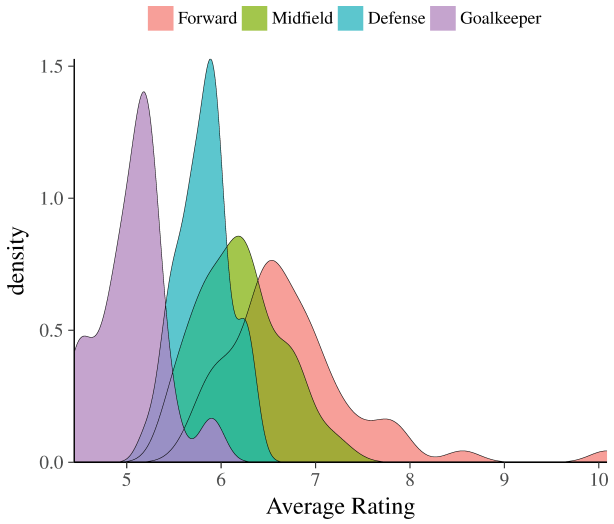


Figure 3.1: The distributions of average ratings by position.

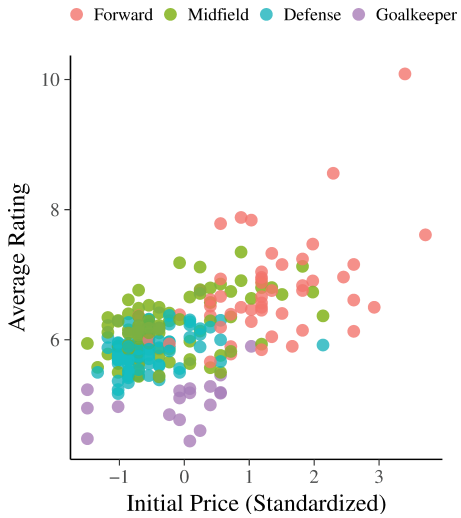


Figure 3.2: The distributions of average ratings versus initial standardized price.

opportunity, some players may move to a new team belonging either to the Italian Serie A or to another league. Although the transfer market could be appealing in terms of players' performances and also in terms of Fantacalcio ratings, only a few players of the current dataset changed team; thus, we do not need to include this issue in the model and we assign to each player the membership team at the beginning of the season.

Notation for observed data

There are $N = 237$ players and $T = 38$ matches in the dataset. When fitting our models we use only the $T_1 = 19$ matches from the first half of the 2015–2016 Serie A season. The remaining $T_2 = 19$ matches are used later for predictive checks. For match $t \in \{1, \ldots, T\}$, let y_{ijkt} denote the value of the total rating for player $i \in \{1, \ldots, N\}$, with position (role on the team) $j \in \{1, \ldots, J\}$, on a team in team-cluster $k \in \{1, \ldots, K\}$. To ease the notational burden, throughout the rest of the chapter the subscripts j and k will often be implicit and we will use y_{it} in place of y_{ijkt}.

The players are grouped into $J = 4$ positions (forward, midfielder, defender, goalkeeper) and $K = 5$ team clusters. The five clusters (Table 3.2) were determined using the official Serie A rankings at the midpoint of the season. The purpose of the team clustering is both to use a grouping structure that has some practical meaning in this context and also to reduce the computational burden somewhat by including cluster-specific parameters rather than team-specific parameters. We experimented with team specific parameters but found that it leads to models that are slower to fit but that yield similar inferences.

There are only two other variables in our limited dataset. We let $h_{it} = 1$ if player i's team plays match t at its home stadium and $h_{it} = 0$ if the match is played at the opponent's stadium. And we use q_i to denote the initial standardized price for player i. These values are assigned by experts and journalists at the beginning of the season based on their personal judgement and then updated throughout the season to reflect each player's performance (`http://www.gazzetta.it/calcio/fantanews/statistiche/serie-a-2015-16/`).

Table 3.2: The $K = 5$ team clusters, from weakest to strongest. Group 5 is headlined by Juventus, the top performing team in Serie A for the past several seasons.

Cluster	Teams
1	Palermo, Frosinone, Carpi, Verona
2	Genoa, Sampdoria, Empoli, Udinese
3	Bologna, Chievo, Atalanta, Torino
4	Milan, Fiorentina, Lazio, Sassuolo
5	Juventus, Roma, Inter, Napoli

3.4 Models

Notation for model parameters

The notation we use for model parameters is similar to the convention adopted by Gelman and Hill (2006) for multilevel models. According to this, the index variables $j[i],\ k[i]$ code group membership. For instance, if $j[1] = 4$, then the first unit in the data ($i = 1$) belongs to position group 4. If $k[1] = 3$, then the first unit belongs to team-cluster 3.

We use α_i for individual random effects corresponding to each player $i = 1, ..., N$. The parameters γ_k and $\beta_{k,t}$ represent, respectively, the team-cluster effect and the team-cluster effect of the team opposing in match t, with $k = 1, ..., K$. As already mentioned, in our simplified framework we set the number of team-clusters $K = 5$. We denote by ρ_j the position-specific parameters, with $j = 1, ..., J$ and $J = 4$. The standardized prices are multiplied by a slope δ_j, which is allowed to vary across the J positions. Because we are interested in detecting trends in player ratings, we also incorporate the average rating up to the game $t - 1$, $\bar{y}_{i,t-1}$, which is multiplied by a factor $\lambda_{j[i]}$ estimated from the data. For the mixture model in Section 3.4.2, the same average rating $\bar{y}_{i,t-1}$ is also multiplied by a coefficient $\zeta_{j[i]}$ in order to model the probability of participating in the match t. We anticipate that in their posteriors λ and ζ (here denoted as vectors) will be meaningfully different from zero. Since we work in a Bayesian framework, all parameters will be assigned prior distributions, which in turn may depend on hyperparameters that are either fixed or themselves estimated from the data.

3.4.1 Hierarchical autoregressive model (HAr)

As above, let y_{it} (with indices j and k implied) denote the total rating (3.1) for player i in match t. For our first model, we code all the missing ratings y as zeros. This makes sense if we are (and, in part, we are) interested in the annual *cumulative* rating of a given player, or of a given subset of players (this is investigated graphically later in Section 3.6). Or, for instance, experts and scouts may be interested in estimating the number of goals that will be scored by Roma's forwards. Since the number of goals heavily depends on the number of games played, it makes sense to assign a value of zero for any missed matches (unobserved player ratings) as they should not contribute to the total number of goals scored. Later, in Section 3.4.3, we will take a different approach in which missing values are actually treated as unobserved and we specify a full joint probability model for both the observed and unobserved ratings.

We begin with a standard hierarchical autoregressive model

$$y_{it} \sim \mathsf{Normal}\left(\eta_{it},\, \sigma_y\right), \tag{3.2}$$

where η_{it} is the linear predictor

$$\eta_{it} = \alpha_0 + \alpha_i + \beta_{k[i],t} + \gamma_{k[i]} + \rho_{j[i]} + \delta_{j[i]} q_i + \lambda_{j[i]} \bar{y}_{i,t-1} + \theta h_{it}, \tag{3.3}$$

α_0 is the intercept, and σ_y is the standard deviation of the error in predicting the outcome. Note that the term autoregressive is used here for indicating the inclusion of the average rating up to the game $t-1$ in the model. As we are fitting our models using Stan (Stan Development Team, 2016*b*), we follow its convention of parameterizing normal distributions in terms of standard deviation rather than the precision or variance.

The individual-level, position-level, and team-cluster-level parameters are given hierarchical normal priors,

$$\begin{aligned}
\alpha_i &\sim \mathsf{Normal}(0,\, \sigma_\alpha), \quad i = 1, \ldots, N \\
\gamma_k &\sim \mathsf{Normal}(0,\, \sigma_\gamma), \quad k = 1, \ldots, K \\
\beta_k &\sim \mathsf{Normal}(0,\, \sigma_\beta), \quad k = 1, \ldots, K \\
\rho_j &\sim \mathsf{Normal}(0,\, \sigma_\rho), \quad j = 1, \ldots, J
\end{aligned} \tag{3.4}$$

with weakly informative prior distributions for the remaining parameters and hyper-

parameters,

$$\begin{aligned}
\alpha_0 &\sim \mathsf{Normal}(0, 5) \\
\theta &\sim \mathsf{Normal}(0, 5) \\
\delta_j &\overset{iid}{\sim} \mathsf{Normal}(0, 5), \quad j = 1, \ldots, J \\
\lambda_j &\overset{iid}{\sim} \mathsf{Normal}(0, 1), \quad j = 1, \ldots, J \\
(\sigma_\theta, \sigma_\alpha, \sigma_\gamma, \sigma_\beta, \sigma_\rho) &\overset{iid}{\sim} \mathsf{Normal}^+(0, 2.5), \\
\sigma_y &\sim \mathsf{Cauchy}^+(0, 5),
\end{aligned}$$

where Normal^+ and Cauchy^+ denote the half-Normal and half-Cauchy distributions. On the choice of these priors for the scale parameters, see Gelman et al. (2006). Note that centering the individual-level, the team-cluster-level, and the position-level parameters in (3.4) at $\mu_\alpha, \mu_\gamma, \mu_\beta$, and μ_ρ would make the model nonidentifiable, because a constant could be added to each of these hyperparameters without changing the predictions of the model. This is the motivation for centering these prior distributions at zero.

3.4.2 Mixture model (MIX)

Even if we found that some players have a tendency to be ejected from matches due to red cards, for instance, or tend to suffer injuries at a high rate, it would still be very challenging to arrive at sufficiently informative probability distributions for these events. Even with detailed player histories over many seasons, it would be hard to predict the number of missing matches in the current season. Nevertheless, we can try to incorporate the *missingness* behavior intrinsic to the game into our models. Assuming that it is very rare for a player to play in every match during a season, we can try to model the overall propensity for missingness. A general way of doing this entails introducing a latent variable, which we denote V_{it} and define as

$$V_{it} = \begin{cases} 1, & \text{if player } i \text{ participates in match } t, \\ 0, & \text{otherwise.} \end{cases}$$

If for each player i we let $\pi_{it} = Pr(V_{it} = 1)$, then we can specify a mixture of a Gaussian distribution and a point mass at 0 (Gottardo and Raftery, 2008)

$$p\left(y_{it} \mid \eta_{it}, \sigma_y\right) = \pi_{it}\, \mathsf{Normal}\left(y_{it} \mid \eta_{it}, \sigma_y\right) + (1 - \pi_{it})\, \delta_0, \tag{3.5}$$

where δ_0 is the Dirac mass at zero and η_{it} is the same linear predictor as before. The probability π_{it} is modeled using a logit regression,

$$\pi_{it} = \text{logit}^{-1}\left(p_0 + \zeta_{j[i]}\bar{y}_{i,t-1}\right), \tag{3.6}$$

which takes into account predictors that are likely to correlate with player participation. The variable $\bar{y}_{i,t-1}$ is the average rating for player i up to match $t-1$, and p_0 is an intercept for the logit model. How to model π_{it} could be the subject of entire papers, but better models would likely require variables beyond what we have in our dataset (e.g., injury histories). Our simplistic model will suffice for our purposes of exploring what we can do with only this dataset. For the new parameters introduced in (3.6) we use the weakly informative priors

$$\begin{aligned} p_0 &\sim \mathsf{Normal}(0, 2.5), \\ \zeta_j &\overset{iid}{\sim} \mathsf{Normal}(0, 1), \quad j = 1, \ldots, J. \end{aligned}$$

The models for the group-level parameters and the hyperpriors are the same as in 3.4.1. The Stan code for the MIX model is in Appendix B.

3.4.3 Refitting the HAr model accounting for missing data

As we have already mentioned, it is difficult to deal with the issue of missing data in such a way as to yield a reasonable estimate of the cumulative ratings over a season. The MIX model may be seen as a natural attempt at modeling the missingness, while, to ease the problem, in the initial HAr model missing values were treated as zeros and not modeled. We have already speculated about the legitimacy of this approach, but we are only partially interested in the cumulative rating over the entire season and are also interested in assessing the predictive accuracy of our models game by game. That is, we also want to answer the question: how will a player perform if they play in the match? One way to do this is by treating each missing player rating as an unknown parameter rather than somewhat arbitrarily fixing it at zero. As broadly outlined in Gelman et al. (2013), Bayesian inference draws no distinction between missing data and parameters, so the target distribution is the joint posterior distribution of the missing data and other model parameters conditional on the observed data.

Let y represent the complete data we could have observed in the absence of missing values; we split our data matrix into two subsets, $y = (y^{obs}, y^{mis})$, where y^{obs} denotes

the observed values and y^{mis} denotes the missing values. We also define I to be the inclusion matrix such that $I_{it} = 1$ if y_{it} is observed and $I_{it} = 0$ if y_{it} is missing. In this setup, y^{obs} are data and y^{mis} are parameters. For convenience, we specify our new augmented model as

$$y_{it} = \begin{cases} y_{it}^{obs}, \text{ if } I_{it} = 1 \\ \xi_{it}, \quad \text{if } I_{it} = 0, \end{cases} \qquad i = 1, ..., N, \ t = 1, ..., T \tag{3.7}$$

where y_{it}^{obs} is an observed rating for player i in match t and each ξ_{it} is a parameter. The toy Stan program reported below shows one way of writing a joint model for the observed data and the missing data.

```
data {
  int N_obs;
  int N_mis;
  real y_obs[N_obs]; // data
}
parameters {
  real eta;
  real<lower=0> sigma;
  real xi[N_mis]; // parameters
}
model {
  vector[N_obs + N_mis] y = append_row(y_obs, xi);
  target += normal_lpdf(y | eta, sigma);  // log density
}
```

The variable `y_obs` represents data and `xi` are parameters. For brevity, in this toy example we leave the default flat priors on `eta` and `sigma`, omit predictor variables, and assume the data is a vector rather than an $N \times T$ matrix. The same idea is then incorporated into the HAr model from 3.4.1. We refer to this modified model as the HAr-mis model.

3.5 Results

3.5.1 Estimates

We fit the models via Markov chain Monte Carlo using RStan Stan Development Team (2016*a*), the R interface to the Stan C++ library Stan Development Team (2016*b*), and monitored convergence as recommended in Stan Development Team (2016*c*). Figure 3.3 shows the parameter estimates.

For all models, the β, γ and δ vectors are almost all shrunk towards their grand mean 0, with little variability. For the position-specific vector ρ, the HAr-mis and MIX models estimate slightly positive values (approximately 0.5) for midfielders (ρ_2) and defenders (ρ_3), while for the HAr model these parameters are shrunk close to zero. The goalkeeper effect (ρ_4) is slightly positive for the HAr model but clearly negative for the HAr-mis and MIX models. For all models these position-level parameters have larger posterior uncertainties than the other parameters. All three models recognize a slight advantage due to playing at home ($\theta > 0$). Also in Figure 3.3 we see that for the λ's, the coefficients on the lagged average ratings, the estimates obtained from the HAr model are much larger than those obtained under the HAr-mis and MIX models, which again give nearly identical estimates. Since for every match day t these coefficients are multiplied by the lagged average rating $\bar{y}_{i,t-1}$, we believe that the larger λ estimates from the HAr model are the result of coding the missing values as zeros.

For the MIX model only there are also additional parameters $\zeta_1, \ldots, \zeta_4$ that scale the lagged average rating in the logit model (3.6). These parameters are all positive —which corresponds to the intuition that higher ratings are associated with higher probabilities of participating in the next match— and they also exhibit non-negligible variation across positions (for goalkeepers, ζ_4, the estimated association is strongest).

3.5.2 Inference through fake data simulation

In this section we give an example of a more interesting comparison focusing on simulating hypothetical players rather than comparing parameter estimates.

Comparing parameter estimates across models is standard practice, but we are more interested in the implications of the parameters for the outcome variable rather than the parameters themselves. For our purposes it should be more informative to

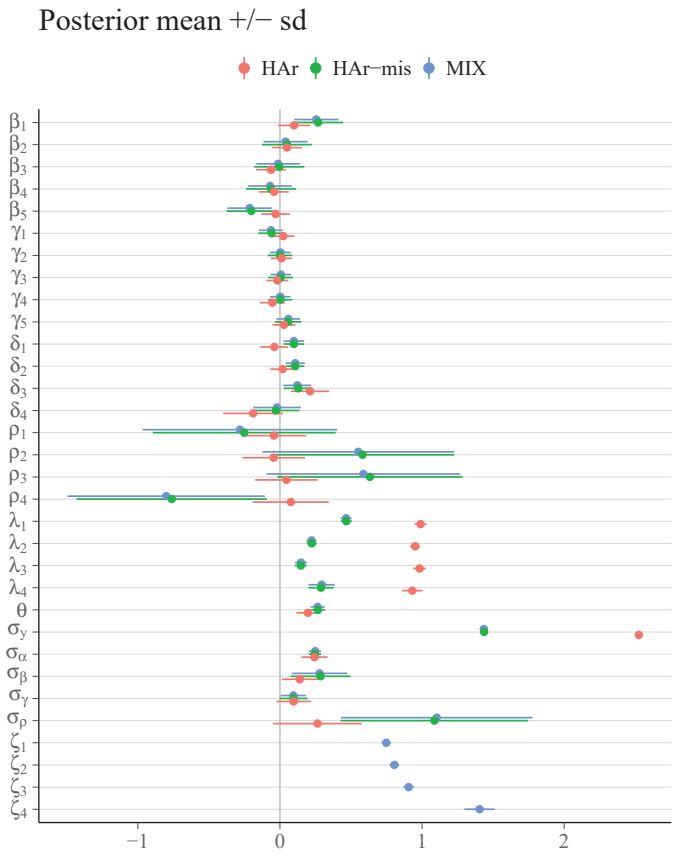


Figure 3.3: Posterior means ± standard deviations for the model parameters common to the HAr, HAr-mis, and MIX models. $\beta_{k,t}$ and γ_k are the parameters for the opposing team-cluster in match t and the player's team-cluster ($k = 1$ the weakest, $k = 5$ the strongest). The parameters δ_j (coefficients on initial price), λ_j (coefficients of the lagged average rating) and ρ_j all vary by position (1 =Forward, 2 =Midfield, 3 =Defender, 4 =Goalkeeper). θ is the coefficient for the home/away predictor. σ_y is the individual-level standard deviation and the other σ's are the hierarchical standard deviation parameters. For the MIX model, the ζ's are the coefficients on the lagged average rating from (3.6).

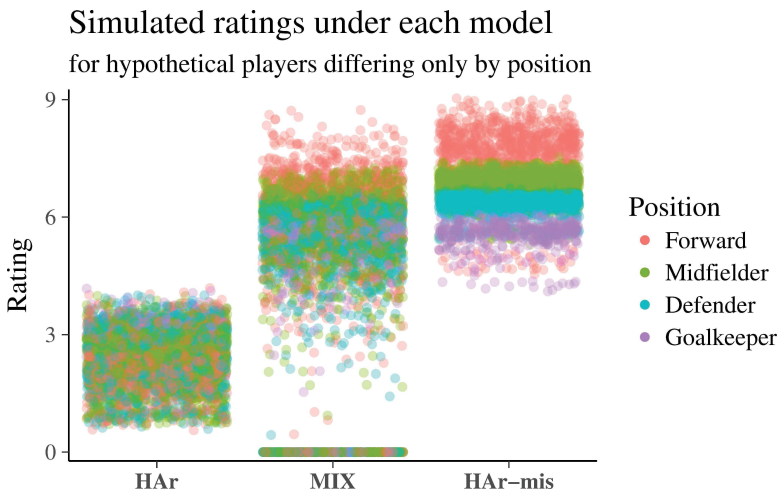


Figure 3.4: Predicted ratings of hypothetical players differing only in their position. Predictions from each of the three models are shown for 19 matches for each of 237 players (the size of our dataset), all playing at home ($h_{it} = 1$), all playing on a team in cluster $k = 3$ against an opponent in cluster $k = 3$, with standardized average position price $\bar{q}_{j[i]},\ j = 1, \ldots, J,\ i = 1, \ldots, N$.

simulate outcomes under each of the models for players differing only in their position. We can then directly compare the variability in the ratings for these hypothetical players. Note that comparing predictions rather than parameter estimates would be even more essential if we were fitting logistic regression models (or other GLMs) rather than Gaussian linear models.

