Identification of Potential Risk Factors for Back Pain in Horses: An Analysis Using Additive Bayesian Networks

---

Master Thesis in Biostatistics (STA495)

by

Oliver John
15-732-118

supervised by

Prof. Dr. Reinhard Furrer
Dr. Marie Dittmann

Zurich, March 15, 2021

Identification of Potential Risk Factors for Back Pain in Horses: An Analysis Using Additive Bayesian Networks

Oliver John

Version March 15, 2021

Contents

Preface

This thesis has been a very interesting and fun project to me. It is hard to believe that time has passed so quickly, as it feels like I only started this project a week ago. I thoroughly enjoyed the statistical consulting project last semester and am very glad that I was able to work on a similar, applied project for my master's thesis. I'd especially like to thank Prof. Dr. Reinhard Furrer and Dr. Marie-Theres Dittmann for giving me the opportunity to independently decide the shape and form of this project, for their continued support and for their constructive feedback. I would also like to thank my PhD supervisor Michael Hediger for helping me when things got difficult. Also, a big thank-you to the entire UZH biostatistics team. Last but not least, I'd like to thank my family and friends. Thank you mom and dad for always being by my side and raising me to be the person I am today. I know that I would not have come this far if it wasn't for you. I would also like to thank the closest friends I have ever had in my life: Dori, Hinti and Ruben. Our nightly sessions have become sanctuary to me. I will appreciate our friendship forever. Thank you for being my best friends, always. The last person I would like to thank is the only remaining, very special and dear friend of mine, Elena. Thank you for your immense encouragement and for always being there for me.

Oliver John
March 2021

Chapter 1

Introduction

Back pain in horses is a common, serious impediment, which is caused by a plethora of factors and complex interplays thereof. Thus, it is difficult identifying and quantifying the effect of specific factors on back pain and on other factors of interest. The rider, saddle and the horse itself are in constant interplay, and may simultaneously contribute to the surge of back pain (Henson, 2013). It is therefore important to look at the factors involved in each part of the rider-horse system and to analyze the network leading to equine back pain in its entirety. The Equine Sports Medicine Unit of the Vetsuisse Faculty in Zurich Switzerland carried out the Swiss Equine Back Health Study from 2017 to 2018 (Dittmann *et al.*, 2020a). The study aimed at gathering and analyzing possible factors leading to equine back pain, resulting in nine data sets, which have been so far studied independently from each other. The data sets are comprised of a large rider-horse survey, data containing demographic information, data containing orthopedic examinations of the horse, three data sets containing information about the saddle and electronic saddle pressure measurements during riding, one data set based on algometric measures of the horse's back, one data set containing back examinations of the horse, and one data set containing physiotherapeutic examinations of the rider. It was identified, that a substantial proportion of horses were suffering from back pain and low-ranked lameness and had one or more inadequate saddle fit issues (Dittmann *et al.*, 2020a). There is yet much unexplored possible relations between variables from different data sets. In order to gain a deeper understanding of the complex system resulting in equine back pain, an analysis of the in-between influences of variables from these different data sets is of paramount importance.

Additive Bayesian network modelling allows for an intricate and intuitive view in the complex interplays between variables. This modelling approach is data driven and does not explicitly require expert opinion/prior information for designing the model, even though it may also be incorporated. It is usually accompanied by variable selection processes and by data imputation. The result of such a model is typically a directed acyclic graph. Each node in the directed acyclic graph is in nature a generalized linear model describing the node the arrow is pointing to. Such network models allow us to combine and examine the data sets at hand in a more holistic and systematic way. They account for not only the influence of explanatory variables on one or several outcomes (equine back pain in our case), but also for the influence of explanatory variables on each other, all while keeping the benefits of an intuitive interpretation of a generalized linear model (Kratzer *et al.*, 2019). Thus, additive Bayesian network modelling has the advantage of better representing the entirety of the rider-horse system reflected by all data sets. The aim of this thesis was to explore the influences of variables from the nine data sets on each other and on a horse's back pain score, posterior to adequate variable selection processes, variable ranking methods and data imputation, as well as comparing these results to a complete-case analysis. This thesis also contains a simulation analysis, which explores a simple, new process to obtain robust results, while also exploring the effect of random forest imputation and complete-case analyses on additive Bayesian networks.