We predict ratings for several players at different positions on the field and with the average position price in virtual money, all on the same cluster team, all playing against the same cluster team, and all playing at their home stadium. Figure 3.4 shows the predicted ratings from each of the models for 19 new matches and $N = 237$ (the size of our dataset) hypothetical players. For the HAr model, the position variability appears to be very small when compared with the variability in the predictions from the the HAr-mis model and, a bit less, the MIX model. Moreover, the predicted values for the HAr model are shrunk together and turn out to be too much low: this failure of the HAr model can be explained by the fact that it treats missed matches as zeros and, then, it will tend to favor players with fewer zeros.

Conversely, the simulations from the HAr-mis model are more clearly separated

into strata corresponding to the different positions and the hierarchy of positions is correct: forwards tend to register the highest simulated ratings, then midfielders, defenders, and goalkeepers. The MIX model is less able to clearly separate the positions in the predictions but it does get the correct ordering on average. As expected, it also predicts a non-negligible number of zeros (missing values).

Here we only show the comparison made by varying a player's position, but analogous visualizations can be made to explore the effect of changing other variables.

3.6 Posterior predictive checks and predictions

Now that we have estimated all of the models, we turn our attention to evaluating the fit of the models to the observed data as well as the predictive performance of the models on hold-out data. We use the 19 match days comprising the first half of the Serie A season —the *andata*— as training data, and for every player in the dataset we make in-sample predictions for those 19 matches as well as out-of-sample predictions for the remaining 19 matches —the *ritorno*. As usual in a Bayesian framework, the prediction for a new dataset may be directly performed via the posterior predictive distribution for our unknown set of observable values. Following the notation of Gelman et al. (2013), we denote by $\tilde{y}$ a generic unknown observable. Its distribution conditional on the observed y is

$$p(\tilde{y}|y) = \int_{\Theta} p(\tilde{y}, \theta|y)\, \mathrm{d}\theta = \int_{\Theta} p(\tilde{y}|\theta)\, p(\theta|y)\, \mathrm{d}\theta, \tag{3.8}$$

where the independence of y and $\tilde{y}$ conditional on θ is assumed. We are also implicitly conditioning on the observed predictors. Sampling from this posterior predictive distribution will allow us to both assess the fit of the model to observed data and also make out-of-sample predictions that average over the posterior.

3.6.1 In-sample posterior predictive checks

To assess how well the models fit the training data, for each draw of the parameters from the posterior distribution we draw a dataset from the posterior predictive distribution of the outcome under each of the models. We should expect the in-sample predictive performance to be better than performance on out-of-sample prediction tasks (Gelman, Hwang and Vehtari, 2014; Vehtari et al., 2017). Figure 3.5 shows an

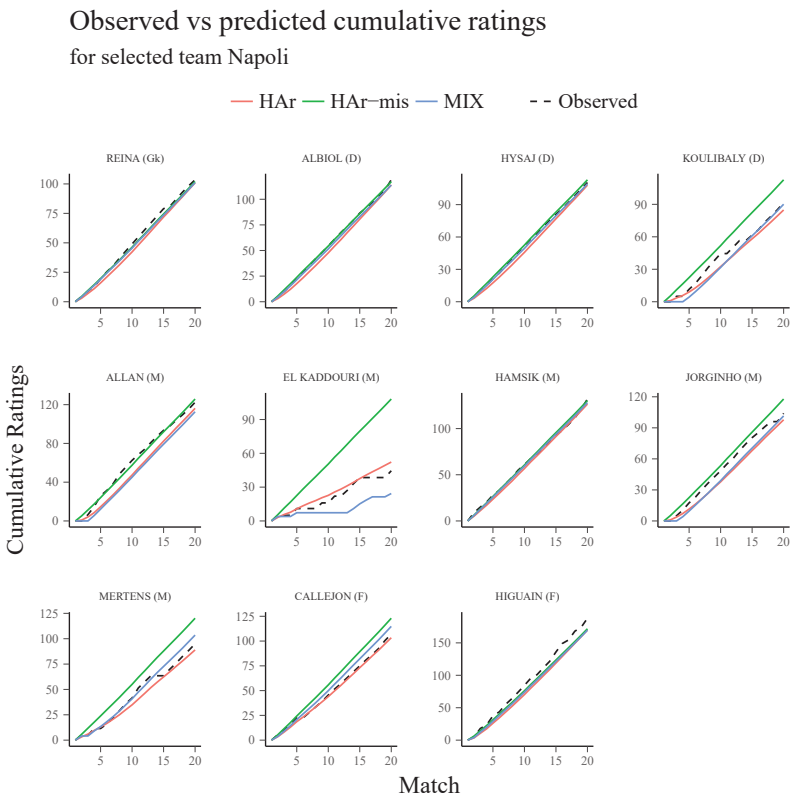


Figure 3.5: Observed vs. median predicted cumulative ratings for selected team Napoli during the first half of the 2015–2016 Serie A season.

example of a graphical posterior predictive check focusing on the *cumulative* ratings for each player over the matches in the training data. For illustration purposes, here we only show the results for one team, Napoli, but equivalent plots could be made analogously for all the other teams. The dashed black lines represent the observed values, while the red, green, blue lines represent predictions from the HAr, HAr-mis and MIX models, respectively.

For many of the players all of the models make reasonable predictions. However, for players with many missed matches the HAr and MIX models outperform the HAr-mis model (see the plot for El Kaddouri, for instance). The HAr-mis model will perform well on many of the predictive tasks, but it is not designed to predict in-sample cumulative ratings. The cumulative rating is very sensitive to the number of missing values, but for each missing value the HAr-mis will predict a plausible rating for if the player had played instead of a zero.

Figure 3.6 provides a different graphical check of the model fitting. Each row of

plots shows the distribution of a test statistic $T(y^{rep})$ computed over the replicated datasets y^{rep} generated from the posterior predictive distribution under each of the models. The vertical black lines indicate the value of $T(y)$, the statistic computed from the observed data. If we consider the distributions of these statistics —mean, median, minimum, maximum, and standard deviation— we immediately notice that the three models differ in their ability to replicate many of these features of the data. According to the mean, the median and standard deviation, the MIX model seems to be best at capturing these aspects of the training data.

In the fourth row we can see that the HAr model severely underestimates the minimum rating in the data, the HAr-mis model predicts a reasonable distribution of the minimum, and for the MIX model the distribution for the minimum is highly concentrated around 0, which is due to the nature of the model.

On the other hand, it is the MIX and the HAr-mis models that substantially underestimate the maximum rating, while the HAr model is able to predict plausible maximums when compared to the observed value. However, Figure 3.7 reveals that although the HAr-mis model fails to predict the overall maximum, it does predict reasonable maximum values for defenders and goalkeepers. Its failure to reproduce the maximums for the forwards and midfielders is explained by the rarity of their maximums (17 and 14, respectively) in the training data. Only one rating as high as 17 was observed in the first half of the season and there were only three ratings of at least 17 observed over the full season (about 1 in every 2000 observed ratings). To allow the HAr-mis model to predict such extreme values it may be possible to use a t-distribution instead of a Gaussian model, but for our purposes in this chapter the ability of a model to replicate these very rare ratings is not so essential.

3.6.2 In-sample and out-of-sample calibration

We are also interested in the calibration of the models on both the training and hold-out data. In Figs. 3.8, 3.9, and 3.10 we display the median predictions and 50% posterior predictive intervals under the HAr, MIX and HAr-mis models for our selected team Napoli, overlaying the observed data points. In a broader analysis we could plot and analyze these graphs for each team in Serie A under each of the models.

In a well-calibrated model we expect half of the observed values to lie outside

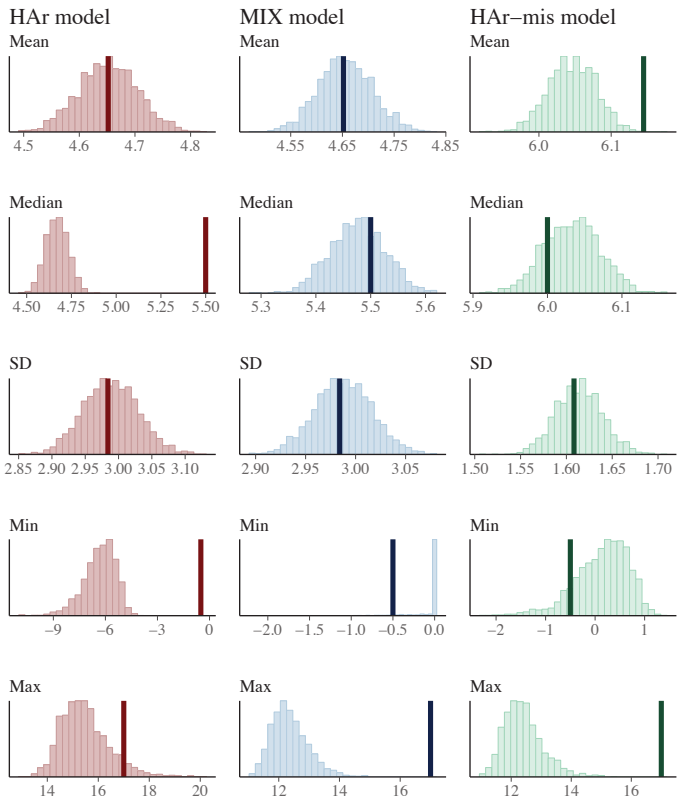


Figure 3.6: In-sample posterior predictive checks of test statistics for the HAr, MIX and HAr-mis models. For a particular test statistic T the plots show $T(y^{rep})$ (histogram) and $T(y)$ (thick vertical line). Each column corresponds to one of the three models, and each row to a different statistic T (mean, median, sd, minimum, maximum). We can see that the HAr model predicts much lower minimum values than the observed minimum. On the other hand, under the MIX model the distribution for the minimum is highly concentrated around zero.

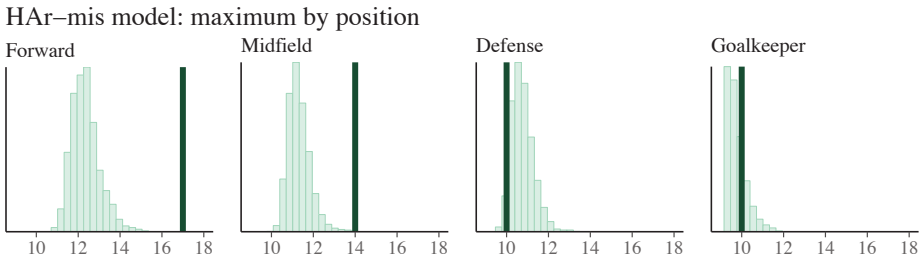


Figure 3.7: Posterior predictive check for $T(y)$=max over different positions for the HAr-mis model. The thick vertical line is the observed value.

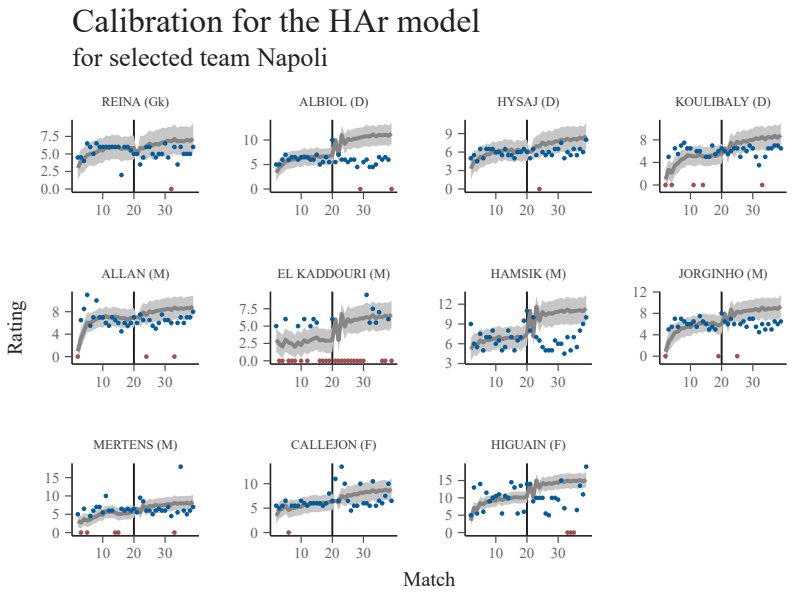


Figure 3.8: Calibration check for the HAr model for selected team Napoli. Blue points are observed values y^{obs}, red points are the zeros. The light gray ribbons represent 50% posterior predictive intervals and the overlaid dark gray lines are the median predictions. The vertical black lines separate the in-sample predictions from the out-of sample predictions.

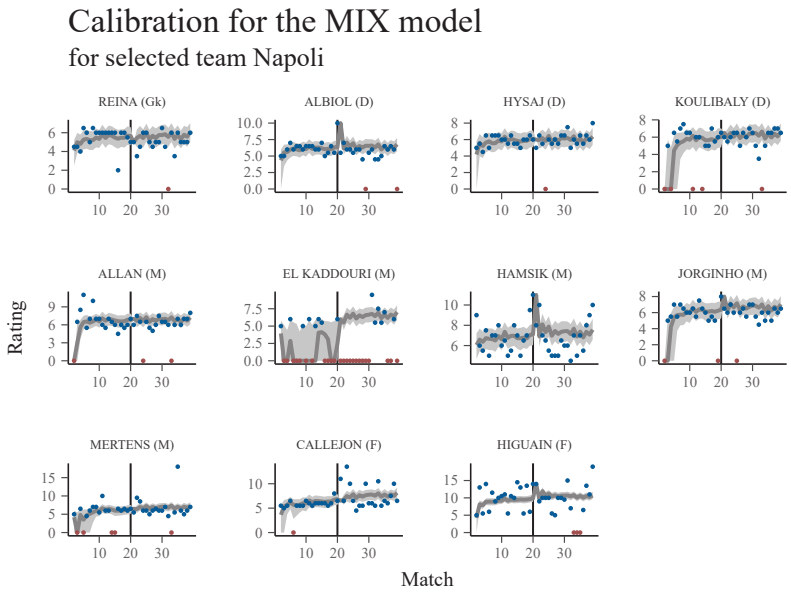


Figure 3.9: Calibration check for the MIX model for selected team Napoli. Blue points are observed values y^{obs}, red points are the missing values. The light gray ribbons represent 50% posterior predictive intervals and the overlaid dark gray lines are the median predictions. The vertical black lines separate the in-sample predictions from the out-of sample predictions.

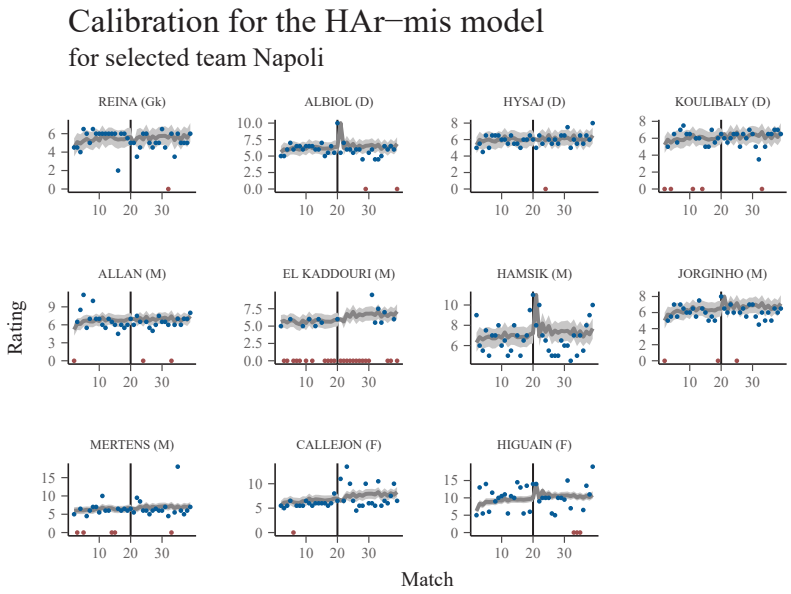


Figure 3.10: Calibration check for the HAr-mis model for selected team Napoli. Blue points are observed values y^{obs}, red points are the missing values. The light gray ribbons represent 50% posterior predictive intervals and the overlaid dark gray lines are the median predictions. The vertical black lines separate the in-sample predictions from the out-of sample predictions.

the corresponding 50% intervals. By this measure we can see in the plots that the HAr-mis and MIX model have decent but not excellent calibration, since for many of the players —particularly the goalkeeper and defenders— the 50% intervals cover more than 50% of the observed (blue) points. Conversely, for the volatile superstar Higuaín (an outlier even among forwards) only a many fewer points fall inside the intervals. Although the HAr model seems to be generally better calibrated, its main flaw consists in overestimating the defenders (and some other players) in the second part of the season, as already alluded in Section 3.5.2. Furthermore, the HAr model appears to identify an increasing trend in the ratings that is not actually supported by the data. As will be clear in Section 3.6.3, the out-of-sample predictions from the HAr model will in fact tend to be unreliable, while the MIX and the HAr-mis models tend to both better detect the best players *on average*.

3.6.3 Out-of-sample predictive checks

RMSE on hold-out data

For out-of-sample prediction we fit the models over the $T = 19$ matches in the first half of the season and then generate predictions for the $T^\star = 19$ matches in the second half of the season. For each player $i = 1, \ldots, N$ and for each posterior predictive simulation $s = 1, \ldots, S$ we compute the root mean square error (RMSE) over the matches $T+1, ..., T+T^\star$ in the held out data (corresponding to matches 20 through 38 of the season),

$$\text{RMSE}_i^{(s)} = \sqrt{\frac{\sum_{t=T+1}^{T+T^\star} \left(\tilde{y}_{it}^{(s)} - y_{it}\right)^2}{T^\star}}. \tag{3.9}$$

In the above equation $\tilde{y}_{it}^{(s)}$ is the sth simulation from the posterior predictive distribution of the predicted rating for player i at match t, and y_{it} is the corresponding observation. From this we obtain an RMSE *distribution* for each player.

Averaging over the simulations for each player and then averaging over players within positions we compute

$$\overline{\text{RMSE}}_j = \frac{\sum_{i=1}^{\#(i\in j)} S^{-1} \sum_{s=1}^{S} \text{RMSE}_i^{(s)}}{\#(i \in j)}, \quad j = 1, ..., J,$$

where $\#(i \in j)$ is the number of observations of position group j. Figure 3.11 shows these position-average RMSE values under each of the three models. The trend is the

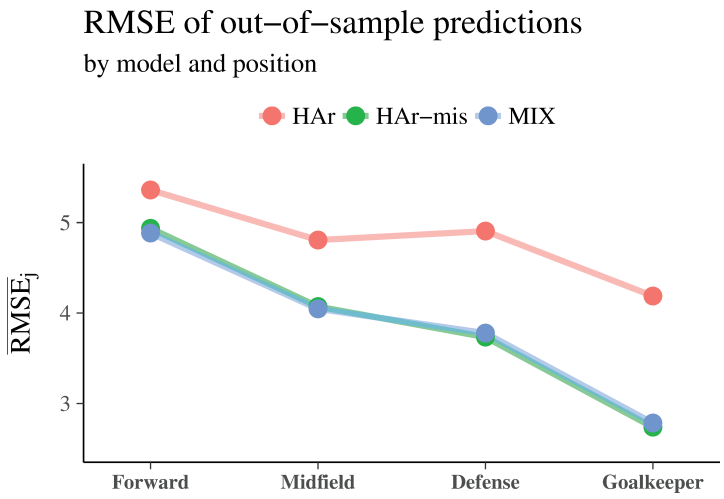


Figure 3.11: Average RMSE for the different positions for each model. The trend is the same across models: better predictions are obtained for goalkeepers, followed by defenders, midfielders, and finally forwards. The HAr-mis and MIX models register the lowest RMSE.

same across all models and suggests that our predictive ability is best for goalkeepers, followed by defenders, midfielders, and then forwards. Comparing across models, the missing data models (MIX and HAr-mis) perform better than the HAr model. This is further confirmation that modeling the missing values is important for predictive accuracy on hold-out data.

It is worth noting that in a dynamic framework, where the models could be updated between matches, the RMSEs would almost certainly be much lower than the RMSEs computed for the second half of the season in one batch. For instance, fitting our models at time t and projecting for match $t+1$, we could account for the disqualification of certain players, injuries, etc. If we know in advance that a player is disqualified for the next match we would have $y_{i,t+1} = \tilde{y}_{i,t+1} = 0$, and the corresponding RMSE would be zero.

Roster selection

Based on average predicted ratings for the held-out data from the second half of the 2015–2016 Serie A season, Figure 3.12 displays the best teams of eleven players that can be assembled from the available players according to each of the models

using their posterior medians. Also shown is the best team assembled using the observed ratings from the same set of matches. Here we assume that, in addition to a single goalkeeper, a team is comprised of four defenders, three midfielders, and three forwards. This is a common structure, although certainly many other formations are also used. As is evident at a first glance, the predictions obtained from the HAr model are quite inefficient. As we saw in the calibration plots in Figure 3.8, the HAr model tends to overestimate the player ratings, and we can see here that the projected ratings for the top players are quite far from their averages computed from the observed ratings in the hold-out data.

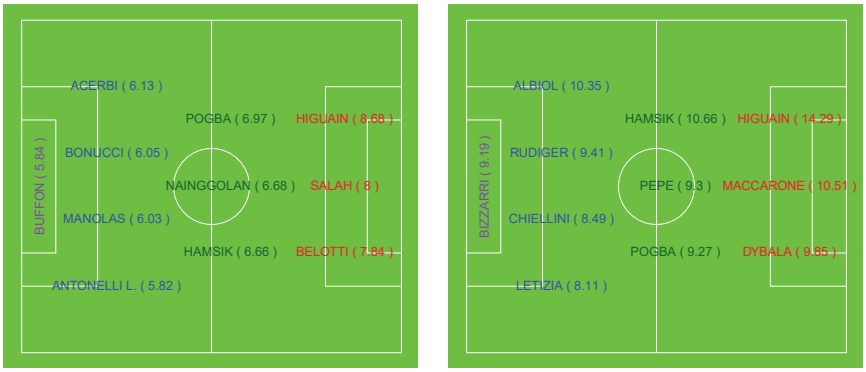
The rosters assembled based on the predictions from the HAr-mis and MIX models are identical except for the ordering of the players within the positions. Four of the eleven players (Acerbi, Pogba, Hamsik, Higuaín) from the team based on the actual ratings are included in the HAr-mis and MIX teams and, of the players that don't match, several are close. Dybala, the third best forward according to the models, is also rated highly (fifth best) according to the observed ratings. Rudiger, the second best defender according to both models, also has high observed mean rating (the eighth best among the 90 defenders). And Bonucci, one of the defenders included based on the observed ratings is also ranked highly by the HAr-mis model (ninth best) and MIX model (eleventh best).

Informally, this is further evidence that modeling the missingness allows us to obtain better out-of-sample predictions. Unlike the HAr model, the rosters selected by the MIX and the HAr-mis models appear to be quite competitive, which confirms the better performance we saw earlier in both the RMSE and the calibration comparisons.

3.7 Discussion

Although we are interested in our predictions for their own sake, our primary goal in this chapter has been of an exploratory rather than confirmatory nature. Given the lack of published research on modeling this kind of data within a Bayesian framework, we hope our proposed models and process will be useful to other researchers interested in working on individual-level predictions in the presence of noisy soccer data.

We proposed various hierarchical models for predicting player ratings and fit them according to two different scenarios: in the first scenario the HAr treated the missing values as zeros; in the second scenario the MIX and the extended HAr-mis models



(a) Observed (b) HAr

(c) MIX (d) HAr-mis

Figure 3.12: Best teams according to out-of-sample prediction of average player ratings for the HAr, MIX and HAr-mis model compared to the observed best team for the second part of the season. The averaged ratings are computed for those players who played at least 15 matches in the second half of the season.

allow for modeling the missing values themselves. We think the second framework is appealing in theory and we found in practice that the predictive predictive performance is good both in-sample and out-of-sample.

As expected, we found that a player's position is, in most cases, an important factor for predicting the Fantacalcio ratings. However, it is somewhat counterintuitive that the inferences from these models suggest that the quality of a player's team, the opposing team, and the initial fantasy price do not account for much of the variation in the ratings (net of the other variables). It is also notable that the association between the current and lagged performance ratings —expressed by the average lagged rating—- is slightly different from zero after accounting for the other inputs into the models. Future research should consider whether other functional forms for describing associations over time are more appropriate, to what extent the inclusion of additional information in the models (e.g. injury data) improves the predictive performance, and if more informative priors can be developed at the position and team levels of the models. As is, the models may be over-shrinking these parameters. Another question to assess in the future is the division into training and testing datasets. In this chapter we split the season in half, but these models should also be useful dynamically, using data available through match day t to predict rating for match day $t + 1$.

The recent successes in the soccer analytics industry are due in large part to the increasing number of available metrics for analyzing and describing the game. However, even as the amount and variety of publicly available soccer data grows —particularly data at the individual player/match level— the interpretability and predictive relevance of that data will remain a question. In fact, it is not straightforward to identify whether a player is or is not performing well —in the Fantacalcio framework this is translated in collecting more point scores— based on metrics such as the total distance run over the course of a match, the number (or percentage) of passes successfully completed, the total number of shots, the number of shots on target, or the number of "dangerous" attacks. According to our current knowledge, the only attempt to using these and many other metrics for measuring player performance is the OPTA index, which positively weights certain game features (e.g., goals, assists, shots, minutes) and negatively weights others (e.g., missed passes, yellow cards, missed goals, etc.). At least we are not aware of other attempts but we do not have proprietary information about what teams and other companies are doing

(see www.optasports.com for further details about the firm and its activity). Despite its appeal, the weighting used for the index appears not to be formulated using statistical methodology and tools like principal component analysis, cluster analysis, or any kind of regression analysis.

Compared to attempts like the OPTA index, our ratings may be crude approximations to player performance since they gloss over many games events. But the formulation of an index based on as many variables as possible for describing the players' performances has not been the aim of this chapter. The attractiveness of our approach —not necessarily all of our particular choices in model construction but our approach in general— is that it is based on a coherent statistical framework: we have an outcome variable y (the player rating) that is actually available, probability models relating the outcome to predictors, the ability to add prior information into an analysis in a principled way, and the ability to propagate our uncertainty into the predictions by drawing from the posterior predictive distribution. Our approach is also transparent, fits naturally into powerful statistical frameworks for model criticism (e.g., posterior predictive checking), and can easily be modified by anyone who has different ideas about the form of the relationship between the outcome and predictors.

Chapter 4

Modeling the soccer outcome using bookmakers' information

4.1 Introduction

In recent years the challenge of modeling football outcomes has gained increasing attention, in large part due to the potential for making substantial profits in betting markets. According to the current literature, this task may be achieved by adopting essentially two different kinds of modeling strategies: the so called *direct* models for the number of goals scored by two competing teams and the *indirect* models for estimating the probability of the categorical outcome Win/Draw/Loss —hereafter, *three-way* process.

The basic assumption is that the numbers of goals scored by the two teams follow two Poisson distributions. Their dependence structure and the specification of their parameters are the most relevant further assumptions according to the literature and for this reason the first framework is of interest for us. The scores' dependence issue is in fact much debated, and the discussion can not yet be concluded. As one of the first contributors to the football scores' modeling, Maher (1982) used two conditionally independent Poisson distributions for the goals scored by the home team and the away team. Dixon and Coles (1997) started from the Maher's work and extended his model introducing a parametric dependence between the scores. This represents also the justification for the bivariate Poisson model, introduced in Karlis and Ntzoufras (2003) in a frequentist perspective and in Ntzoufras (2011) under a Bayesian perspective. On the other hand, Baio and Blangiardo (2010) assumed the

conditional independence within hierarchical Bayesian models, on the ground that the goals' correlation is already taken into account by the hierarchical structure. Similarly, Groll and Abedieh (2013) showed that up to a certain amount the scores' dependence of two competing teams may be explained by the inclusion of some specific teams' covariates in the linear predictors. On the other hand, Dixon and Robinson (1998) noted that modeling the dependence along a single match is worth: in such a case, a temporal structure in the ninety minutes is required.

The second common assumption is the inclusion in the models of some teams' effects to allow for the attack and the defense strengths of the competing teams. Generally, they are used for modeling the scoring rate of a given team, and in much of the mentioned literature they do not vary over time. Of course, this is a major limitation of these models. Dixon and Coles (1997) tried to overcome this problem by downweighting the likelihood exponentially over time, in order to reduce the impact of matches far from the current time of evaluation. Whereas Rue and Salvesen (2000) assumed that the teams' specific effects for the attack and the defense vary over the time, and they frame their analysis in a Bayesian context.

For our aims the scores' dependence assumption may be relaxed, and in this chapter we adopt a conditional independence assumption between the scores. This choice allows in fact for a simpler formulation for the likelihood function and simplifies the inclusion of the bookmakers' odds in our model. Concerning the dynamic assumption of the teams-specific effects, we use an autoregressive model by centering the effect of seasonal time τ at the lagged effect in $\tau - 1$ plus a fixed effect.

Whatever are the choices for the two assumptions discussed above, the model proposed in this context was built with both a descriptive and a predictive goal, and its parameters' estimates/model probabilities were often used for building efficient betting strategies (Dixon and Coles, 1997; Londono and Hassan, 2015). In fact, the well known expression 'beating the bookmakers' is often considered a cornerstone for whoever tries to predict soccer —or more generally, sports— results. As mentioned by Dixon and Coles (1997), to win money from the bookmakers requires a determination of probabilities which is sufficiently more accurate than those obtained from the odds. On the other hand, it is empirically known that betting odds are the most accurate source of information for forecasting sports performances (Štrumbelj, 2014). However, at least two issues deserve a deep analysis: how to determine probability forecasts from the raw betting odds and how to use this huge source of information within a

forecasting model (e.g., for the number of goals). Concerning the first point, it is well known that the betting odds do not correspond to precise probabilities; in fact, to make a profit, bookmakers set unfair odds, and they have a 'take' of 5-10%. In order to derive a set of fair probabilities from these odds, many authors used the so called *basic normalization* procedure, by normalizing the inverse odds up to their sum. Alternatively, Forrest et al. (2005) and Forrest and Simmons (2002) proposed a regression model-based approach, modeling the betting probabilities through an historical set of betting odds and match outcomes. But Štrumbelj (2014) showed that the so called Shin's procedure (Shin, 1991, 1993) gives overall the best results, being preferable both to the basic normalization and to the regression approaches. Concerning the second issue, a very sparse literature focused on using the existing betting odds as a *part* of a statistical model for improving the predictive accuracy and the model fit. Londono and Hassan (2015) used the betting odds for eliciting the hyperparameters of a Dirichlet distribution, and then updated them based on observations of the categorical three-way process. No author tried to implement a similar strategy within the framework of the direct models.

In this chapter we tried to fill the gap creating a bridge between the betting odds —and betting probabilities— on one hand and the statistical modeling of the scores on the other hand. Once we transform the betting odds into precise probabilities, we develop a procedure for (i) infer from these the implicit scoring intensities according to the bookmakers (ii) use these implicit intensities directly in the conditionally independent Poisson model for the scores, within a Bayesian perspective. We are interested both in the estimation of the models parameters, and in the prediction of a new bunch of matches. Intuitively, the latter task is much harder than the former one, since football is *per se* noisy and hardly predictable. However, we believe that the combination of the betting odds with an historical set of data may give predictions much more accurate than those obtained from a single source of information.

In Section 4.2 we introduce two methods proposed by the current literature for transforming the three-way bookmakers' betting odds in precise probabilities. In Section 4.3 we introduce the full model, along with the implicit scoring rates. The results and predictive accuracy of the model on the top-four European leagues — Bundesliga, Premier League, La Liga and Serie A— are presented in Section 4.4 and summarized through posterior probabilities and graphical checks. Section 4.6 concludes.

4.2 Transforming the betting odds in precise probabilities

The connection between betting odds and probabilities has been broadly investigated over the last decades. Before proceeding, we introduce the formal definition of odd and the related notation we are going to use in the rest of the chapter. An odd about a given event is usually specified as the amount of money we would win if we bet one unit on that event. Thus, the odd 2.5 corresponds to 2.5 euro (or dollars) we would win betting 1 euro. The inverse odd —usually denoted as 1:2.5— corresponds to the unfair probability associated to that event. Let $O_m = \{o_{Win}, o_{Draw}, o_{Loss}\}$, $\Pi_m = (\pi_{Win}, \pi_{Draw}, \pi_{Loss})$ and $\Delta_m = \{\text{'Win'}, \text{'Draw'}, \text{'Loss'}\}$ denote respectively the vector of the inverse betting odds, the vector of the estimated betting probabilities and the set of the three-way possible results for the m-th game. As widely known, the betting odds do not correspond to precise probabilities. In fact, the sum of the inverse odds for a single match is greater than one (Dixon and Coles, 1997) in order to guarantee the bookmakers' profit.

As mentioned by Štrumbelj (2014), there is empirical evidence that the betting odds are the most accurate available source of probability forecasts for sports; in other words, forecasts based on odds-probabilities have been shown to be better or at least as good as statistical models which use sport-specific predictors and/or expert tipsters.

However, some issues remain open. Among these, there is a strong debate on which method to use for inferring a set of unbiased probabilities from the raw betting odds. We can transform them into unbiased probabilities by using two procedures proposed in the literature: the *basic normalization* —dividing the inverse odds by the booksum, as broadly explained in Štrumbelj (2014)— and the *Shin's procedure* described in Shin (1991, 1993). Štrumbelj (2014), Cain et al. (2002, 2003) and Smith et al. (2009) showed that the Shin's probabilities improve over the basic normalization: in Štrumbelj (2014) this result has been achieved by the application of the Ranked Probability Score (RPS) (Epstein, 1969), which may be defined as a discrepancy measure between the probability of a three-way process outcome and the actual outcome.

In this chapter we do not actually focus on comparing these two procedures; rather,

we are interested in using the probabilities derived from them for statistical and prediction purposes, as will be more clear later.

(A) *Basic normalization*

$$\pi_i = \frac{o_i}{\beta},\ i \in \Delta_m, \tag{4.1}$$

where $\beta = \sum_i o_i$ is the so called booksum (Štrumbelj, 2014). The method gained a great popularity due to its simplicity.

(B) *Shin's procedure*

In the model proposed by Shin (1993), the bookmakers specify their odds in order to maximize their expected profit in a market with uninformed bettors and insider traders. The latters are those particular actors which, due to a superior information, are assumed to *already* know the outcome of a given event —e.g. football match, horse race, etc— before the event takes place. Their contribute in the global bets volume is quantified by the percentage z. Jullien et al. (1994) used the Shin's model for explicitely working out the expression for the betting probabilities:

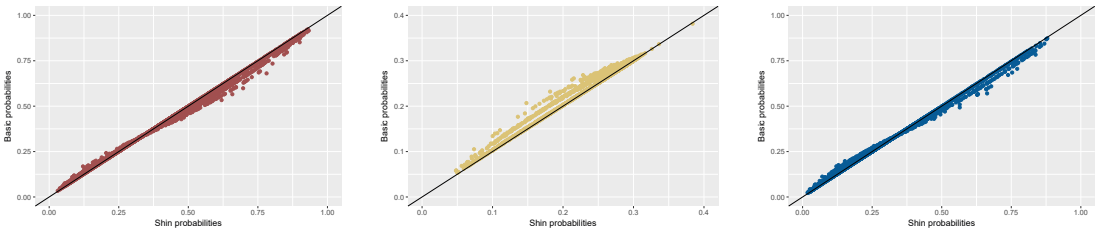
$$\pi(z)_i = \frac{\sqrt{z^2 + 4(1-z)\frac{o_i^2}{\sum_i o_i}} - z}{2(1-z)},\ i \in \Delta_m, \tag{4.2}$$

so that $\sum_{i=1}^{3} \pi(z)_i = 1$. The current literature refers to these as the Shin's probabilities. The formula above is a function depending on the insider trading rate z, which Jullien et al. (1994) suggested to estimate by nonlinear least squares as:

$$\underset{z}{\text{Argmin}}\{\sum_{i=1}^{3} \pi(z)_i - 1\}.$$

The value here obtained may be defined as the minimum rate of insider traders that yields precise probabilities corresponding to the vector of inverse betting odds O.

Both these methods yield precise probabilities, with the difference that the Shin's procedure is a function of the insider traders rate and needs to be minimized for every



(a) Home win (b) Draw (c) Away win

Figure 4.1: Comparison between Shin probabilities (x-axis) and basic normalized probabilities (y-axis) for the Spanish La Liga championship (seasons from 2007/2008 to 2016/2017), according to seven different bookmakers.

match. Figure 4.1 displays the three-way betting probabilities obtained through the two procedures described above. As may be noticed, the Draw probabilities obtained with the basic normalization tend to be higher than those obtained with the Shin's procedure. Conversely, as a Home win and an Away win tend to become more likely, the Shin's procedure tends to favor them.

As is intuitive, a higher probability of an home win should somehow be associated with a greater number of goals scored by the home team, and the same for an away team.

4.3 Model

4.3.1 Model for the scores

Let $\boldsymbol{y} = (y_{m1}, y_{m2})$ denote the vector of observed scores, where y_{m1} and y_{m2} are respectively the number of goals scored by the home team and by the away team in the m-th match of the dataset. According to the motivations provided by Baio and Blangiardo (2010), in this chapter we adopt a conditional independence assumption between the scores. This choice allows in fact for a simpler formulation for the likelihood function and for the direct inclusion of the bookmakers odds into the model through the Skellam distribution (Karlis and Ntzoufras, 2009). The model for the scores is then specified as

$$\begin{aligned} y_{m1}|\theta_{m1} &\sim \text{Poisson}(\theta_{m1}) \\ y_{m2}|\theta_{m2} &\sim \text{Poisson}(\theta_{m2}), \\ y_{m1} &\perp y_{m2}|\theta_{m1}, \theta_{m2}, \end{aligned} \tag{4.3}$$

where $\boldsymbol{y}$ is modelled as *conditionally* independent Poisson and the joint parameter $\boldsymbol{\theta} = (\theta_{m1}, \theta_{m2})$ represents the scoring intensities in the m-th game, respectively for the home team and for the away team. In what follows, we will refer to (4.3) as the *basic* model, which is estimated using the past scores. The main novelty of this chapter consists in enriching this specification by including the extra information which stems from the bookmakers betting odds. Thus, for each pair of match m and bookmaker $s,\ s = 1, ..., S$ the betting probabilities $\pi^s_{i,m},\ i \in \Delta_m$ derived with one of the methods in Section 4.2 may be used for finding out the values $\hat{\boldsymbol{\theta}}^s = (\hat{\theta}^s_{m1}, \hat{\theta}^s_{m2})$ which solve the following nonlinear system of two equations:

$$\begin{aligned} \pi^s_{Win,m} + \pi^s_{Draw,m} &= P(y_{m1} \geq y_{m2}|\theta^s_{m1}, \theta^s_{m2}) \\ \pi^s_{Loss,m} &= P(y_{m1} < y_{m2}|\theta^s_{m1}, \theta^s_{m2}). \end{aligned} \tag{4.4}$$

The existence of these values is guaranteed by the fact that, under (4.3), $y_{m1} - y_{m2} \sim PD(\theta_{m1}, \theta_{m2})$, where PD denote the Poisson-Difference distribution, also known as Skellam distribution, with parameters θ_{m1}, θ_{m2} and mean $\theta_{m1} - \theta_{m2}$. In such a way, we obtain for each pair (m, s) the *implicit* scoring rates $\hat{\theta}^s_{m1}, \hat{\theta}^s_{m2}$, somehow inferring the scoring intensities implicit in the three-way bookmakers' odds. Now, we consider our augmented dataset by including as auxiliary data the observed $\hat{\theta}^s_{m1}, \hat{\theta}^s_{m2}$: for every m, our new data vector is represented by

$$(\boldsymbol{y}, \hat{\boldsymbol{\theta}}^s) = (y_{m1}, y_{m2}, \hat{\theta}^s_{m1}, \hat{\theta}^s_{m2},\ s = 1, ..., S).$$

Now, from Equation (4.3) we move to the following specification:

$$\begin{aligned} y_{m1}|\theta_{m1}, \lambda_{m1} &\sim \text{Poisson}(p_{m1}\theta_{m1} + (1 - p_{m1})\lambda_{m1}) \\ y_{m2}|\theta_{m2}, \lambda_{m2} &\sim \text{Poisson}(p_{m2}\theta_{m2} + (1 - p_{m2})\lambda_{m2}), \end{aligned} \tag{4.5}$$

where $\lambda_{m1},\ \lambda_{m2}$ are bookmakers parameters introduced for modeling the additional data $\hat{\theta}^s_{m1}, \hat{\theta}^s_{m2},\ s = 1, ..., S$, as explained in the next section. Parameters p_{m1}, p_{m2} are

assigned a non-informative prior distribution, with hyper-parameters a and b, e.g. $p_{m\cdot} \sim B\text{eta}(a, b)$.