Chapter 2

Methods

The entirety of all analyses were conducted using the `R` programming language (R Core Team, 2020) with base packages and the subsequent analysis-specific packages: `dplyr`, `plyr`, `purr`, `tidyverse`, `tidyr`, `knitr`, `kableExtra`, `ggplot2`, `varrank`, `missForest`, `abn`.

2.1 Study Details

The explorative back health survey study in sports medicine for horses lead from 2017 to 2018 is comprised of a total of nine data sets. Each study subject is comprised of a unique horse-rider pair. Horse owners/riders were invited to participate in the study and contacted using the national Swiss database for horse owners via post-office. Data acquisition was conditional on a signed declaration of consent from the part of the horse's owner/rider. Data acquisition was also conditional on the horse being ridden mainly by the registered owner, the owner being at least 18 years old, the horse being between 5 and 18 years old, the horse being ridden at least twice weekly, the horse not suffering under any sort of acute diseases and on the horse being used for either dressage, endurance, leisure, western riding, gait and/or jumping competitions. Additional publications on related studies are listed here: Dittmann *et al.* (2021), Dittmann *et al.* (2019), Dittmann *et al.* (2020b). The full study was divided in an online survey and several subsequent examination days. The quantification of equine back pain is difficult, as it can manifest itself through different reactions or behaviors, which are very individual. In this study, an attempt to quantify equine back pain was made through back palpitations by an experienced veterinarian. An overall back pain score was calculated based on the horse's reaction to being palpated. However, for the sake of simplicity and interpretability, a log-transformed pain score was used as an outcome of interest for this analysis. The individual data sets resulting from this study are listed here in more detail:

1. BH_Reiter_Pferd_DEF (RiderHorse): descriptive data about horse and rider height, weight, breed, etc... (237 obs. of 41 variables)
2. Survey_definitiv (Survey): survey questions regarding husbandry, feeding, equipment, training, etc..., filled out by the horse's owner (248 obs. of 562 variables)
3. BH_ORTHO (Ortho): orthopaedic examinations of the horse regarding lameness, movement asymmetry, limb problems, etc ..., carried out by two experienced veterinarians (237 obs. of 89 variables)
4. BH_Englischsattel_DEF_190220 (ESaddle): manual inspection of English saddles including brand, type and fit, carried out by a single experienced veterinarian (237 obs. of 78 variables)
5. BH_Westernsattel (WSaddle): manual inspection of Western saddles including brand, type and fit, different examination criteria/variables to English saddles, carried out by a single veterinarian (237 obs. of 26 variables)
6. BH_spm_max_pressure (SaddleP): different variables of electronically measured pressure beneath the saddle at different gaits while ridden by the owner (237 obs. of 18 variables)
7. BH_Algometrie (Algo): algometry measurements along different locations on the horse's back, potential additional measurements for equine back pain (237 obs. of 114 variables)
8. BH_Ruecken_Pferd (HorseBack): examination of the horse's back, assessing painfulness of muscles, ligaments and the spine through palpitation by two experienced veterinarians, potential additional measurements for equine back pain (237 obs. of 188 variables)
9. Physio_DEF (Physio): physiotherapeutic examination of the rider, assessing mobility, strength, endurance and reactivity (237 obs. of 112 variables)

2.2 Data Preparation and Description

All nine data sets were originally saved as an xlsx-file. To ease data import and data formatting, all nine data sets were first saved as a csv-file and then sequentially imported into `R`. Data preparation generally consisted in correctly formatting factors and numeric variables and renaming them. An initial expert-driven variable selection was also performed as an early step: from almost each data set, a selection of the most relevant variables were chosen to undergo further analysis. A series of few newly coded summary variables (a combination of other variables) were also created with the help of expert opinion. As a final step, the data sets were merged together with an outer join and sequentially trimmed to remove any missing values produced by the merging procedure. The data set to be analyzed ultimately contained 30 variables. Table 2.1 provides a short description of each selected variable for this analysis. Figure 2.1 and Figure 2.2 show the different histograms and proportions of variables for numeric and categorical variables, respectively.

Table 2.1: Variable selection and description.