4.3.2 Model for the rates

Equation (4.5) introduced a convex combination for the Poisson parameters, accounting for both the scoring rates $\theta_{\cdot 1}, \theta_{\cdot 2}$ and the bookmakers' parameters $\lambda_{\cdot 1}, \lambda_{\cdot 2}$. Denoting with T the number of teams, the common specification for the scoring intensities is a log-linear model in which for each $t,\ t = 1, ..., T$:

$$\begin{aligned} \log(\theta_{m1}) &= \mu + att_{t[m]1} + def_{t[m]2} \\ \log(\theta_{m2}) &= att_{t[m]2} + def_{t[m]1} \end{aligned} \tag{4.6}$$

with the nested index $t[m]$ denoting the team t in the m-th game. The parameter μ represents the well known football advantage of playing at home, and is assumed to be constant for all the teams and over time, as in the current literature. The attack and the defence strengths of the competing teams are summarized respectively by the parameters att and def. Baio and Blangiardo (2010) and Dixon and Coles (1997) assume that these team-specific effects do not vary over the time, and this represents a major limitation in their models. In fact, Dixon and Robinson (1998) showed that the attack and the defence effects are not static and vary even during a single match; thus, a static assumption is often not reliable for making prediction and represents a crude approximation of the reality. Rue and Salvesen (2000) proposed a generalized linear Bayesian model in which the team-effects at match time τ are drawn from a Normal distribution centered at the team-effects at the match time $\tau - 1$, and with a variance term depending on the time difference. Their choice is appealing and more realistic, and we make a similar assumption considering the effects for the season τ following a Normal distribution centered at the previous seasonal effect plus a fixed component. For each $t = 1, \ldots, T,\ \tau = 2, \ldots, \mathcal{T}$:

$$\begin{aligned} att_{t,\tau} &\sim \mathsf{N}(\mu_{att} + att_{t,\tau-1}, \sigma^2_{att}) \\ def_{t,\tau} &\sim \mathsf{N}(\mu_{def} + def_{t,\tau-1}, \sigma^2_{def}), \end{aligned} \tag{4.7}$$

while for the first season we assume:

$$\begin{aligned} att_{t,1} &\sim \mathsf{N}(\mu_{att}, \sigma^2_{att}) \\ def_{t,1} &\sim \mathsf{N}(\mu_{def}, \sigma^2_{def}). \end{aligned} \tag{4.8}$$

As outlined in the literature, we need to impose a 'zero-sum' identifiability constraint within each season to these random effects

$$\sum_{t=1}^{T} att_{t,\tau} = 0, \qquad \sum_{t=1}^{T} def_{t,\tau} = 0, \quad t = 1, \ldots, T, \ \tau = 1, \ldots \mathcal{T},$$

whereas μ and the hyperparameters of our model are assigned weakly informative priors:

$$\begin{aligned} \mu, \mu_{att}, \mu_{def} \sim & \mathsf{N}(0, 10) \\ \sigma_{att}, \sigma_{def} \sim & \mathsf{Cauchy}^{+}(0, 2.5), \end{aligned}$$

where Cauchy^{+} denotes the half-Cauchy distribution, centered in 0 and with scale 2.5.[1] The team-specific effects modeled through Equation (4.7) and (4.8) are estimated from the past scores of the dataset. As expressed in (4.5), we add a level to the hierarchy, by including the implicit scoring rates as a separate data model. Given then a further level which consists of S bookmakers, it is natural to consider $\lambda_{m1}, \lambda_{m2}$ as the model parameters for the observed $\hat{\theta}^s_{m1}, \hat{\theta}^s_{m2}$. More precisely, they represent the means of two truncated Normal distributions for the further implicit scoring rates model:

$$\begin{aligned} \hat{\theta}^1_{m1}, ..., \hat{\theta}^S_{m1} &\sim \text{trunc}\mathsf{N}(\lambda_{m1}, \tau^2_1, 0, \infty) \\ \hat{\theta}^1_{m2}, ..., \hat{\theta}^S_{m2} &\sim \text{trunc}\mathsf{N}(\lambda_{m2}, \tau^2_2, 0, \infty), \end{aligned} \tag{4.9}$$

where $\text{trunc}\mathsf{N}(\mu, \sigma^2, a, b)$ is the common notation for the density of a truncated Normal with parameters $\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+$ and defined in the interval $[a, b]$. $\lambda_{m1}, \lambda_{m2}$ are in turn assigned two truncated Normal distributions:

$$\begin{aligned} \lambda_{m1} &\sim \text{trunc}\mathsf{N}(\alpha_1, 10, 0, \infty) \\ \lambda_{m2} &\sim \text{trunc}\mathsf{N}(\alpha_2, 10, 0, \infty), \end{aligned} \tag{4.10}$$

with hyperparameters α_1, α_2.

[1] On the choice of the half-Cauchy distribution for scale parameters, see Gelman et al. (2006).

4.4 Applications and results: top-four European leagues

4.4.1 Data

We collected the exact scores for the top-four European professional leagues —Italian Serie A, English Premier League, German Bundesliga and Spanish La Liga— from season 2007/2008 to 2016/2017. Moreover, we also collected all the betting odds —three-way odds and Over/Under odds— for the following bookmakers: Bet365, Bet&Win, Interwetten, Ladbrokes, Sportingbet, VC Bet, William Hill. All these data have been downloaded from the public available page `http://www.football-data.co.uk/`. We are both interested in (a) posterior predictive checks in terms of replicated data under our models and (b) out-of-sample predictions for a new dataset. According to point (b), which appears to be more appealing for fans, betters and statisticians, let $\mathcal{T}_r$ denote the *training set*, and $\mathcal{T}_s$ the *test set*. Our training set contains the results of 9 seasons for each professional league, and our test set contains the results of the 10-th season.

The model coding has been implemented in `WinBUGS` (Spiegelhalter et al., 2003) and in `Stan` (Stan Development Team, 2016*a*). We ran our MCMC simulation for $H = 5000$ iterations, with a burn-in period of 1000, and we monitored the convergence using the usual MCMC diagnostic (Gelman, Carlin, Stern and Rubin, 2014).

4.4.2 Parameter estimates

As broadly explained in Section 4.3, the model in (4.5) combines historical information about the scores and betting information about the odds. We acknowledge that the scoring rate is a convex combination that *borrows strengths* from both the sources of information. Figures 4.2- 4.5 display the posterior estimates for the attack and the defense parameters associated to the teams belonging to the top-four European leagues during the test set season 2016-2017. The bigger is the team-attack parameter, and the greater is the estimated attacking quality for that team; conversely, the lower is the team-defense parameter, and the better is estimated the defense power for that team. As a general comment, after reminding that these quantities are estimated using only the historical results, the pattern seems to reflect the actual

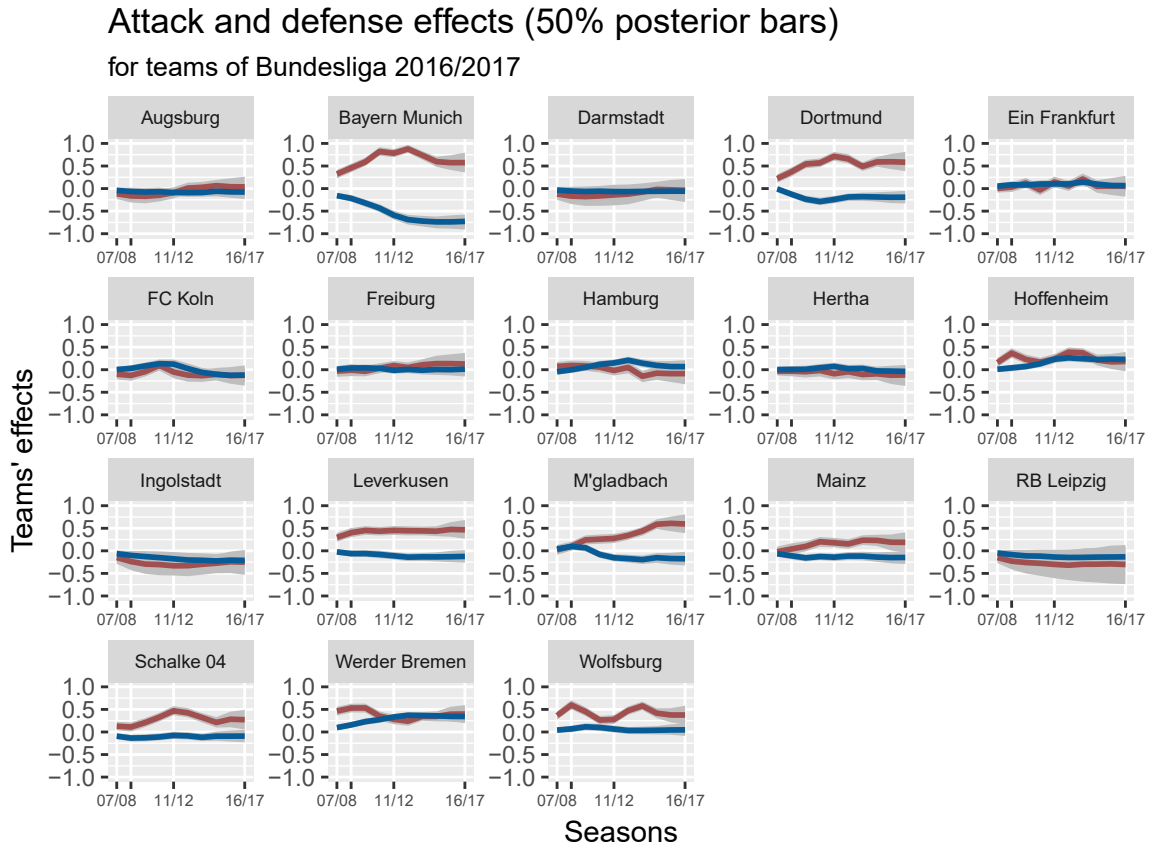


Figure 4.2: Posterior 50% confidence bars for the attack (red) and the defense (blue) effects along the ten seasons for the teams belonging to the Bundesliga 2016/2017. Wider posterior bars are associated to teams with fewer observations.

strength of the teams across the seasons. For example Juventus (Serie A), Bayern Munich (Bundesliga), Barcelona and Real Madrid (La Liga), Chelsea and Manchester City (Premier League) register the highest effects for the attack and the lowest for the defense across the nine considered seasons: consequently, the out-of-sample estimates for the tenth season mirror the previous performance. Conversely, weaker teams are associated with an inverse pattern: see for instance Ingolstadt (Bundesliga), Middlesbrough (Premier League), Osasuna (La Liga) and Pescara (Serie A). It is worth noting that some wide posterior bars are associated to those teams with fewer seasonal observations: in fact, for simplicity we do not account for a relegation system and some teams are less observed during the considered seasons.

Figure 4.6 displays the ordered 50% confidence bars for the marginal posteriors of the probabilities parameter $p_{m1}, p_{m2}, m = 1, \ldots, M$ which appear in (4.5), computed

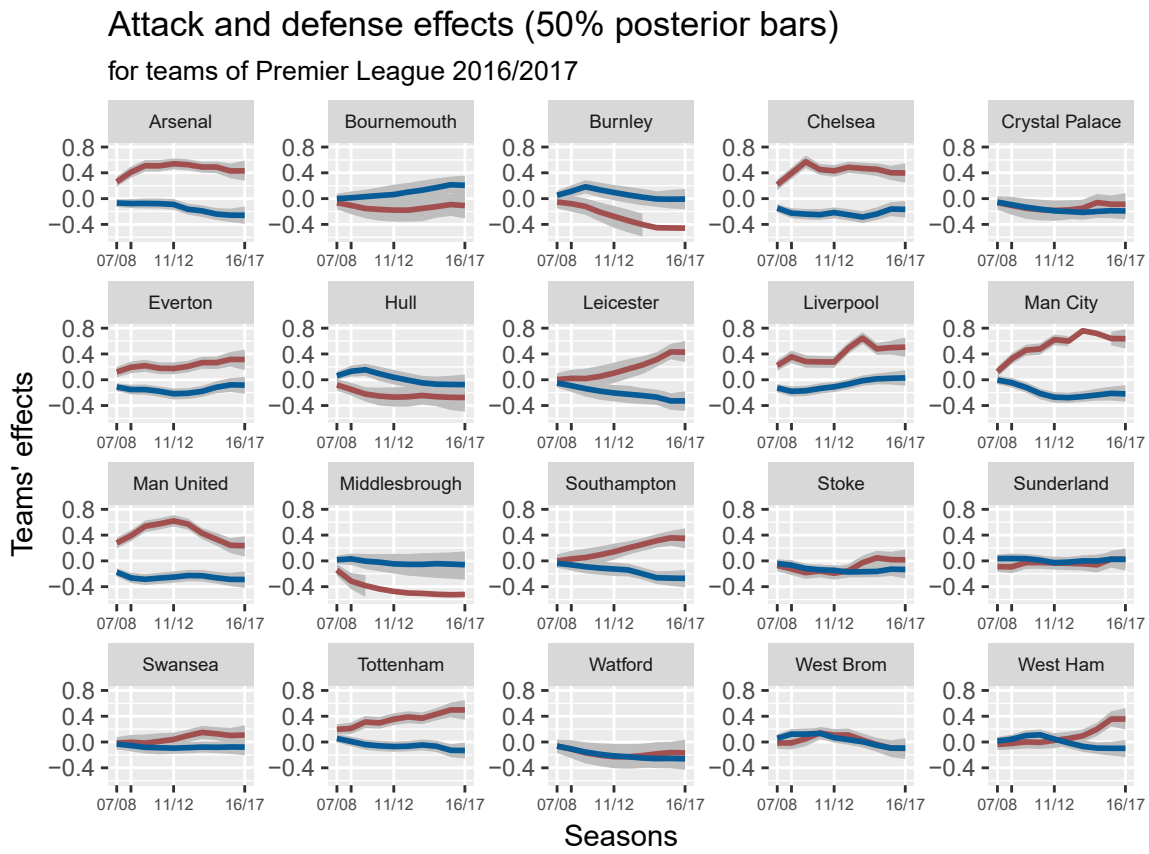


Figure 4.3: Posterior 50% confidence bars for the attack (red) and the defense (blue) effects along the ten seasons for the teams belonging to the Premier League 2016/2017. Wider posterior bars are associated to teams with fewer observations.

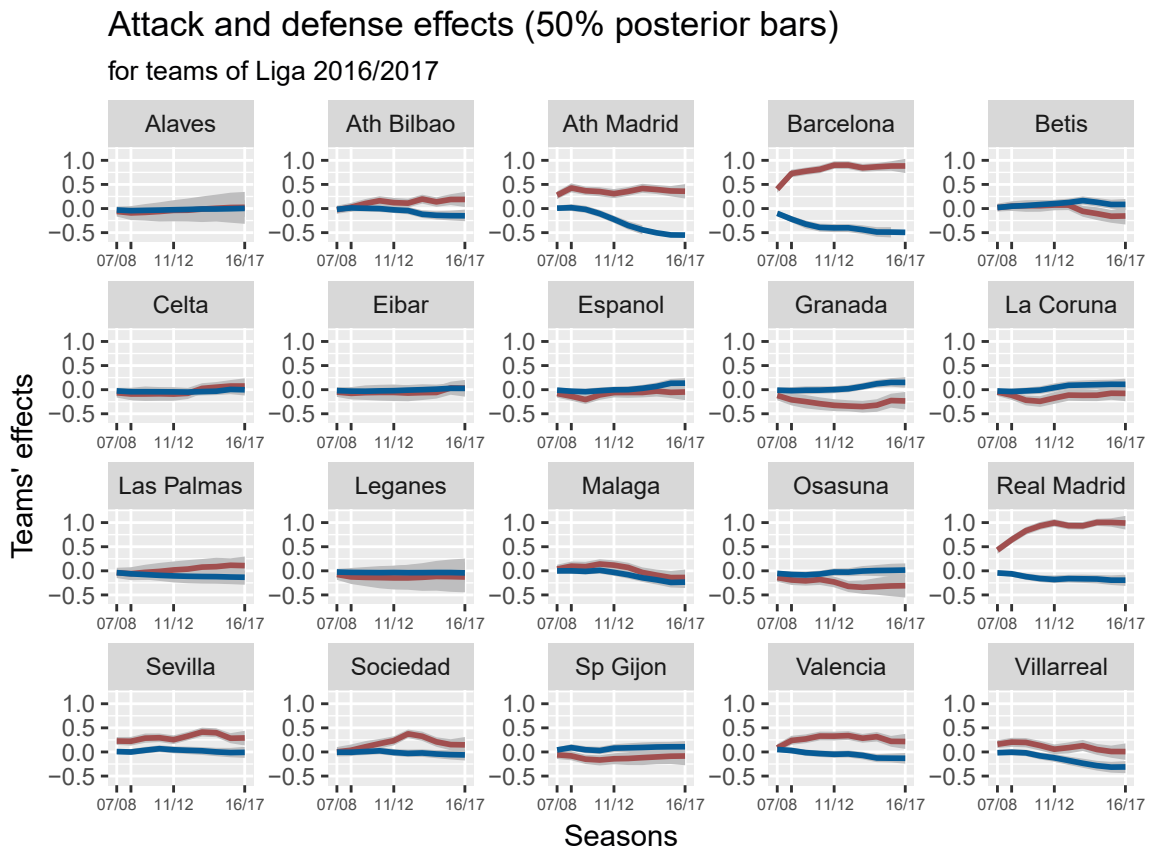


Figure 4.4: Posterior 50% confidence bars for the attack (red) and the defense (blue) effects along the ten seasons for the teams belonging to the La Liga 2016/2017. Wider posterior bars are associated to teams with fewer observations.

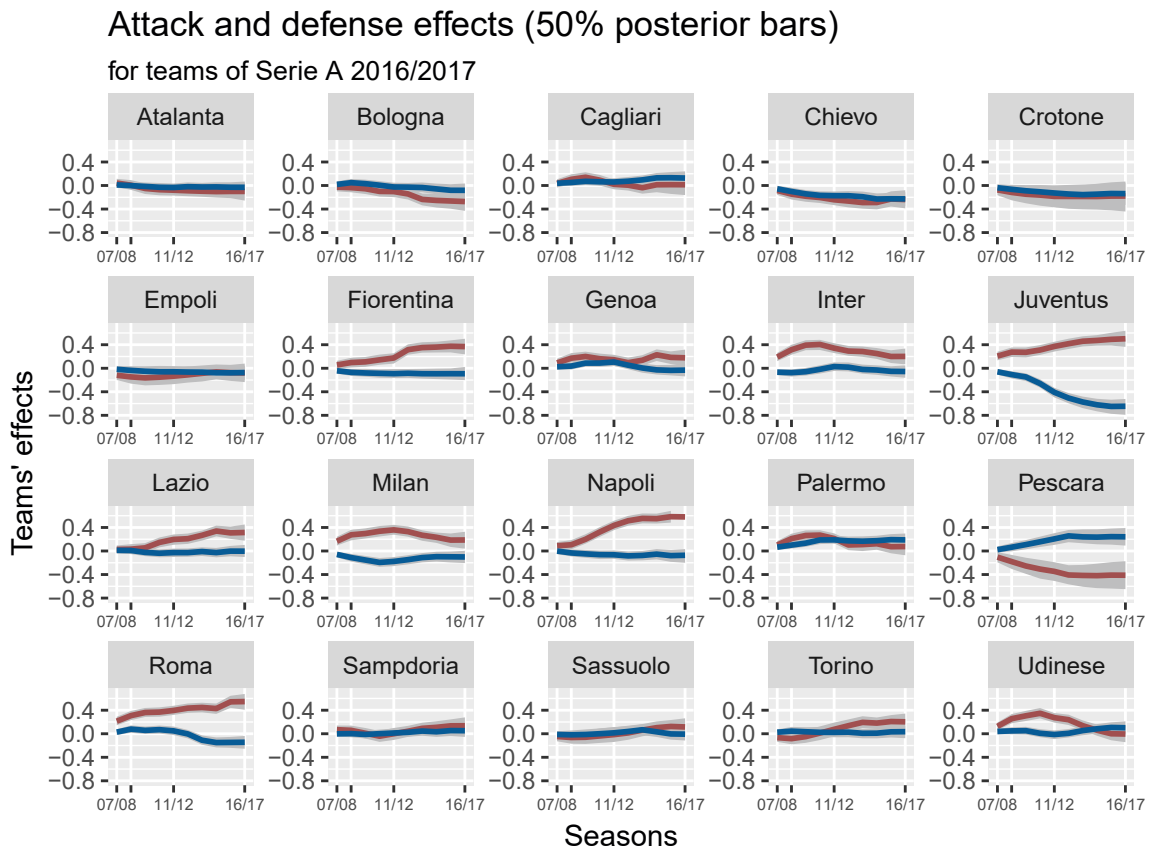
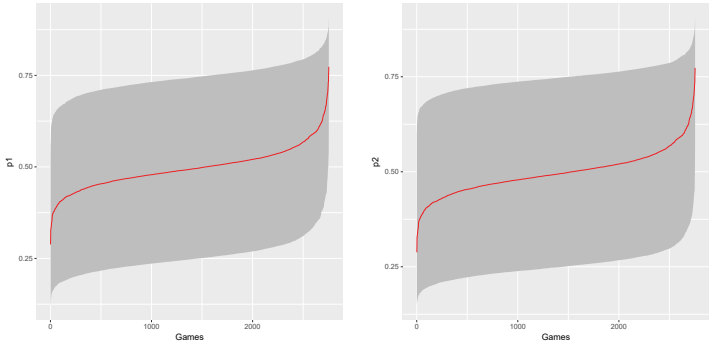


Figure 4.5: Posterior 50% confidence bars for the attack (red) and the defense (blue) effects along the ten seasons for the teams belonging to the Serie A 2016/2017. Wider posterior bars are associated to teams with fewer observations.