	Data Set	Distribution	Summary Variable	Meaning	%NA
Horse Height	Survey	Gaussian	No	in centimeters	0
Breed Warm Blood	Survey	Binomial	No	binary indicator of breed status	0
Horse Age	Survey	Gaussian	No	in years	0
Pasture Score	Survey	Gaussian	No	quantification of frequency and duration of a horse's weekly meadow time	0
License	Survey	Binomial	No	rider license available	0
Main Usage	Survey	Multinomial	No	dressing (D), spare time (ST), military (M), jumping (J), other (O)	0
Saddle Rides	Survey	Poisson	No	weekly frequency with which the horse is ridden with a saddle	0
Training Freq	Survey	Binomial	No	weekly guided training frequency	0
Saddle Type	Survey	Binomial	No	Western saddle + other (WS/O), English saddle (ES)	0
Lambskin	Survey	Binomial	No	binary indicator of lambskin underneath the saddle	0
Number Symptoms	Survey	Poisson	No	number of total pain and lameness symptoms	0
Algo Measures	Algo	Gaussian	Yes	the mean of all mean algometric measures for each back location	4.8
Conformation	HorseBack	Binomial	No	binary indicator of conformation issues	4.8
Brach Hyper	HorseBack	Binomial	No	binary indicator of Brachialis muscular hypertrophy	4.8
Pain Rank log	HorseBack	Log-Gaussian	No	log-transformed sum of three distinct numeric pain ranked scores	6.5
Limited ROM Back	HorseBack	Binomial	Yes	combination of two distinct binary measures of limited range of movement	5.2
Limited ROM Ilium	HorseBack	Binomial	Yes	combination of two distinct binary measures of limited range of movement	5.6
Broken Hoof	Ortho	Binomial	Yes	binary indicator whether at least one hoof was broken towards the rear	4.4
Habit Patfix	Ortho	Binomial	Yes	binary indicator of habitual patellafixation	4.4
Tail Pos	Ortho	Binomial	No	a possible reflection of an underlying orthopedic issue	10.9
Lameness	Ortho	Binomial	No	binary indicator of lameness	5.2
Defensive Pull	Ortho	Binomial	No	binary indicator whether the horse pulled defensively during measurements	4.4
Mean ROM	Physio	Gaussian	Yes	mean of all range of movement measures of the rider	4.4
Mean Force	Physio	Gaussian	Yes	mean of all force measures of the rider	4.4
Mean Velo	Physio	Gaussian	Yes	mean of all velocity measures of the rider	4.4
Coord Rider	Physio	Poisson	No	number of successful coordination exercises of the rider	4.4
Dressed BW Rider	RiderHorse	Gaussian	No	dressed bodyweight of the rider	4.8
Saddle Pressure	SaddleP	Gaussian	Yes	mean of all saddle pressure measurements across all location for different gaits	4.8
Saddle Issues	SaddleE	Poisson	No	number of total saddle issues	12.5
Waist	SaddleE	Binomial	No	binary indicator of a narrow saddle waist cut	22.6

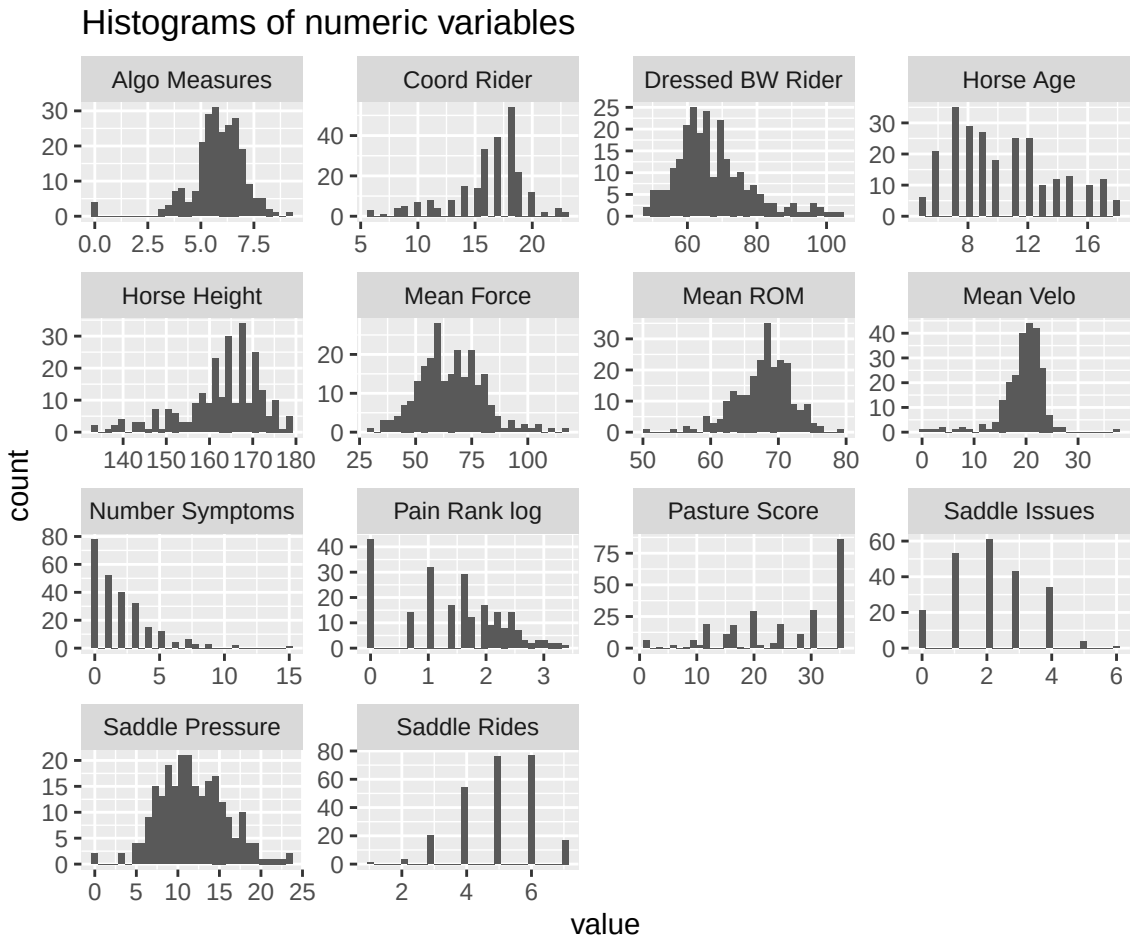

Figure 2.1: Histogram of numeric variables.