(a) p_1 (b) p_2

Figure 4.6: Ordered posterior 50% confidence bars for parameters p_{m1}, p_{m2} for German Bundesliga (from 2007-2008 to 2015-2016), 2754 matches.

for the German Bundesliga. Despite an high variability, these plots suggest that the amount of information which stems from the bookmakers is comparable with that arising from historical information. Perhaps, the convex combination in (4.5) seems to be an adequate option for our purposes.

4.4.3 Model fit

As broadly explained in Gelman, Carlin, Stern and Rubin (2014), once we obtain some estimates from a Bayesian model we should assess the fit of this model to the data at hand and the plausibility of such model given the initial purposes for which we built it. The principal tool designed for achieving this task is the *posterior predictive checking*. This post-model procedure consists in verifying whether some additional replicated data under our model are similar to the observed data. Thus, we draw simulated values y^{rep} from the joint predictive distribution of replicated data

$$p(y^{rep}|y) = \int_{\Theta} p(y^{rep}, \theta|y)d\theta = \int_{\Theta} p(\theta|y)p(y^{rep}|\theta)d\theta.$$

It is worth noting that the symbol y^{rep} used here is different from the symbol $\tilde{y}$ used in the next section. The former is just a replication of y, the latter is any future observable value.

Then, we define a test statistic $T(y)$ for assessing the discrepancy between the model and the data. A lack of fit of the data with respect to the posterior predictive

distribution may be measured by tail-area posterior probabilities, or Bayesian p-values

$$p_B = P(T(y^{rep}) > T(y)|y). \tag{4.11}$$

As a practical utility, we usually do not compute the integral in (4.4.3), but we compute the posterior predictive distribution through simulation. If we denote with $\theta^{(s)}$, $s = 1, ..., S$ the s-th MCMC draw from the posterior distribution of θ, we just draw y^{rep} from the predictive distribution $p(y^{rep}|\theta^{(s)})$. Hence, an estimate for the Bayesian p-value is given by the proportion of the S simulations for which the quantity $T(y^{rep\ (s)})$ exceeds the observed quantity $T(y)$. From an interpretative point of view, an extreme p-value —too close to 0 or 1— suggests a lack of fit of the model compared to the observed data.

Rather than comparing the posterior distribution of some statistics with their observed values (Gelman, Carlin, Stern and Rubin, 2014), we propose a slightly different approach allowing for a broader comparison of the replicated data under the model. Figures 4.7 and 4.8 display a probability plot for comparing the observed goals' difference $y_1 - y_2$ with the replicated distribution of $y_1^{rep} - y_2^{rep}$. The top-row graphs provide a predictive check over the training set, where darker regions are associated with a greater posterior probability. In correspondence of an observed goals' difference on the x-axis —here we consider only goals' differences bounded between -3 and 3— this graphical tool provides the most likely replicated goals' difference. A perfect — overfitting — model should display the black regions along the bisector. In fact, let consider the bottom-row graphs that display the observed distributions of the goals' difference for the leagues plotted above. These distributions are slightly asymmetric, but the maximum is always concentrated at zero. In our dataset a goals' difference amounting at zero is more likely than a goals' difference of, say, three. Perhaps, the top-row checks are comforting, since they suggest a good fit in correspondence of likely goals'difference — e.g. -1,0,1 — and a poorer fit when the goals difference turns out to be more rare –say, 3.

Figure 4.9 displays the replicated distributions $y_1^{rep} - y_2^{rep}$ (gray areas) and the observed goals' difference (red horizontal line) for the top-four European leagues. From this plots the fit of the model seems good: in other words, the replicated data under the model are plausible and close to the data at hand. As it may be noted, the variability of the replicated goals' difference amounting at -1, 0, 1 is greater than the variability for a goals' difference of -3 or 3. Moreover, the observed goals' differences

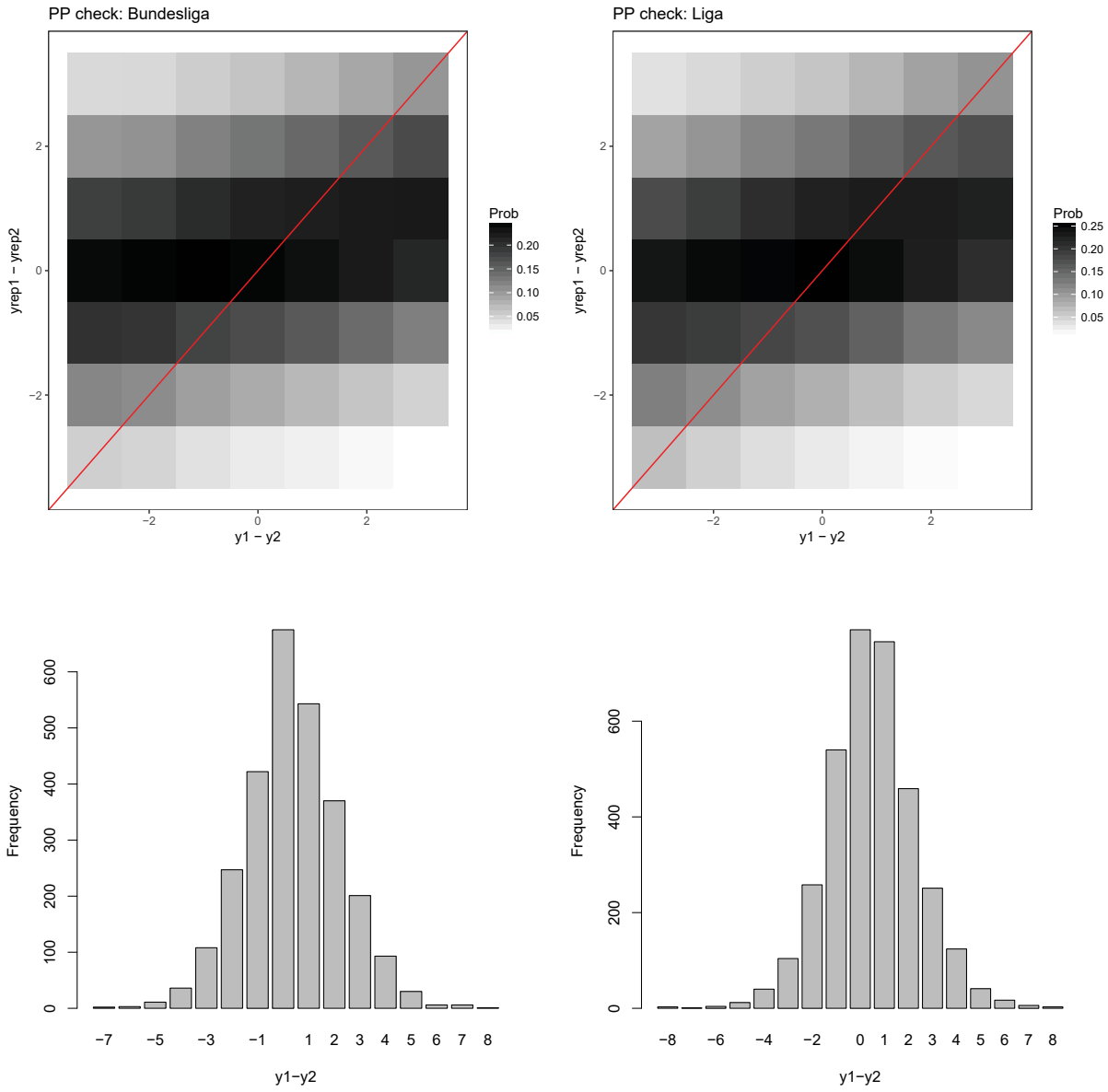


Figure 4.7: PP check: probability plot (top row), with darker regions associated to higher posterior probabilities and distribution of the observed goals' difference (bottom row). For the top graphs: on x-axis the observed goals' difference, on the y-axis the predicted goals' difference.

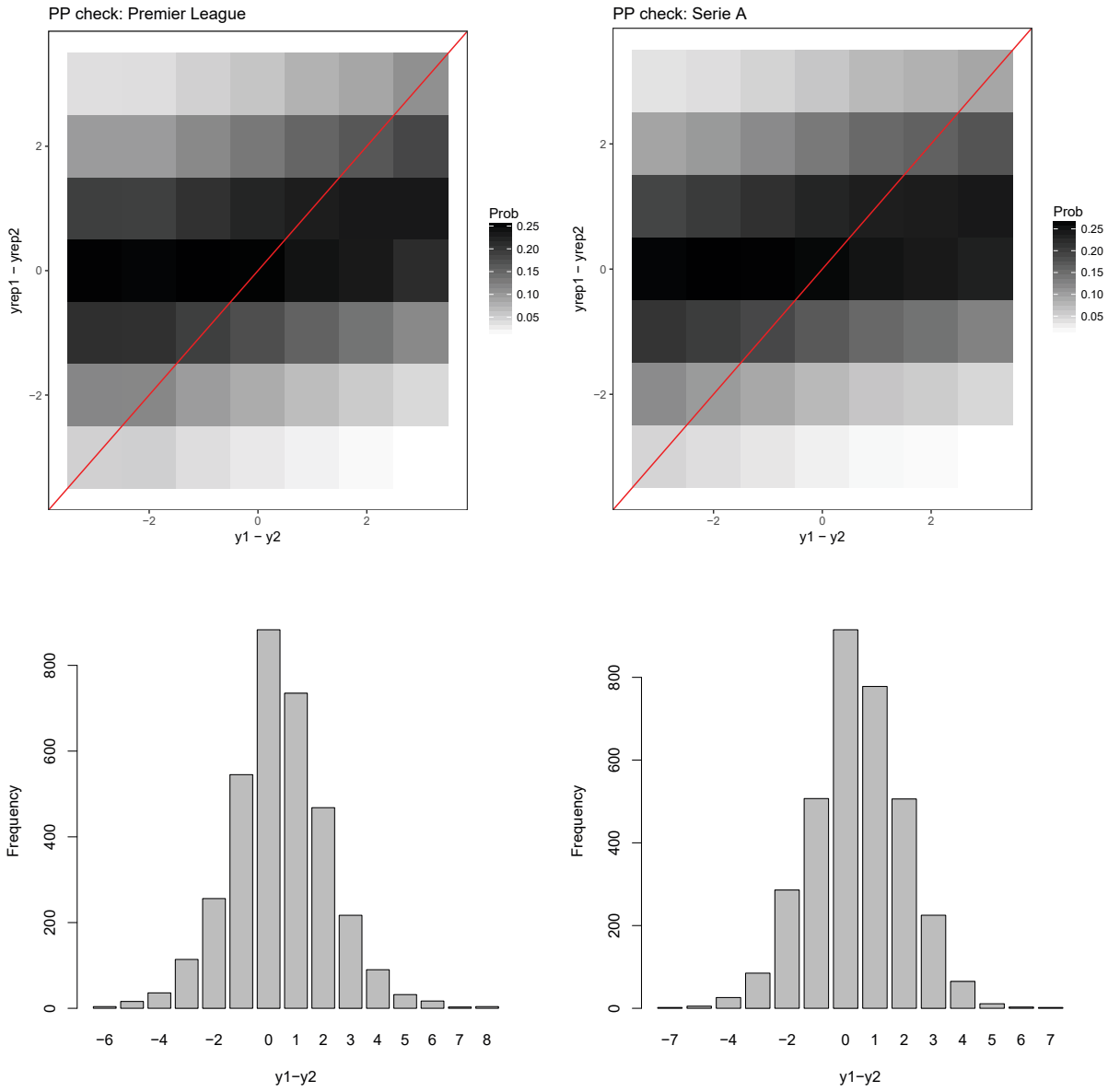
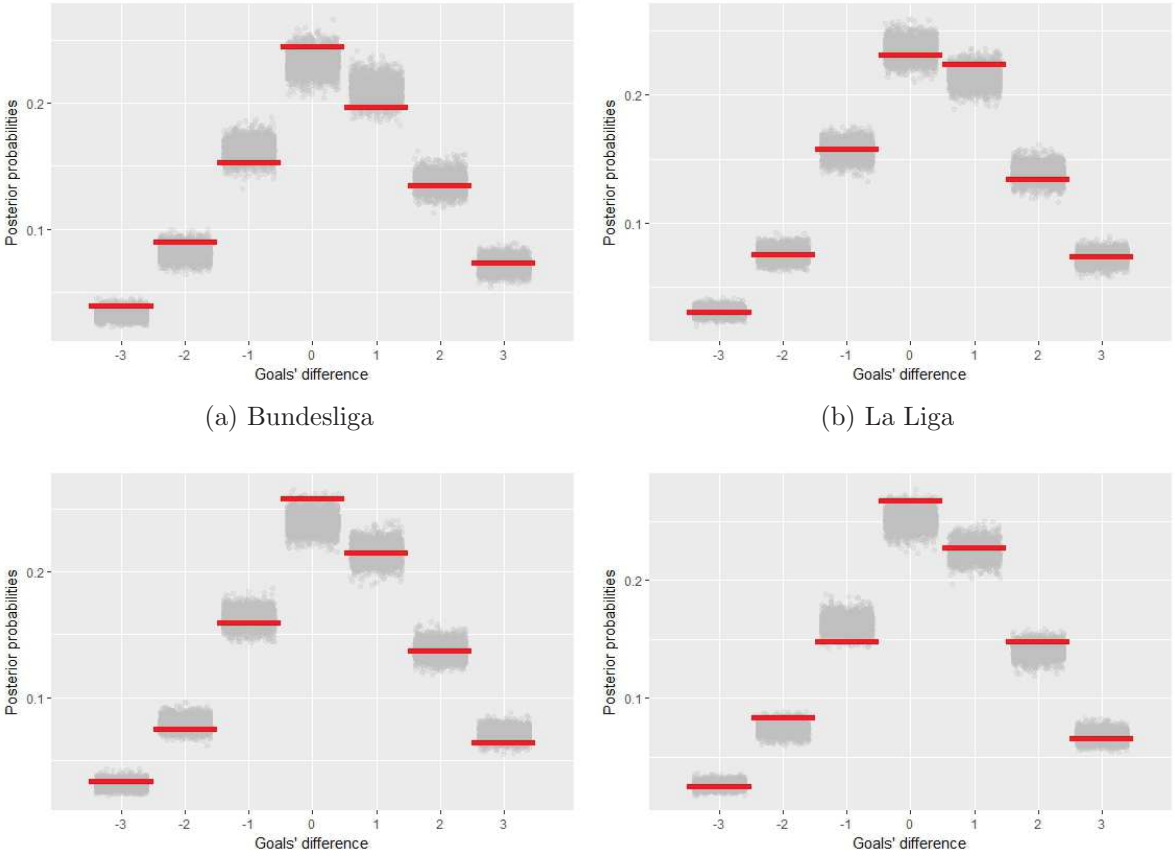


Figure 4.8: PP check: probability plot (top row), with darker regions associated to higher posterior probabilities and distribution of the observed goals' difference (bottom row). For the top graphs: on x-axis the observed goals' difference, on the y-axis the predicted goals' difference.



(a) Bundesliga

(b) La Liga

(c) Premier League

(d) Serie A

Figure 4.9: PP check for the goals' difference $y_1 - y_2$ against the replicated goals' difference $y_1^{rep} - y_2^{rep}$ for the top-four European leagues . For each league, the graphical posterior predictive checks show an excellent fit of the model to the data.

always fall within the replicated distributions. In correspondence of a draw —goal difference of 0— the observed goals' differences register an high posterior probability if compared with the corresponding replicated distribution.

4.4.4 Prediction and posterior probabilities

The main appeal of a statistical model relies on its predictive accuracy. As usual in a Bayesian framework, the prediction for a new dataset may be directly performed via the posterior predictive distribution for our unknown set of observable values. Following the same notation of Gelman, Carlin, Stern and Rubin (2014), let us denote with $\tilde{y}$ a generic unknown observable. Its distribution is then conditional on the

observed y,

$$p(\tilde{y}|y) = \int_{\Theta} p(\tilde{y}, \theta|y)d\theta = \int_{\Theta} p(\theta|y)p(\tilde{y}|\theta)d\theta,$$

where the conditional independence of y and $\tilde{y}$ given θ is assumed. Let $([\tilde{y}_{m1}]_h, [\tilde{y}_{m2}]_h)$ denote the outcome for the m-th match at the h-th iteration of the Markov chain, with $h = 1, ..., H$. Figure 4.10 displays the posterior predictive distribution for Real Madrid-Barcelona, Spanish La Liga 2016/2017, and for Sampdoria-Juventus, Italian Serie A. The dashed blue line in the top row plots indicates the observed result, respectively (2,3) for the first match and (0,1) for the second match. According to the model, the most likely result for the first game is (2,1), with an associated posterior probability slightly greater than 0.08. Whereas the most likely result coincide with the actual result (0,1) for the second game. The bottom-raw plots display the posterior predictive distribution of a given match as well: here a red square is in correspondence of the observed result and darker regions are associated to higher posterior probabilities.

These plots are not actually suggesting a most likely result: it would be smart betting on an event with an associated probability about 0.09? Maybe, not. Rather, these plots provide a picture that acknowledges the large uncertainty of the predictive distribution. We are not really interested on a model that often indicates as most likely a rare result that has been actually observed; we suspect, in fact, that a model which would favor the outcome (2,3) as most (or quite) likely, probably is not a good model. Rather, being aware of the unpredictable nature of football, we would like to grasp the posterior uncertainty of a match outcome in such a way that the actual result is not extreme in the predictive distribution.

Table 4.1 and Table 4.2 report respectively the estimated posterior probabilities for each team being the first, the second, the third and the first relegated, the second relegated and the third relegated for each of the top-four leagues, together with the observed rank and the achieved points. At the beginning of the 2016-2017 season, Bayern Munich had an estimated probability 0.8168 of winning the German league, as it actually did; in Italy, Juventus had an high probability of being the first (0.592) as well. Conversely, Chelsea had a low associated probability to win the league at the beginning of the season, and this is mainly due to the bad results obtained by Chelsea in the last years. Of course, the model does not account for the players'/managers'

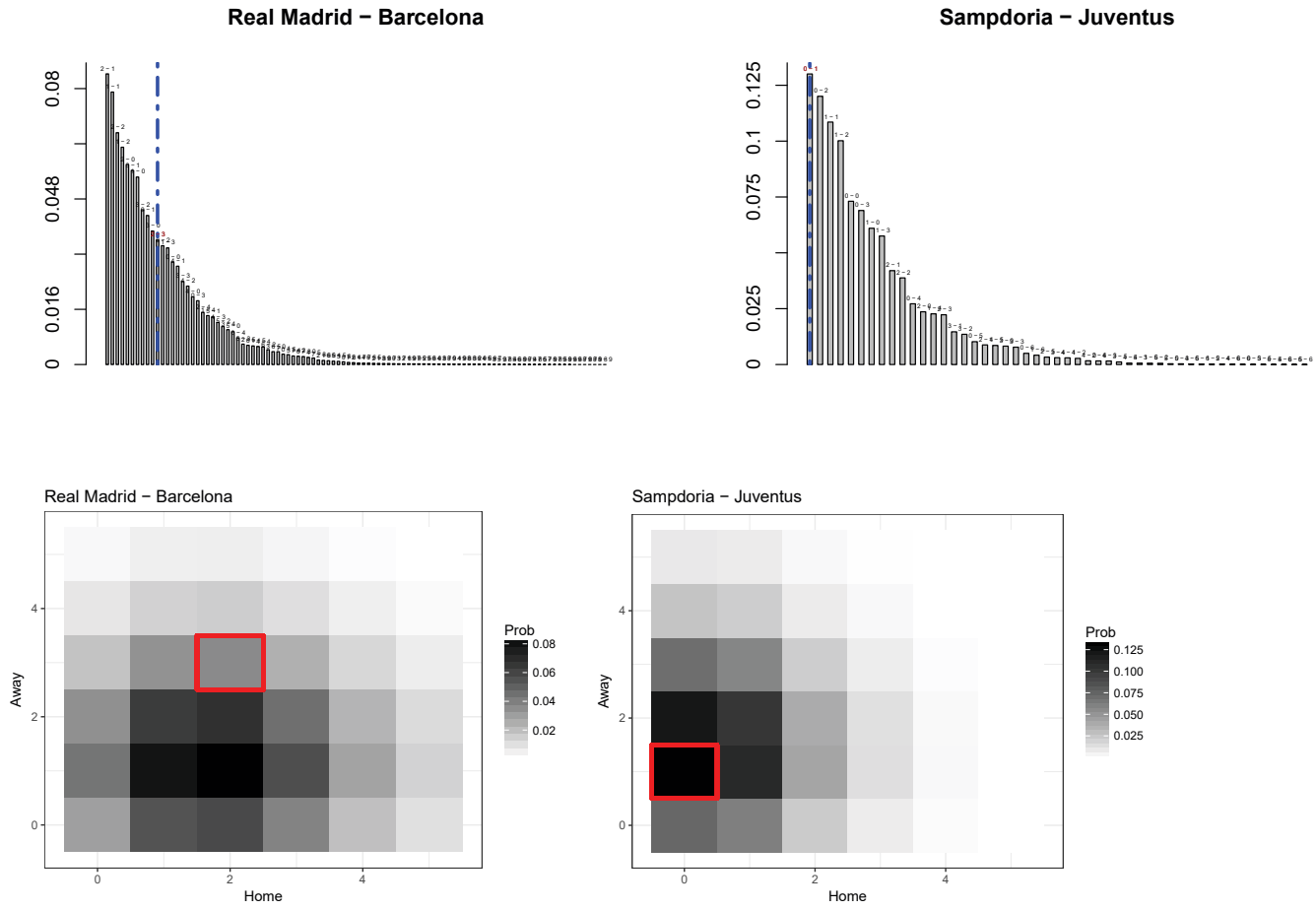


Figure 4.10: Posterior predictive distribution of the possible results for the match Real Madrid-Barcelona, Spanish La Liga 2016/2017, and Sampdoria-Juventus, Italian Serie A 2016-2017. Both the plots report the posterior uncertainty related to the exact predicted outcome. In the bottom row plots, darker regions are associated with higher posterior probability and the red square is in correspondence of the observed result.

Table 4.1: Estimated posterior probabilities for each team being the first, the second and the third in the Bundesliga, Premier League, La Liga and Serie A 2016-2017 together with the observed rank and the number of points achieved.

	Team	P(1st)	P(2nd)	P(3rd)	Actual rank	Points
	Bayern Munich	0.8168	0.1508	0.0248	1	82
	RB Leipzig	0.008	0.0284	0.0608	2	67
	Dortmund	0.1332	0.4712	0.1856	3	64
	Chelsea	0.1396	0.1592	0.1584	1	93
	Tottenham	0.1096	0.132	0.1424	2	86
	Man City	0.3904	0.2004	0.1388	3	78
	Real Madrid	0.3868	0.4844	0.1076	1	93
	Barcelona	0.5652	0.3536	0.0728	2	90
	Ath Madrid	0.046	0.1348	0.5556	3	78
	Juventus	0.592	0.2335	0.107	1	91
	Roma	0.1535	0.263	0.2595	2	87
	Napoli	0.206	0.2965	0.213	3	86

transfer market occurring in the summer period. In July 2016, Chelsea hired Antonio Conte, one of the best European managers, who won the English Premier League at his first attempt. For what concerns the relegated teams, it is worth noting the high estimated probability associated to Pescara of being the worst team of the Italian league (0.46). Globally, the model appears able to identify the teams with an associated high relegation's posterior probability.

Figure 4.11 provides posterior 50 % confidence bars (gray ribbons) for the predicted achieved points for each team in top-four European leagues 2016-2017 at the end of the respective seasons, together with the observed final ranks. At a first glance, the four predicted posterior ranks appear to detect a pattern similar to the observed ranks, with only a few exceptions. As may be noticed for Bundesliga (Panel (a)), Bayern Munich's prediction mirrors his actual strength in the 2016-2017 season, whereas RB Leipzig was definitely underestimated by the model. Still, the model can not handle the budget's information, and RB Leipzig was one of the richest teams in the Bundesliga 2016-2017. In the English Premier League (Panel (b)) Chelsea was definitely underestimated by the model, whereas Manchester City actually gained the

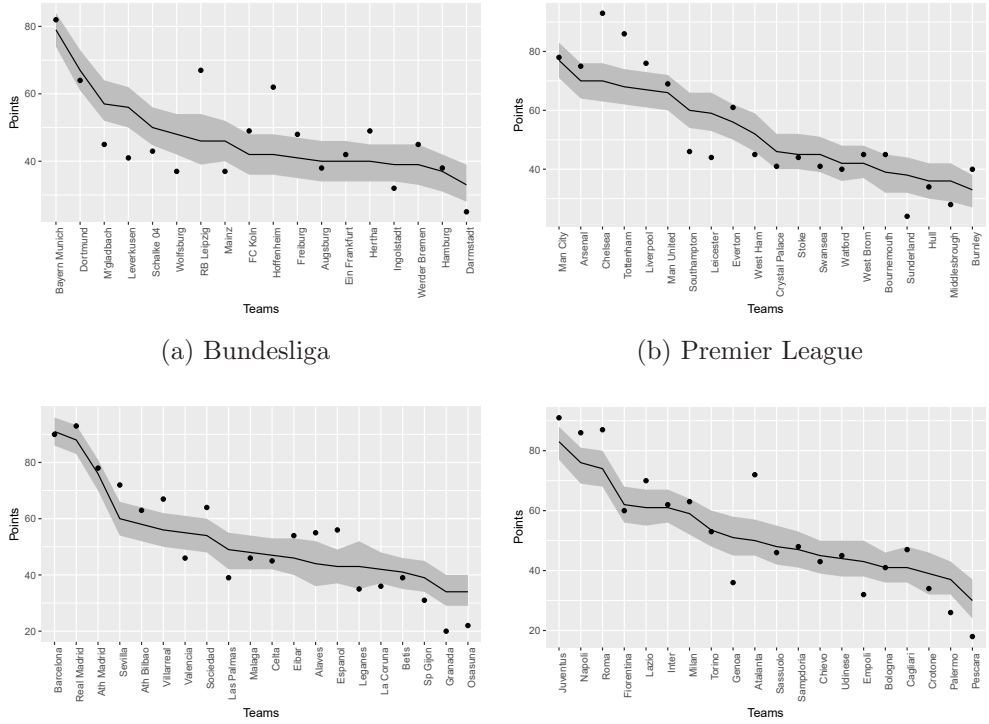
Table 4.2: Estimated posterior probabilities for each team being the first, the second and the third relegated team in the Bundesliga, Premier League, La Liga and Serie A 2016-2017, together with the observed rank and the number of points achieved.

	Team	P(1st rel)	P(2nd rel)	P(3d rel)	Actual rank	Points
	Wolfsburg	0.0212	0.0236	0.0064	16	37
	Ingolstadt	0.0952	0.0904	0.0912	17	32
	Darmstadt	0.1192	0.1552	0.2528	18	25
	Hull	0.1384	0.1512	0.1428	18	34
	Middlesbrough	0.118	0.1448	0.1812	19	28
	Sunderland	0.1272	0.1228	0.1144	20	24
	Sp Gijon	0.1132	0.1112	0.1016	18	31
	Osasuna	0.1464	0.174	0.228	19	22
	Granada	0.138	0.1748	0.2476	20	20
	Empoli	0.0795	0.066	0.0415	18	32
	Palermo	0.132	0.1765	0.1205	19	26
	Pescara	0.1215	0.178	0.46	20	18

predicted number of points (78). The predicted pattern for the Spanish La Liga (Panel (c)) is extremely close to the observed one, apart for the winner (our model favored Barcelona, second in the observed rank). The worst teams (Sporting Gijon, Osasuna and Granada) are correctly predicted to be relegated. Also for the Italian Serie A the predicted ranks globally match the observed ranks. The outlier is represented by Atalanta, a team that performed incredibly well and gained the Europa League's qualification at the end of the last season. As a general comment, we may conclude that these plots show a good model calibration, since more or less half of the observed points fall in the posterior 50 % confidence bars.

4.5 Betting strategy

In this section we provide a real betting experiment, assessing the performance of our model compared to the existing betting odds. In a betting strategy, two main questions arise: it is worth betting on a given single match? If so, how much is worth betting?



(a) Bundesliga (b) Premier League

(c) La Liga (d) Serie A

Figure 4.11: Posterior 50% confidence bars (gray ribbons) for the achieved final points of the top-four European leagues 2016-2017. Black points are the observed points. Black lines are the posterior medians. At a first glance, the pattern of the predicted ranks appears to match the pattern of the observed ranks and the model calibration appears satisfying.

4.5.1 Three-way bets

In Section 4.2 we described two different procedures for inferring a vector of betting probabilities Π from the inverse odds vector O. The common expression 'beating the bookmakers' may be interpreted in two distinct ways: from a probabilistic point of view and from a profitable point of view. According to the first definition, which is more appealing for statisticians, a bookmaker is beaten whenever our matches' probabilities are more favorable than his probabilities. Let $\pi^s_{i,m}$ denote as before the betting probability provided by the s-th bookmaker for the m-th game, with $i \in \Delta_m = \{\text{'Win'}, \text{'Draw'}, \text{'Loss'}\}$. On the other hand, let Y_{m1} and Y_{m2} denote the random variables representing the number of goals scored by two teams in the m-th match. From our model in (4.5) we can compute the following three-way model's

Table 4.3: Three-way bets: average correct probability $\bar{p}$ obtained through our model, Shin probabilities and basic probabilities (here we take the average of the seven considered bookmakers). Greater values indicate better predictive accuracy.

	Model	Shin	Basic
Bundesliga	0.4010	0.4100	0.4072
Premier League	0.4349	0.4516	0.4480
La Liga	0.4553	0.4584	0.4549
Serie A	0.4430	0.4554	0.4507

posterior probabilities: $p_{Win,m} = P(Y_{m1} > Y_{m2})$, $p_{Draw,m} = P(Y_{m1} = Y_{m2})$, $p_{Loss,m} = P(Y_{m1} < Y_{m2})$ for each $m \in \mathcal{T}_s$, using the results of the Skellam distribution outlined in Section 4.3. In fact, $Y_{m1} - Y_{m2} \sim PD(\hat{\gamma}_{m1}, \hat{\gamma}_{m2})$, where $\hat{\gamma}_{m1} = \hat{p}_{m1}\hat{\theta}_{m1} + (1-\hat{p}_{m1})\hat{\lambda}_{m1}$ and $\hat{\gamma}_{m2} = \hat{p}_{m2}\hat{\theta}_{m2} + (1-\hat{p}_{m2})\hat{\lambda}_{m2}$ are the convex combinations of the posterior estimates obtained through the MCMC sampling. Thus, the global average probability of a correct prediction for our model may be defined as

$$\bar{p} = \frac{1}{M}\sum_{m=1}^{M}\prod_{i\in\Delta_m} {p_{i,m}}^{\delta_{im}}, \tag{4.12}$$

where δ_{im} denotes the Kronecker's delta, with $\delta_{im} = 1$ if the observed result at the m-th match is i, $i \in \Delta_m$. This quantity serves as a global measure of performance for comparing the predictive accuracy between the posterior match probabilities provided by the model and those obtained from the bookmakers' odds. As reported in Table 4.3, our model is very close to the bookmakers' probabilities (Shin's method and basic procedure). At a first glance, one may be tempted to say that according to this measure our model does not improve the bookmakers' probabilities. However, as explained below, this index is only an averaged measure of the predictive power, which does not take into account the possible profits for the single matches.

According to the second definition, beating the bookmaker means earn money through our model's probabilities. In what follows we drop the bookmaker's subscripts s for easing the notation. Let us introduce the expected profit at the m-th game under our model as

$$\text{Exp-profit}_m = \sum_{i\in\Delta_m} p_{i,m}/o_{i,m}. \tag{4.13}$$

The betting strategy A is the following: for each match, bet one unit on the three-way match outcome with the highest expected return Exp-profit$_{i,m}$, by solving the simple expression

$$\max_{i \in \Delta_m} \{p_{i,m}/o_{i,m}\}.$$

In this way we are uniformly betting one unit for each match. But we could need something more sophisticated, because different matches may require different bets, due for instance to their variability. Rue and Salvesen (2000) suggest to put different amounts basing each bet on the match's profit variability. Let Var-profit$_m$ denote the variance of the profit for the match m. They found that the optimal bet for the m-th game is given by

$$\beta_{i,m} = \max\{0, \text{Exp-profit}_{i,m}/\text{Var-profit}_{i,m}\}, \tag{4.14}$$

where they choose to bet on that outcome $i \in \Delta_m$ such that $\beta_{i,m}$Exp-profit$_{i,m}$ is the greatest. We adopted both the strategies and the expected profits are reported in Table 4 and 5. At a first glance, it is evident that betting with our posterior model probabilities yields high positive returns for each league and each bookmaker; conversely, if we bet with the betting odds probabilities we would always incur in a sure loss. This is an empirical confirm for the performance of the model and suggests that the measure $\bar{p}$ alone does not mean necessarily nothing in terms of profitable strategies. As a second consideration, strategy B yields higher profits than strategy A.

4.5.2 Over/Under bets

The over/under (O/U) bets represent one the greatest source of the football betting. An O/U bet is a wager consisting in the prediction of a specific game's statistic. In football this is often translated in guessing whether the sum score of a match is lower or greater than two. The greatest appeal of a direct model is predicting all the games' features connected with the number of goals; in this case, the posterior probabilities for the total number of goals $Y_{m1}+Y_{m2}$ are: $p_{Over,m} = P(Y_{m1}+Y_{m2} > 2)$ and $p_{Under,m} = P(Y_{m1} + Y_{m2} \leq 2)$. From probability theory, we know that

$$Y_{m1} + Y_{m2} \sim \text{Poisson}(\gamma_{m1} + \gamma_{m2}),$$

Table 4.4: Strategy A: expected profits (%/100) for the seven considered bookmakers, for each of the top-four European leagues.

		Bet365	Bwin	Interwetten	Ladbrokes	Sportingbet	VC Bet	W. Hill
Bundesliga	Model	0.180	0.157	0.135	0.148	0.226	0.169	0.177
	Shin	-0.032	-0.032	-0.038	-0.046	-0.008	-0.034	-0.022
	Basic	-0.047	-0.050	-0.054	-0.061	-0.019	-0.047	-0.039
Premier League	Model	0.242	0.188	0.157	0.198	0.247	0.216	0.248
	Shin	-0.011	-0.026	-0.043	-0.029	-0.007	-0.036	-0.011
	Basic	-0.048	-0.050	-0.055	-0.059	-0.020	-0.046	-0.036
La Liga	Model	0.100	0.085	0.07	0.073	0.138	0.095	0.104
	Shin	-0.032	-0.035	-0.0335	-0.0386	-0.006	-0.034	-0.026
	Basic	-0.028	-0.048	-0.055	-0.046	-0.020	-0.044	-0.027
Serie A	Model	0.180	0.154	0.088	0.14	0.228	0.156	0.199
	Shin	-0.028	-0.037	-0.051	-0.039	-0.003	-0.027	-0.024
	Basic	-0.048	-0.050	-0.074	-0.060	-0.020	-0.046	-0.040

Table 4.5: Strategy B: expected profits (%/100) for the seven considered bookmakers, for each of the top-four European leagues. The value for the variance profit is set to one.

		Bet365	Bwin	Interwetten	Ladbrokes	Sportingbet	VC Bet	W. Hill
Bundesliga	Model	0.209	0.184	0.164	0.175	0.266	0.199	0.200
	Shin	-0.028	-0.027	-0.035	-0.043	-0.005	-0.031	-0.018
	Basic	-0.047	-0.050	-0.054	-0.060	-0.019	-0.046	-0.038
Premier League	Model	0.295	0.232	0.206	0.244	0.300	0.266	0.299
	Shin	-0.007	-0.022	-0.041	-0.026	-0.004	-0.034	-0.007
	Basic	-0.028	-0.048	-0.055	-0.046	-0.020	-0.044	-0.027
La Liga	Model	0.127	0.111	0.113	0.092	0.162	0.120	0.121
	Shin	-0.028	-0.032	-0.028	-0.034	-0.003	-0.031	-0.024
	Basic	-0.048	-0.049	-0.055	-0.058	-0.020	-0.046	-0.036
Serie A	Model	0.241	0.197	0.136	0.189	0.287	0.200	0.253
	Shin	-0.023	-0.034	-0.046	-0.034	0.001	-0.022	-0.020
	Basic	-0.048	-0.050	-0.074	-0.060	-0.020	-0.046	-0.040

Table 4.6: O/U bets: average correct probability $\bar{p}$ obtained through our model and through the basic probabilities (we take here the average of the seven considered bookmakers). In the last column, the expected profits (%/100). Greater $\bar{p}$ values indicate better predictive accuracy.

	Model	Basic	Profit
Bundesliga	0.514	0.512	0.045
Premier League	0.513	0.514	0.0498
La Liga	0.532	0.533	0.039
Serie A	0.517	0.521	0.065

where γ_{m1}, γ_{m2} are the convex parameters in (4.5). As in the three-way process, we may compute the average correct probability for these bets. Table 4.6 reports the average correct probability measure $\bar{p}$ for the O/U bets. The model predictions and the bookmakers predictions are very close, around 0.5. As evident from the last column, here the profits —playing with the analogous Strategy A of the tree-way process— are positive but lower than the three-way process profits. These results suggest a slightly good ability to predict the total number of goals, being this prediction comparable to no much more than flipping a coin. Still, even if the $\bar{p}$ measure under the model is not often greater than the same measure for the bookmakers basic probabilities, the expected profit are always positive. Here also, it is worth keep in mind that if we played with the bookmakers probabilities, we would incur in a sure loss.

Figure 4.12 displays a comparison between the O/U probabilities implied by the bookmakers and by our model. For Premier League and Serie A the model tends to predict greater probabilities for the Under, while for Bundesliga and La Liga the model and the bookmakers tend to be closer. Together with the expected profits in Table 4.6, these graphs seem to provide an empirical suggestion that the expected profits are greater in correspondence of higher Under probabilities provided by the model.