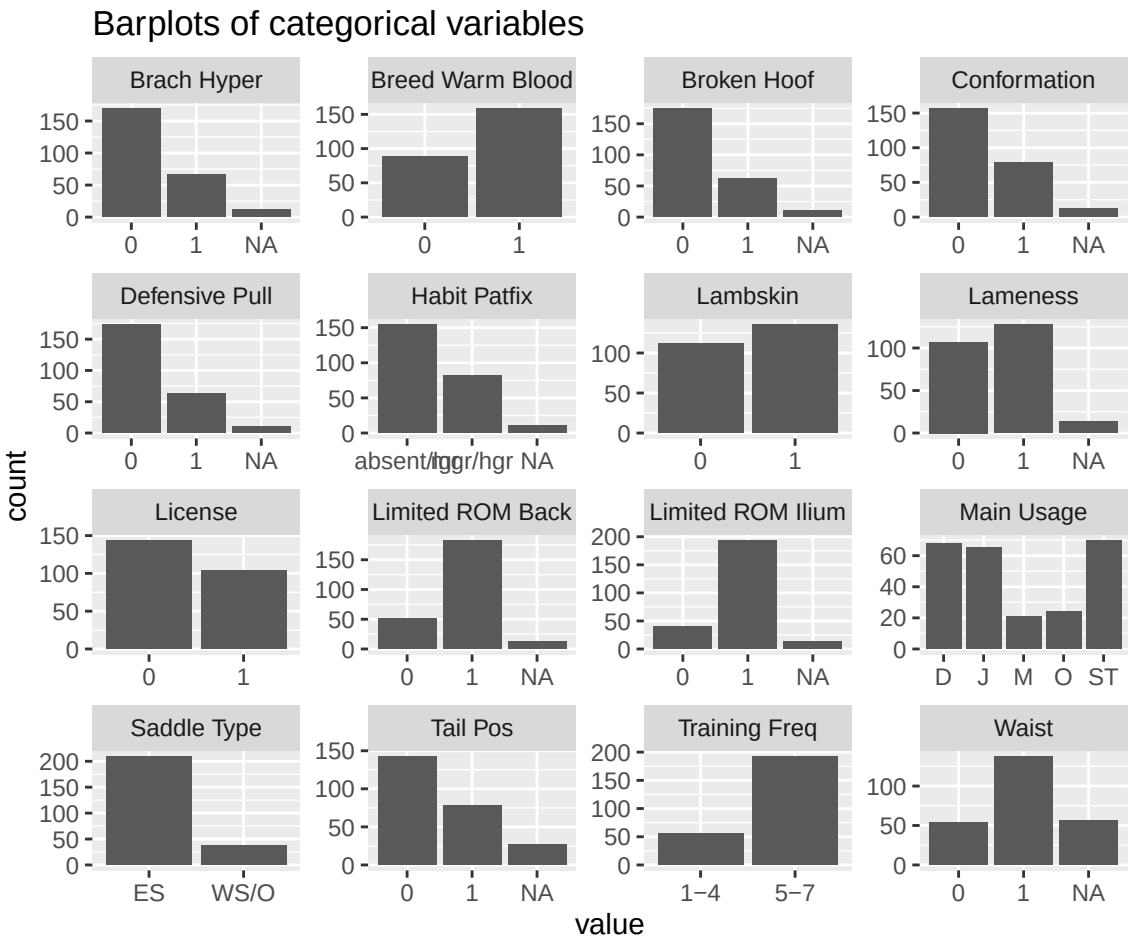

Figure 2.2: Barplots of categorical variables. A zero indicates no findings or issues.

2.3 Variable Ranking and Selection

Variable ranking is a useful method for reducing the dimensionality of a data set. It allows for a simple, yet still representable analysis/model of a system, whenever the addition of too many variables decreases the model's general predictive effectiveness. Furthermore, especially in terms of additive Bayesian network modelling, dimensionality reduction can greatly increase computational speed, interpretability and visualization of results. Variable ranking was performed in this analysis using the package `varrank`. The package uses filter-based methods to heuristically measure variable ranks using mutual information against a set of variables. In contrast to wrapper-based or embedded methods, filter-based methods compute variable ranks without the need of a pre-specified model, which means that the analysis is solely data-driven. The mutual information algorithm employed is particularly well-suited for a filter-based selection approach. It is a measure of minimum redundancy and maximum relevance (mRMRe) of a variable against a set of relevant variables of interest, which one desires to be in the final model. The mRMRe algorithm sequentially selects variables with the highest mutual information score possible. This score is penalized by the amount of redundant information, which is already present in preceding selected variables. The following Equation (2.1) quantifies the score of a variable to be selected (f_i) as an expression of the previously selected variables (f_s), a total set of variables ($\boldsymbol{F}$), a set of relevant variables ($\boldsymbol{C}$) and a set of selected variables from previous iterations ($\boldsymbol{S}$) (Kratzer and Furrer, 2018).

$$\underbrace{g(\alpha, \beta, \boldsymbol{C}, \boldsymbol{S}, f_i)}_{\text{Score of } f_i} = \overbrace{\text{MI}(f_i; \boldsymbol{C})}^{\text{Relevance}} - \sum_{f_s \in \boldsymbol{S}} \underbrace{\alpha(\beta, f_i, f_s, \boldsymbol{C}, \boldsymbol{S})}_{\text{Scaling factor}} \overbrace{\text{MI}(f_i; f_s)}^{\text{Redundancy}} \tag{2.1}$$

For this analysis, $\alpha(\beta, f_i, f_s, \boldsymbol{C}, \boldsymbol{S}) = 1/|\boldsymbol{S}|$ was set by defining `method="peng"` in the function `varrank()`. This method in itself has been named min-redundancy/max-relevance measure, even though `varrank` possesses several other methodologies for measuring redundancy/relevance (Kratzer and Furrer, 2018). A different formulation of Equation (2.1) is given by Equation (2.2). In this case, $\beta = 1/|\boldsymbol{S}|$ and $\alpha(f_i, f_s, \boldsymbol{C}, \boldsymbol{S}) = 1$ was set for the chosen methods for this analysis (Kratzer and Furrer, 2020).

$$g(\alpha, \beta, \boldsymbol{C}, \boldsymbol{S}, f_i) = \text{MI}(f_i; \boldsymbol{C}) - \beta \sum_{f_s \in \boldsymbol{S}} \text{MI}(f_i; f_s) \tag{2.2}$$

The mutual information MI is given by Equation (2.3). It is a formulation of the probability distributions of the random variables X and Y, where N and M are the respective number of levels within X and Y (Kratzer and Furrer, 2020). It is a measure of two categorical variables, which is why `varrank` uses discretization rules to categorize continuous variables. Sturges' discretization rule is the default of `varrank`, however, several other, as well as user-defined discretization, are possible. Relevance and redundancy measures are typically summarized and combined by either taking the difference (default) or the quotient of measurements (Kratzer and Furrer, 2018). For this and the remaining options to set in `varrank`, the default was chosen.

$$\text{MI}(X; Y) = \sum_{n=1}^{N} \sum_{m=1}^{M} P(x_n, y_m) \log\Big(\frac{P(x_n, y_m)}{P(x_n) P(y_m)}\Big) \tag{2.3}$$

Even after heavy initial variable selection with the help of an expert, the amount of variables remaining for an additive Bayesian network analysis remained too high. The resulting directed acyclic graph from an analysis incorporating all selected variables would have been too complex, difficult to interpret, time-consuming and computationally expensive. As a result, two analyses were performed: an additive Bayesian network analysis for high-ranking variables and one for low ranking variables. Out of the remaining 30 variables (including the outcome), rank/relevance of variables were determined with respect to the outcome *Pain Rank log*. The data set was split into two by sub-setting it to contain 16 high-ranking and 15 low-ranking variables, both including the outcome itself.

2.4 Random Forest Imputation

Data imputation is often a requirement for the analysis of large-scale data sets, since the dependency of an analysis frequently relies on algorithms that require a complete set of observations. However, many imputations methods often restrict themselves to either continuous of categorical variables, thus disregarding mixed-type data sets including both. Random forest imputation offers a pragmatic, nonparametric solution for the imputation of mixed-type data sets (Stekhofen, 2012).

Random forests are used a useful discrimination and classification tool, often applied in prediction. Each separate tree, which is a decision tree in nature, sequentially chooses random variables available in a data set as a predictor. Prediction is then based on the majority consensus of all trees present (Breinman, 2001). Thus, random forest imputation uses the majority consensus of all trees present in a forest to determine the missing value for each variable containing missing values. Random forest imputation provides the advantage of being robust against non-linear relations and complex interactions between variables, contrary to parametric imputation methods. As a result, random forest imputation is well-suited for whenever any normality assumptions might be violated. The package `missForest` performs a single imputation with the help of many random forest classifier trees. It trains a classifier forest using the observed cases and imputes missing values iteratively. `missForest` additionally supplies measures of accuracy regarding the imputation. The out-of-bag error rates are calculated for either the entire data set or for each variable separately. The classification error is defined as the mean squared error (MSE) for continuous variables (Equation (2.4)) and as the proportion of misclassified entries (PFC) for categorical variables (Stekhofen, 2012).

$$\text{MSE} = \frac{1}{n}\sum_{n=1}^{N}(X_{obs} - X_{pred})^2 \tag{2.4}$$

The total amount of missing values in the entire data set is 4.13% (3.93% and 4.49% for high- and low-ranking variable sets, respectively). In order to maximize the efficiency of random forest imputation, the imputation was performed on the entire data set of 30 variables, prior to splitting the data set according high- and low-ranking variables as described in Section 2.3. A seed was set to ensure reproducibility and 100 trees (default) were trained to impute missing values using `missForest`. Poisson-distributed variables run the chance of being imputed with values that are not integers. As a result, all Poisson-distributed variables were reverted to integers whenever this was the case. The out-of-bag error rates were calculated for each variable separately by setting `variablewise=TRUE` in `missForest()`.

2.5 Additive Bayesian Network Models

The package `abn` provides functions to create, select, modify, compare and plot additive Bayesian network models. In essence, the output of additive Bayesian network modelling, which is a data-driven analysis, is a network that is comprised of a collection of interconnected, easily interpretable generalized linear models. Such a network can help identifying and quantifying the many possible relationships between explanatory variables in a data set, while also allowing the inclusion of several outcome variables. The network itself takes the shape of a directed acyclic graph, comprised of nodes and interconnecting edges. Figure 2.3 shows a simple representation of such a graph. Each node represents a random variable in the data set, while the edges represent the relationship between nodes, i.e. the corresponding predictors. The distance between nodes does not reflect the importance or strength of a particular association. A directed acyclic graph is in itself a representation of the probabilistic model behind additive Bayesian network models. A general formulation of the probabilistic model is shown in Equation (2.5). It is based on a factorization of the probabilities of variables X_j, conditional on their parents $\boldsymbol{Pa}_j$; $j \in 1, \ldots, n$, where n is the number of variables in the data set (Kratzer *et al.*, 2020).