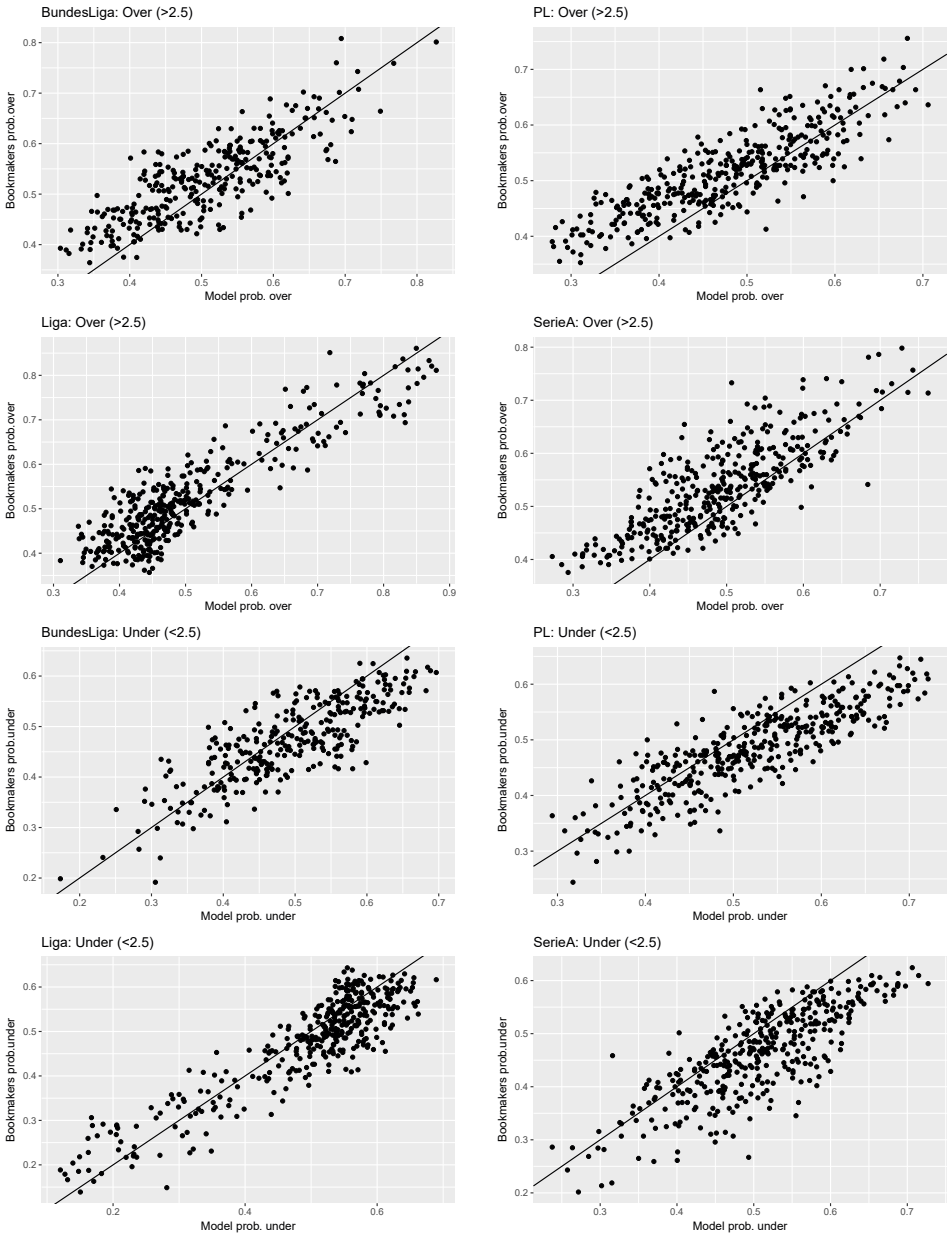


Figure 4.12: Model and bookmakers' O/U probabilities for each of the top-four European leagues

4.6 Discussion and further work

We have proposed a new hierarchical Bayesian Poisson model in which the rates are convex combinations of parameters accounting for two different sources of data: the bookmakers' betting odds and the historical match results. We transformed the betting odds in precise probabilities and we worked out the bookmakers scoring rates through the Skellam distribution. A wide graphical and numerical analysis for the top-four European leagues has shown a good predictive accuracy for our model and surprising results in terms of expected profits. These results confirm on one hand that the information contained in the betting odds is relevant in terms of football prediction; on the other hand, combining it with historical data allows for a natural extension of the existing models for the football scores.

Appendix A

Proof of Theorem 1

Due to distribution-constant definition, $g(T(\boldsymbol{y})|\boldsymbol{\theta}) = g(T(\boldsymbol{y}))$ and then

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|T(\boldsymbol{y}))/g(T(\boldsymbol{y})|\boldsymbol{\theta}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|T(\boldsymbol{y})).$$

Furthermore, $\pi(\boldsymbol{\theta}|T(\boldsymbol{y})) \propto g(T(\boldsymbol{y})|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta})$. □

Proof of Theorem 2

Proof. For simplicity of notation we denote with α the baseline prior $\pi_b(\theta)$, with γ the informative prior $\pi(\theta)$ and with β the mixture prior $\varphi(\theta) = \psi_{m^*}\pi_b(\theta) + (1 - \psi_{m^*})\pi(\theta)$. Furthermore, we abbreviate the weight ψ_{m^*} as ψ. Unless otherwise stated, the dependence of the quantities introduced in Section 2.4 on the parameter $\theta \in \mathbb{R}$ is here implicit. We compute the negative second log-derivative for the mixture prior (2.17) in general terms as

$$D_\varphi = -\frac{d^2 \log\{\varphi(\theta)\}}{d\theta^2} = -\frac{d^2 \log\{\psi\pi_b(\theta) + (1-\psi)\pi(\theta)\}}{d\theta^2} = \tag{15}$$

$$= -\frac{d}{d\theta}\left[\frac{\psi\alpha^{'} + (1-\psi)\gamma^{'}}{\psi\alpha + (1-\psi)\gamma}\right] = \tag{16}$$

$$= \frac{(\psi\alpha^{'} + (1-\psi)\gamma^{'})^2 - (\psi\alpha^{''} + (1-\psi)\gamma^{''})(\psi\alpha + (1-\psi)\gamma)}{(\psi\alpha + (1-\psi)\gamma)^2}. \tag{17}$$

After some simple expansions we can rewrite (17) and apply some minorations:

$$D_\varphi = \frac{\psi^2[(\alpha^{'})^2 - \alpha^{''}\alpha] + (1-\psi)^2(\gamma^{'})^2 + 2\psi(1-\psi)\gamma^{'}\alpha^{'}}{(\psi\alpha + (1-\psi)\gamma)^2} -$$

$$-\frac{\psi(1-\psi)\alpha^{''}\gamma + \psi(1-\psi)\alpha\gamma^{''} + (1-\psi)^2\gamma\gamma^{''}}{(\psi\alpha + (1-\psi)\gamma)^2} \leq$$

$$\leq \left[\frac{(\alpha^{'})^2 - \alpha^{''}\alpha}{\alpha^2}\right] + \frac{(1-\psi)^2(\gamma^{'})^2 - (1-\psi)^2\gamma\gamma^{''}}{(1-\psi)^2\gamma^2} +$$

$$+\frac{2\psi(1-\psi)\gamma^{'}\alpha^{'}-\psi(1-\psi)\alpha^{''}\gamma-\psi(1-\psi)\alpha\gamma^{''}}{\psi^2\alpha^2}=$$

$$=D_\alpha+K_1, \tag{18}$$

where K_1 collects all the terms which do not enter in D_α. Analogously, we can find another minoration:

$$D_\varphi\leq\left[\frac{(\gamma^{'})^2-\gamma^{''}\gamma}{\gamma^2}\right]+\frac{\psi^2(\alpha^{'})^2-\psi^2\alpha\alpha^{''}+2\psi(1-\psi)\gamma^{'}\alpha^{'}}{\psi^2\alpha^2}-$$
$$-\frac{\psi(1-\psi)\alpha^{''}\gamma+\psi(1-\psi)\alpha\gamma^{''}}{\psi^2\alpha^2}=$$

$$=D_{\gamma+K_2}. \tag{19}$$

From (18) and (19) it stems that

$$K_1-K_2=\left[\frac{(\gamma^{'})^2-\gamma^{''}\gamma}{\gamma^2}\right]-\left[\frac{(\alpha^{'})^2-\alpha^{''}\alpha}{\alpha^2}\right]=D_\gamma-D_\alpha,$$

with $D_\gamma-D_\alpha>0$ for assumption (see Table 2.1). In what follows we abbreviate D_φ as D. Hence we have found the following conditions

$$\begin{cases}\mathbf{A}\ D\leq D_\alpha+K_1\\ \mathbf{B}\ D\leq D_\gamma+K_2\end{cases} \tag{20}$$

Condition **B** implies $D\leq D_\gamma+K_2+(K_1-K_2)=D_\gamma+K_1$ and yields the further condition

$$\mathbf{C}\ D\leq D_\gamma+K_1.$$

Thus, we may collect the three conditions already found

$$\begin{cases}\mathbf{A}\ D\leq D_\alpha+K_1\\ \mathbf{B}\ D\leq D_\gamma+K_2\\ \mathbf{C}\ D\leq D_\gamma+K_1\end{cases} \tag{21}$$

Now we may distinguish three separate cases which satisfy the condition $K_1-K_2>0$:

(a) $K_1>K_2>0$

We use conditions $\mathbf{B}, \mathbf{C}$

$$\begin{cases} \mathbf{B}\ D \leq D_\gamma + K_2 \\ \mathbf{C}\ D \leq D_\gamma + K_1 \end{cases} \rightarrow \begin{cases} 2D \leq 2D_\gamma + 2K_2 \\ D \leq D_\gamma + 2K_1 \end{cases} \rightarrow \begin{cases} D \leq D_\gamma + 2(K_2 - K_1) \leq D_\gamma \\ - \end{cases} \tag{22}$$

and we conclude that $D \leq D_\gamma$.

(b) $K_1 > 0,\ K_2 < 0$
By applying condition $\mathbf{B}$, it follows $D \leq D_\gamma$.

(c) $K_1 < 0,\ K_2 < 0$
By applying condition $\mathbf{B}$ or $\mathbf{C}$, it follows $D \leq D_\gamma$.

We have proved that for any possible sign of $K_1,\ K_2$, $D \leq D_\gamma$. By definition of effective sample size from Morita et al. (2008) we know that

$$\begin{aligned} ESS(\varphi(\theta)) &= \underset{m \in \mathbb{N}}{\text{Argmin}}\{\delta(m, \bar{\theta}, \varphi, q_m)\} = \\ &= \underset{m \in \mathbb{N}}{\text{Argmin}}\{|D - D_{q_m}(\bar{\theta})|\}, \end{aligned}$$

evaluated in the plug-in estimate $\bar{\theta} = E_\pi(\theta)$. From Table 2.1 we also know that the observed information of the baseline posterior D_{q_m} is a linear function of the sample size m and is increasing:

$$\frac{dD_{q_m}}{dm} > 0,\ \forall m \in \mathbb{N}.$$

Thus we may conclude that from $D \leq D_\pi$ it follows:

$$\begin{aligned} ESS(\varphi(\theta)) &= \underset{m \in \mathbb{N}}{\text{Argmin}}\{|D_\varphi(\bar{\theta}) - D_{q_{m(\theta|y)}}(\bar{\theta})|\} \leq \\ &\leq \underset{m \in \mathbb{N}}{\text{Argmin}}\{|D - D_{q_m}(\bar{\theta})|\} = ESS(\pi(\theta))\ .\ \square \end{aligned}$$

Logistic regression for phase I trial

Algorithm for computing the ESS (Morita et al., 2008)

- According to the definitions in (2.22), we compute the following quantities: $D_{\pi,1} = (\tilde{\sigma}_{\mu}^2)^{-1},\ D_{\pi,2} = (\tilde{\sigma}_{\beta}^2)^{-1}$.
- We need to compute $D_{q,1}(m, \boldsymbol{\theta}, X_m, \boldsymbol{y}_m) = \sum_{i=1}^{m} \pi(X_i, \boldsymbol{\theta})\{1 - \pi(X_i, \boldsymbol{\theta})\}$, $D_{q,2}(m, \boldsymbol{\theta}, X_m, \boldsymbol{y}_m) = \sum_{i=1}^{m} X_i^2 \pi(X_i, \theta)\{1 - \pi(X_i, \theta)\}$.
- It turns out that $\int D_{q_m,j} f(\boldsymbol{y}_m) d\boldsymbol{y}_m$ —where $f(\boldsymbol{y}_m)$ is the likelihood (2.43) evaluated in correspondence of fixed values for $\boldsymbol{\theta}$ and $\boldsymbol{X}$— cannot be computed analytically and need to be computed through Monte Carlo simulation. Before of proceeding, let us notice that $D_{q,1}(m, \boldsymbol{\theta}, X_m, \boldsymbol{y}_m)$ and $D_{q,2}(m, \boldsymbol{\theta}, X_m, \boldsymbol{y}_m)$ depend on X_m but not on $\boldsymbol{y}_m$, and this simplifies the simulation procedure. We may replace them respectively with the new notations $D_{q,1}(m, \boldsymbol{\theta}, X_m)$ and $D_{q,2}(m, \boldsymbol{\theta}, X_m)$.
- Assuming a uniform distribution for the doses, we draw $X_1^{(t)}, ..., X_6^{(t)}$ independently from $\{X_1, ..., X_6\}$ with probability 1/6 each, for $t = 1, ..., 100000$.
- Use the Monte Carlo average $T^{-1} \sum_{t=1}^{T} D_{q,j}(m, \boldsymbol{\theta}, X_m)$ in place of $\int D_{q_m,j} f(\boldsymbol{y}_m) d\boldsymbol{y}_m$, for $j = 1, 2$.
- Compute $\delta_1(m_\mu, \bar{\boldsymbol{\theta}}, \pi_\mu, q_{m_\mu})$, $\delta_2(m_\beta, \bar{\boldsymbol{\theta}}, \pi_\beta, q_{m_\beta})$ and $\delta(m, \bar{\boldsymbol{\theta}}, \pi, q_m)$.
- $ESS(\pi(\mu)), ESS(\pi(\beta))$ and $ESS(\pi(\boldsymbol{\theta}))$ are the interpolated values of the sample sizes m_μ, m_β, m minimizing δ_1, δ_2 and δ respectively.

Appendix B

Stan code for the MIX model

```
data{
  // Dimensions
  int<lower=0> N;  // Number of players
  int<lower=1> J;  // Number of positions
  int<lower=1> T;  // Number of matches
  int<lower=1> K;  // Number of team clusters
  int<lower=1> D;  // Number of mixture components

  // Variables
  vector[T] y[N];                          // Outcome
  int<lower=1,upper=J> position[N];        // Position
  int<lower=1,upper=K> team[N];            // Team cluster
  int<lower=1,upper=K> opp_team[N, T-1];   // Opponent team cluster for each game
  int<lower=0,upper=1> home[N, T-1];       // Home/Away variable (0=Away, 1=Home)
  vector[N] price_std;                     // Initial price for every player
  real avg_rating[N, T-1];                 // Lagged average ratings

  // Out-of-sample stuff
  int<lower=1> T_twiddle;                                  // Number of games
  int<lower=0,upper=1> home_twiddle[N, T_twiddle];         // Home/Away
  int<lower=1,upper=K> opp_team_twiddle[N, T_twiddle];     // Opponent team cluster
}
parameters {
  // For non-centered parameterizations
```

```
  vector[N] alpha_raw;                // Player intercepts
  vector[K] gamma_raw;                // Team-cluster incercepts
  vector[K] beta_raw;                 // Opponent team-cluster intercepts
  vector[J] rho_raw;                  // Position intercepts

  vector[J] lambda;                   // Coefs on lagged average rating
  vector[J] delta;                    // Coefs on standardized price
  real alpha0;                        // Global intercept
  real theta;                         // Coef on home/away indicator

  // Scale parameters
  real<lower=0> sigma_y;
  real<lower=0> sigma_alpha;
  real<lower=0> sigma_beta;
  real<lower=0> sigma_gamma;
  real<lower=0> sigma_rho;

  // Parameters in logit submodel
  vector[J] zeta;
  real pzero;
}
transformed parameters{
  // Non-centered parameterizations
  vector[N] alpha = alpha_raw * sigma_alpha;
  vector[K] beta = beta_raw * sigma_beta;
  vector[K] gamma = gamma_raw * sigma_gamma;
  vector[J] rho = rho_raw * sigma_rho;

  vector[T] eta[N];
  for (n in 1:N) {
    eta[n, 1] = 0; // just needs some value
    for (l in 2:T) {
      eta[n,l] =
         alpha0
```

```
        + alpha[n]
        + delta[position[n]] * price_std[n]
        + (gamma[team[n]] + beta[opp_team[n, l-1]])
        + rho[position[n]]
        + theta * home[n, l-1]
        + lambda[position[n]] * avg_rating[n, l-1];
    }
  }
}
model{
  // Mixture
  for (n in 1:N) {
    for (l in 2:T) {
      real pi_eta = pzero + zeta[position[n]] * avg_rating[n, l-1];
      target +=
        log_mix(
          inv_logit(pi_eta),
          normal_lpdf(y[n,l] | eta[n,l], sigma_y),
          normal_lpdf(y[n,l] | 0, 0.1) // 0
        );
    }
  }

  // Log-priors
  target += normal_lpdf(alpha0 | 0, 5);
  target +=
    normal_lpdf(alpha_raw | 0, 1)
    + normal_lpdf(sigma_alpha | 0, 2.5);
  target +=
    normal_lpdf(rho_raw | 0, 1)
    + normal_lpdf(sigma_rho | 0, 2.5);
  target +=
    normal_lpdf(beta_raw | 0, 1)
    + normal_lpdf(sigma_beta | 0, 2.5);
```

```
  target +=
    normal_lpdf(gamma_raw | 0, 1)
    + normal_lpdf(sigma_gamma | 0, 2.5);

  target += normal_lpdf(theta | 0, 2.5);
  target += normal_lpdf(lambda | 0, 1);
  target += normal_lpdf(delta | 0, 5);
  target += cauchy_lpdf(sigma_y | 0, 5);

  target += normal_lpdf(zeta | 0, 1);
  target += normal_lpdf(pzero | 0, 2.5);
}
generated quantities{
  vector[T] y_rep[N];                    // In-sample replications/predictions
  vector[T-1] pi_eta_rep[N];
  int<lower=0, upper=1> V[N, T-1];

  vector[T_twiddle] eta_twiddle[N];
  real y_twiddle[N, T_twiddle];          // Out-of-sample predictions
  vector[T_twiddle] pi_eta_twiddle[N];
  int<lower=0, upper=1> V_twiddle[N, T_twiddle];
  real avg_rating_twiddle[N, T_twiddle];

  avg_rating_twiddle[,1] = avg_rating[,T-1];
  y_twiddle[,1] = y[,T];
  y_rep[,1] = y[,1];

  for (n in 1:N) {
    for (l in 2:T) {
      pi_eta_rep[n, l-1] =
        pzero + zeta[position[n]] * avg_rating[n,l-1];
      V[n, l-1] = bernoulli_logit_rng(pi_eta_rep[n,l-1]);
      y_rep[n,l] = (V[n, l-1] == 1) ? normal_rng(eta[n,l], sigma_y) : 0;
    }
```

```
}

for (n in 1:N) {
  pi_eta_twiddle[n,1] =
    pzero+ zeta[position[n]] * avg_rating_twiddle[n,1];
    V_twiddle[n,1] = bernoulli_logit_rng(pi_eta_twiddle[n,1]);
      eta_twiddle[n,1] =
                        alpha0
                        + alpha[n]
                        + (gamma[team[n]]+beta[opp_team_twiddle[n,1]])
                        + delta[position[n]]*price_std[n]
                        + theta * home_twiddle[n,1]+rho[position[n]]
                        + lambda[position[n]] * avg_rating[n,T-1];
}

for (n in 1:N) {
  for (l in 2:T_twiddle) {
    avg_rating_twiddle[n,l] =
      sum(y_twiddle[n,1:(l-1)]) / size(y_twiddle[n,1:(l-1)]);
    pi_eta_twiddle[n,l] =
      pzero+ zeta[position[n]] * avg_rating_twiddle[n,l];
      V_twiddle[n,l] = bernoulli_logit_rng(pi_eta_twiddle[n,l]);
            eta_twiddle[n,l] =
                              alpha0
                              + alpha[n]
                              +(beta[opp_team_twiddle[n,l]]+gamma[team[n]])
                              + delta[position[n]]*price_std[n]
                              + theta * home_twiddle[n,l]+rho[position[n]]
                              + lambda[position[n]] * avg_rating_twiddle[n,l-1];

    y_twiddle[n,l] =
      (V_twiddle[n,l] == 1) ? normal_rng(eta_twiddle[n,l], sigma_y  ) : 0;
  }
}}
```

Appendix C

JAGS code for German Bundesliga

```
mod.BundesLiga.mixture.hier.B<-"model{
# Likelihood:

    for (n in 1:ngames_train){
          for (s in 1:agenzie){

          theta1_bm[n,s] ~dnorm(lambda1_book[n], tau1_book[n])T(0,)
          theta2_bm[n,s] ~dnorm(lambda2_book[n], tau2_book[n])T(0,)
           }

          theta1_hat[n]<-pClust1[n,1]*thetaofClust[n,1,1] +
                                      pClust1[n,2]*thetaofClust[n,1,2]
          theta2_hat[n]<-pClust2[n,1]*thetaofClust[n,2,1] +
                                      pClust2[n,2]*thetaofClust[n,2,2]



          score1[n] ~ dpois(theta1_hat[n])
          score2[n] ~ dpois(theta2_hat[n])

 # Average Scoring intensities (accounting for mixing components)

  log(thetaofClust[n,1,1])<-mu+att[team1[n], season[n]]+def[team2[n], season[n]]
  log(thetaofClust[n,1,2])<-log(lambda1_book[n])
  log(thetaofClust[n,2,1])<-att[team2[n], season[n]]+def[team1[n], season[n]]
```

```
  log(thetaofClust[n,2,2])<-log(lambda2_book[n])

#priors for lambda1_book, lambda2_book

        lambda1_book[n]~ dnorm(theta1_bm_mean[n], 0.01)T(0,)
        lambda2_book[n]~dnorm(theta2_bm_mean[n], 0.01)T(0,)

        tau1_book[n]<-pow(sigma1.y[n],-2)
        tau2_book[n]<-pow(sigma2.y[n],-2)

        sigma1.y[n]~dnorm(0, alpha)T(0,)
        sigma2.y[n]~dnorm(0, beta)T(0,)

}

# Predictive distribution for the number of goals scored

 for (n in 1:ngames_test){
    for (s in 1:agenzie){

       theta1_bm_prev[n,s] ~dnorm(lambda1_book_prev[n], tau1_book[n])T(0,)
       theta2_bm_prev[n,s] ~dnorm(lambda2_book_prev[n], tau2_book[n])T(0,)
 }


   theta1_hat_prev[n]<-pClust1_prev[n,1]*thetaofClust_prev[n,1,1]+
                       pClust1_prev[n,2]*thetaofClust_prev[n,1,2]
   theta2_hat_prev[n]<-pClust2_prev[n,1]*thetaofClust_prev[n,2,1]+
                        pClust2_prev[n,2]*thetaofClust_prev[n,2,2]

   score1_prev[n] ~ dpois(theta1_hat_prev[n])
   score2_prev[n] ~ dpois(theta2_hat_prev[n])
```

```
   pClust1_prev[n,1:2] ~ ddirch(onesRepNclust_prev[1:2])
   pClust2_prev[n,1:2] ~ ddirch(onesRepNclust_prev[1:2])

   log(thetaofClust_prev[n,1,1])<-mu+att[team1_prev[n], season_prev[n]]+
                                  def[team2_prev[n], season_prev[n]]
   log(thetaofClust_prev[n,1,2])<-log(lambda1_book_prev[n])
   log(thetaofClust_prev[n,2,1])<-att[team2_prev[n], season_prev[n]]+
                                  def[team1_prev[n], season_prev[n]]
   log(thetaofClust_prev[n,2,2])<-log(lambda2_book_prev[n])

#priors for lambda1_book, lambda2_book

  lambda1_book_prev[n]~dnorm(theta1_bm_mean_prev[n], 0.01)T(0,)
  lambda2_book_prev[n]~dnorm(theta2_bm_mean_prev[n], 0.01)T(0,)

}

# Prior: MODEL FOR HYPERPARAMETERS
for (t in 1:nteams){

  att.star[t,1]~ dnorm(mu.att, tau.att)
  def.star[t,1]~ dnorm(mu.def, tau.def)

     att[t,1] <- att.star[t,1] - mean(att.star[,1])
     def[t,1] <- def.star[t,1] - mean(def.star[,1])

for (h in 2:T){

    att.star[t,h] ~ dnorm(mu.att+att.star[t,h-1],tau.att)
    def.star[t,h] ~ dnorm(mu.def+def.star[t,h-1],tau.def)

       att[t,h] <- att.star[t,h] - mean(att.star[,h])
       def[t,h] <- def.star[t,h] - mean(def.star[,h])
   }
```

```
 }

# priors on the random effects

mu.att ~ dnorm(0,0.0001)
mu.def ~ dnorm(0,0.0001)
tau.att ~ dgamma(.01,.01)
tau.def ~ dgamma(.01,.01)
mu~ dnorm(0,0.0001)
alpha~dunif(0,10)
beta~dunif(0,10)
}"
```

Bibliography

Albert, J. (1992), 'A Bayesian analysis of a poisson random effects model for home run hitters', *The American Statistician* **46**(4), 246–253.

Baio, G. and Blangiardo, M. (2010), 'Bayesian hierarchical model for the prediction of football results', *Journal of Applied Statistics* **37**(2), 253–264.

Becker, A. and Sun, X. A. (2016), 'An analytical approach for fantasy football draft and lineup management', *Journal of Quantitative Analysis in Sports* **12**(1), 17–30.

Berger, J. and Berliner, L. M. (1986), 'Robust Bayes and empirical Bayes analysis with ε-contaminated priors', *The Annals of Statistics* pp. 461–486.

Berger, J. et al. (2006), 'The case for objective Bayesian analysis', *Bayesian analysis* **1**(3), 385–402.

Bhattacharyya, A. (1946), 'On a measure of divergence between two multinomial populations', *Sankhyā: the indian journal of statistics* pp. 401–406.

Bonomo, F., Durán, G. and Marenco, J. (2014), 'Mathematical programming as a tool for virtual soccer coaches: a case study of a fantasy sport game', *International Transactions in Operational Research* **21**(3), 399–414.

Borovkov, A. and Moullagaliev, A. (1998), 'Mathematical statistics.', *Gordon Breach, Amsterdam* .

Cain, M., Law, D. and Peel, D. (2002), 'Is one price enough to value a state-contingent asset correctly? Evidence from a gambling market', *Applied Financial Economics* **12**(1), 33–38.

Cain, M., Law, D. and Peel, D. (2003), 'The favourite-longshot bias, bookmaker margins and insider trading in a variety of betting markets', *Bulletin of Economic Research* **55**(3), 263–273.

Carlin, B. P. and Louis, T. A. (2000), *Bayes and empirical Bayes methods for data analysis*, Vol. 17, Chapman & Hall/CRC Boca Raton, FL.

Cole, S. R., Chu, H. and Greenland, S. (2013), 'Maximum likelihood, profile likelihood, and penalized likelihood: a primer', *American journal of epidemiology* **179**(2), 252–260.

Darnieder, W. F. (2011), Bayesian methods for data-dependent priors, PhD thesis, The Ohio State University.

Dixon, M. J. and Coles, S. G. (1997), 'Modelling association football scores and inefficiencies in the football betting market', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **46**(2), 265–280.

Dixon, M. and Robinson, M. (1998), 'A birth process model for association football matches', *Journal of the Royal Statistical Society: Series D (The Statistician)* **47**(3), 523–538.

Efron, B. (2012), *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Vol. 1, Cambridge University Press.

Epstein, E. S. (1969), 'A scoring system for probability forecasts of ranked categories', *Journal of Applied Meteorology* **8**(6), 985–987.

Evans, M., Moshonov, H. et al. (2006), 'Checking for prior-data conflict', *Bayesian Analysis* **1**(4), 893–914.

Forrest, D., Goddard, J. and Simmons, R. (2005), 'Odds-setters as forecasters: The case of english football', *International journal of forecasting* **21**(3), 551–564.

Forrest, D. and Simmons, R. (2002), 'Outcome uncertainty and attendance demand in sport: the case of english soccer', *Journal of the Royal Statistical Society: Series D (The Statistician)* **51**(2), 229–241.

Garthwaite, P. H., Kadane, J. B. and O'Hagan, A. (2005), 'Statistical methods for eliciting probability distributions', *Journal of the American Statistical Association* **100**(470), 680–701.

Gelman, A. (2016*a*), 'Data-dependent prior as an approximation to hierarchical model'.
URL: *http://andrewgelman.com/2016/03/25/28321/*

Gelman, A. (2016*b*), 'Prior choice recommendations wiki !'.
URL: *http://andrewgelman.com/page/2/*

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013), *Bayesian Data Analysis*, third edn, Chapman & Hall/CRC.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2014), *Bayesian data analysis*, Vol. 2, Chapman & Hall/CRC Boca Raton, FL, USA.

Gelman, A. and Hill, J. (2006), *Data analysis using regression and multilevel/hierarchical models*, Cambridge University Press.

Gelman, A., Hwang, J. and Vehtari, A. (2014), 'Understanding predictive information criteria for Bayesian models', *Statistics and Computing* **24**(6), 997–1016.

Gelman, A., Jakulin, A., Pittau, M. G. and Su, Y.-S. (2008), 'A weakly informative default prior distribution for logistic and other regression models', *The Annals of Applied Statistics* pp. 1360–1383.

Gelman, A. and Shalizi, C. R. (2013), 'Philosophy and the practice of Bayesian statistics', *British Journal of Mathematical and Statistical Psychology* **66**(1), 8–38.

Gelman, A., Simpson, D. and Betancourt, M. (2017), 'The prior can generally only be understood in the context of the likelihood', *arXiv:1708.07487* .

Gelman, A. et al. (2006), 'Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)', *Bayesian analysis* **1**(3), 515–534.

Gottardo, R. and Raftery, A. E. (2008), 'Markov chain Monte Carlo with mixtures of mutually singular distributions', *Journal of Computational and Graphical Statistics* **17**(4), 949–975.

Groll, A. and Abedieh, J. (2013), 'Spain retains its title and sets a new record–generalized linear mixed models on European football championships', *Journal of Quantitative Analysis in Sports* **9**(1), 51–66.

Hastie, T., Tibshirani, R. and Friedman, J. (2002), 'The elements of statistical learning: data mining, inference, and prediction', *Biometrics* .

Jeffreys, H. (1998), *The theory of probability*, OUP Oxford.

Jullien, B., Salanié, B. et al. (1994), 'Measuring the incidence of insider trading: a comment on Shin', *Economic Journal* **104**(427), 1418–19.

Karlis, D. and Ntzoufras, I. (2000), 'On modelling soccer data', *Student* **3**(4), 229–244.

Karlis, D. and Ntzoufras, I. (2003), 'Analysis of sports data by using bivariate Poisson models', *Journal of the Royal Statistical Society: Series D (The Statistician)* **52**(3), 381–393.

Karlis, D. and Ntzoufras, I. (2009), 'Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference', *IMA Journal of Management Mathematics* **20**(2), 133–145.

Kohl, M., Ruckdeschel, P. and Kohl, M. M. (2007), 'The distrex package'.

Lomax, R. G. (2006), 'Fantasy sports: history, game types, and research', *Handbook of sports and media* pp. 383–392.

Londono, M. G. and Hassan, A. R. (2015), 'Sports betting odds: a source for empirical Bayes'.

Maher, M. J. (1982), 'Modelling association football scores', *Statistica Neerlandica* **36**(3), 109–118.

Miller, A. (2002), *Subset selection in regression*, CRC Press.

Morita, S., Thall, P. F. and Müller, P. (2008), 'Determining the effective sample size of a parametric prior', *Biometrics* **64**(2), 595–602.

Mutsvari, T., Tytgat, D. and Walley, R. (2016), 'Addressing potential prior-data conflict when using informative priors in proof-of-concept studies', *Pharmaceutical statistics* **15**(1), 28–36.

Ntzoufras, I. (2011), *Bayesian modeling using WinBUGS*, Vol. 698, John Wiley & Sons.

O'Hara, R. B., Sillanpää, M. J. et al. (2009), 'A review of Bayesian variable selection methods: what, how and which', *Bayesian analysis* **4**(1), 85–117.

Park, T. and Casella, G. (2008), 'The Bayesian Lasso', *Journal of the American Statistical Association* **103**(482), 681–686.

Petrone, S., Rizzelli, S., Rousseau, J. and Scricciolo, C. (2014), 'Empirical Bayes methods in classical and bayesian inference', *Metron* **72**(2), 201–215.

R Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Reimherr, M., Meng, X.-L. and Nicolae, D. L. (2014), 'Being an informed Bayesian: assessing prior informativeness and prior likelihood conflict', *arXiv preprint arXiv:1406.5958* .

Richardson, S. and Green, P. J. (1997), 'On Bayesian analysis of mixtures with an unknown number of components (with discussion)', *Journal of the Royal Statistical Society: series B (statistical methodology)* **59**(4), 731–792.

Rue, H. and Salvesen, O. (2000), 'Prediction and retrospective analysis of soccer matches in a league', *Journal of the Royal Statistical Society: Series D (The Statistician)* **49**(3), 399–418.

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D. and Neuenschwander, B. (2014), 'Robust meta-analytic-predictive priors in clinical trials with historical control information', *Biometrics* **70**(4), 1023–1032.

Shin, H. S. (1991), 'Optimal betting odds against insider traders', *The Economic Journal* **101**(408), 1179–1185.

Shin, H. S. (1993), 'Measuring the incidence of insider trading in a market for state-contingent claims', *The Economic Journal* **103**(420), 1141–1153.

Smith, M. A., Paton, D. and Williams, L. V. (2009), 'Do bookmakers possess superior skills to bettors in predicting outcomes?', *Journal of Economic Behavior & Organization* **71**(2), 539–549.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2003), *WinBUGS user manual*.

Stan Development Team (2016*a*), 'RStan: the R interface to Stan, version 2.9.0'. **URL:** *http://mc-stan.org*

Stan Development Team (2016*b*), 'The Stan C++ library, version 2.9.0'. **URL:** *http://mc-stan.org*

Stan Development Team (2016*c*), *Stan Modeling Language User's Guide and Reference Manual, Version 2.14.0*. `http://mc-stan.org/`.

Štrumbelj, E. (2014), 'On determining probability forecasts from betting odds', *International journal of forecasting* **30**(4), 934–943.

Thall, P. and Lee, S.-J. (2003), 'Practical model-based dose-finding in phase I clinical trials: methods based on toxicity', *International Journal of Gynecological Cancer* **13**(3), 251–261.

Thomas, A., Ventura, S. L., Jensen, S. T. and Ma, S. (2013), 'Competing process hazard function models for player ratings in ice hockey', *The Annals of Applied Statistics* pp. 1497–1524.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the Lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

Van der Vaart, A. W. (1998), *Asymptotic statistics*, Vol. 3, Cambridge university press.

Vehtari, A., Gelman, A. and Gabry, J. (2017), 'Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC', *Statistics and Computing* **27**(5), 1413–1432.

Wasserman, L. (2000), 'Asymptotic inference for mixture models by using data-dependent priors', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**(1), 159–180.

Wood, S. N. (2017), *Generalized additive models: an introduction with R*, CRC press.

Leonardo Egidi

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 049 827 4174
e-mail: egidi@stat.unipd.it; leoegidi@hotmail.it.

Current Position

Since November 2014; (expected completion: March 2018)
PhD Candidate in Statistical Sciences, admitted to the final exam.
University of Padova, Department of Statistical Sciences.
Thesis title: *Developments in Bayesian Hierarchical Models and Prior Specification with Application to Analysis of Soccer Data*
Supervisor: Prof. Nicola Torelli
Co-supervisor: Prof. Francesco Pauli

Research interests

- Relabelling in Bayesian mixture models .
- Hierarchical Bayesian models for football data.
- Data-dependent priors.

Education

September 2011 – March 2014
Master degree (*laurea magistrale***) in Statistical and Actuarial Sciences**.
University of Trieste, Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche "Bruno de Finetti" (DEAMS).
Title of dissertation: *A Comparison Between Bayesian Hierarchical Models for the Meta-Analysis of Data from Pre-Electoral Polls*
Supervisor: Prof. Nicola Torelli
Final mark: 110/110 cum laude

September 2008 – July 2011
Bachelor degree (*laurea triennale***) in Statistics, Informatics and Finance**.
University of Trieste, Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche "Bruno de Finetti" (DEAMS).
Title of dissertation: *Some Considerations on Arbitrage Principle*
Supervisor: Prof. Marco Zecchin
Final mark: 106/110

Visiting periods

January 2016-July 2016
Department of Statistics, Columbia University, New York, USA.
Supervisor: Prof. Andrew Gelman

Work experience

June 2014 – August 2014
Employer: Studio Attuariale Visintin & Associati
Position: Junior Actuary, Trieste

Computer skills

- R (profocient), WinBUGS (good), SAS (basic), Visual Basic (basic), Java (good), Excel (good), Stan (good).
- TEX , Microsoft Word.

Language skills

Italian (native); English (good); German (basic).

Publications

Articles in journals

Egidi L., Pappadà R., Pauli F. and Torelli N. (2017). Relabelling in Bayesian Mixture Models by Pivotal Units. *Statistics and Computing*,(), 1-13, DOI 10.1007/s11222-017-9774-2. (arXiv preprint arXiv:1501.05478v2)

Egidi L., Pappadà R., Pauli F. and Torelli N. (2017). Maxima units search (MUS) algorithm: methodology and applications. Accepted in *Studies in Theoretical and Applied Statistics, Springer Book*, 2018. (arXiv preprint arXiv:1611.01069).

Egidi L., Gabry J.S. (2017). Bayesian hierarchical models for predicting individual performance in football (soccer). *Journal of Quantitative Analysis of Sports*. **Accepted with minor revision.**

Not yet submitted to a journal

Egidi L., Pauli F. and Torelli N. (2017). Mixture Data-Dependent Priors. (arXiv preprint:1708.00099).

Conference proceedings

Egidi L., Pappadà R., Pauli F. and Torelli N. (2016). Relabelling in Bayesian Mixture Models by Pivotal Units - Una procedura di relabelling in modelli mistura Bayesiani basata su unità pivotali. Conference Paper, *SIS2016 Proceedings*. ISBN 9788861970618. [48th Scientific Meeting of the Italian Statistical Society, Salerno, 8-10 June 2016].

Egidi L., Gabry J.S. (2017). Bayesian hierarchical models for predicting individual performance in football (soccer). Conference Paper, *MathSport International 2017 Conference*. ISBN 978-88-6938-058-7. [Padova, 26-28 June 2017].

Conference presentations

Egidi L., Pappadà R., Pauli F. and Torelli N. (2016). Relabelling in Bayesian Mixture Models by Pivotal Units. (poster). *ISBA* (International Society for Bayesian Analysis), Santa Margherita di Pula, Italy, 13-17 June 2016.

Egidi L., Pappadà R., Pauli F. and Torelli N. (2016). Relabelling in Bayesian Mixture Models by Pivotal Units. (poster). *MBC2* (Workshop on Model-Based Clustering and Classification), Catania, Italy, 5-7 September 2017.

Egidi L., Pauli F. and Torelli N. (2017). A Hierarchical Bayesian Model for the Football Scores Using the Bookmakers Odds. (poster). *SISBAYES 2017*. Roma, Italy, 7-8 February 2017.

Egidi L., Pauli F. and Torelli N. (2017). Mixture data-dependent priors. (poster) *BISP 10* (Workshop on Bayesian Inference in Stochastic Processes), Milano, Italy, 13-15 June 2017.

Egidi L., Gabry J.S. (2017). Bayesian hierarchical models for predicting individual performance in football (soccer). (talk). *MathSport International 2017 Conference*. Padova, Italy, 26-28 June 2017.

Egidi L., Pauli F. and Torelli N. (2017). A Hierarchical Bayesian Model for the Football Scores Using the Bookmakers Odds. (invited talk). *AUEB Sports Analytics Workshop 2017*. Athens, Greece, 7-8 November 2017.

Teaching experience

27-29 September 2017
Introduction to Bayesian Data Analysis with Stan, 21 hours
Frontal lectures+labs, 21 hours. Joint Research Center (JRC), Sevilla. Instructor: Jonah Gabry (Department of Statistics, Columbia University).

Other Interests

Comic writing: publication of *Blogaritmi newyorkesi*, Leonardo Egidi. Battello Stampatore, 2017. ISBN-13: 9788887208696.

References

Prof. Nicola Torelli
Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche "Bruno de Finetti" (DEAMS), University of Trieste
Via Tigor 22, Trieste
Phone: 040 558 7032
e-mail: nicola.torelli@deams.units.it

Prof. Francesco Pauli
Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche "Bruno de Finetti" (DEAMS), University of Trieste
Via Tigor 22
Phone: 040 558 2518
e-mail: francesco.pauli@deams.units.it