$$P(\boldsymbol{X}) = \prod_{j=1}^{n} P(X_j | \boldsymbol{Pa}_j) \tag{2.5}$$

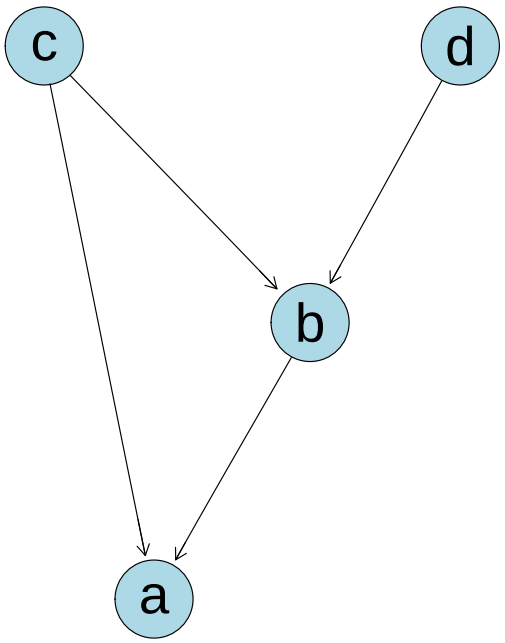

Figure 2.3: An example of a simple directed acyclic graph.

The entire process of creating an additive Bayesian network with `abn` can be divided into the following parts: a model learning, structure learning and a parameter learning part. Equation (2.6) describes the model in a mathematical notation. The model learning itself can be understood as the structure's distribution given the data set, multiplied by the parameter's distribution given both the data set and the model structure. The package `abn` uses score-based methods to select the best-fitting model from a vast selection of possible models. Once scores (marginal likelihoods) have been computed for each possible network, heuristic search algorithms then select the highest possible score, i.e. a network's score is representative of that network's capability of capturing the data's nature. The entire learning process may be solely data-driven, however, a semi-supervised setting is more often the case. This entails expert knowledge by adding some known or expected structure to the directed acyclic graph prior to the score-based model selection process. Parameter learning is performed as a last step, once the network's structure has been determined. The two approaches used for parameter estimation included in `abn` are either maximum likelihood or Bayesian methodologies. As a result, the regression coefficients are interpretable as in generalized linear regression models (Kratzer *et al.*, 2020).

$$\begin{aligned} \overbrace{P(\text{Model}|\text{Data})}^{\text{Model learning}} &= P(\text{Model Parameters}, \text{Structure}|\text{Data}) \\ &= \underbrace{P(\text{Model Parameters}|\text{Structure}, \text{Data})}_{\text{Parameter learning}} \underbrace{P(\text{Structure}|\text{Data})}_{\text{Structure learning}} \end{aligned} \tag{2.6}$$

An additive Bayesian network analysis including only complete cases ($n = 200$ and $n = 190$ for high- and low-ranking variables, respectively) was performed as a first step. An analysis with `abn` utilizes the following `R` commands: `buildscorecache()`, `mostprobable()` and `fitabn()`. The functions create a cache of scores for each possible network configuration, search the computed scores for its optimal value and fit an additive Bayesian network model, respectively. The maximum likelihood approach was used for network score and parameter estimation by setting `method="mle"` in both `buildscorecache()` and `fitabn()`. In order to reduce computation time and increase interpretability, high-ranking variables were analyzed separately from low-ranking variables. Variables were abbreviated to enhance the network's visualization. Table 2.2 and Table 2.3 provide a legend of both high- and low-ranking variable names and their corresponding abbreviation, respectively. Variable distributions were set as described in Table 2.1.

In order to warrant the biological and medical interpretability of the models, a series of banned arcs/relationships was set up. The interplay between high-ranking variables is more complicated than between low-ranking variables, which is why a more intricate banned arc matrix was created. Variable interactions with each other were carefully examined and restricted whenever needed. This was done with the help of expert opinion, thus taking the nature and context of each experiment and measure into account. The resulting interaction-restricting matrix between variables was nevertheless constructed to be as permissive as possible. Table 2.4 is a representation of the restrictions placed upon the possible relations of the network model of high-ranking variables. Each column represents a parent, while each row represents a child. For example, the first row is interpreted as follows: variable *Pain* cannot have itself, variable *V2*, *V4* and *V6* as a parent. Conversely, the fist column is interpreted as follows: variable *Pain* cannot have itself, variable *V2*, *V3*, *V4*, *V5*, *V6*, *V7*, *V9*, *V11*, *V12*, *V13* and *V15* as children. In contrast, low-ranking variables needed only few prior specifications regarding the network's structure: variables *V1*, *V4*, *V5*, *V6*, *V11*, *V13*, *V14* were set as variables without any parents. A variable not having any parent is equivalent to that variable's row including only "X".

As a next step, the maximum number of parent nodes was investigated for each variable set. This was done by penalizing model complexity by scoring the network using the Bayesian information criterion, while sequentially increasing the maximum number of parents and fitting an additive Bayesian network model. Upon finding the right amount of maximum parent nodes for each network, a model was created for each variable set. The exact same procedure was repeated for the variable sets with random-forest-imputed values ($n = 248$). To underline the variables specifically influencing equine back pain, the Markov blanket variables of *Pain Rank log* were specifically colored. The Markov blanket is the set of variables that are needed to fully describe a target variable, i.e. the Markov blanket contains the necessary information to fully perform inference on the target variable. In other words, the Markov blanket of a target variable is comprised of its parents, children and those children parents (Kratzer *et al.*, 2019).

Table 2.2: High-ranking variables abbreviation legend.

Variable	Abbreviation
Pain Rank log	Pain
Algo Measures	V1
Coord Rider	V2
Saddle Type	V3
Mean Velo	V4
Horse Height	V5
Mean ROM	V6
Pasture Score	V7
Number Symptoms	V8
Horse Age	V9
Tail Pos	V10
Mean Force	V11
Saddle Rides	V12
Saddle Issues	V13
Limited ROM Ilium	V14
Saddle Pressure	V15

Table 2.3: Low-ranking variables abbreviation legend.

Variable	Abbreviation
Pain Rank log	Pain
Breed Warm Blood	V1
Defensive Pull	V2
Hyper Brach	V3
Lambskin	V4
Dressed BW Rider	V5
License	V6
Habit Patfix	V7
Conformation	V8
Broken Hoof	V9
Lameness	V10
Main Usage	V11
Limited ROM Back	V12
Training Freq	V13
Waist	V14

Table 2.4: Banned arcs matrix representation.

	Pain	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
Pain	X	O	X	O	X	O	X	O	O	O	O	O	O	O	O	O
V1	O	X	X	O	X	O	X	O	O	O	O	O	O	O	O	O
V2	X	X	X	O	O	X	O	X	X	X	X	O	O	X	X	X
V3	X	X	O	X	O	O	O	X	X	O	X	O	O	X	X	X
V4	X	X	O	O	X	X	O	X	X	X	X	O	O	X	X	X
V5	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
V6	X	X	O	O	O	X	X	X	X	X	X	O	O	X	X	X
V7	X	X	X	X	X	X	X	X	O	O	X	X	O	X	X	X
V8	O	O	X	O	X	O	X	O	X	O	O	O	O	O	O	O
V9	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
V10	O	O	X	O	X	O	X	X	O	O	X	O	O	O	O	O
V11	X	X	O	O	O	X	O	X	X	X	X	X	O	X	X	X
V12	X	X	O	O	O	X	O	X	O	O	X	O	X	O	O	X
V13	X	X	X	O	X	O	X	X	X	X	X	X	O	X	X	X
V14	O	O	O	O	O	O	O	O	O	O	O	O	O	O	X	O
V15	X	X	O	O	O	O	O	O	O	X	X	O	O	O	O	X

2.6 Additive Bayesian Network Models: Robustness Analysis

Additive Bayesian network modelling is often accompanied by overfitting, resulting in an overly complex network and biased coefficient estimates (Kratzer *et al.*, 2019). In order to control for overfitting and to explore the impact of random forest imputation on the entire network in a thorough manner, a simulation in which missing values (5% missing values entries, same magnitude as in the original data set) are introduced into the complete-case data set and sequentially imputed with random forests was conducted. In order to improve computational efficiency, this simulation/robustness analysis was conducted with the eleven highest ranking variables. This analysis thus simulates the complete-case data set ($n = 221$) as the "true" data set and compares the resulting model with an additive Bayesian network model analysis after random forest imputation ($n = 221$) and after an analysis using only complete cases of the missing-value-imposed "true" data set ($n = 120$, essentially a complete-case analysis of the complete cases after introducing missing values). Without accounting for overfitting, some difference in both structure and coefficients between the three resulting models would be expected. In order to ensure robust results, a simple and pragmatic resampling method was conceptualized. For each of the aforementioned data sets, 50% was resampled to create a network model. This process was repeated 100 times. Robustness was then achieved by constructing a model consisting of only associations, which were present in at least 50% of all resampling iterations, e.g. a pruned additive Bayesian network.

Chapter 3

Results

3.1 Varrank Results

Figure 3.1 shows the results of variable ranking using `varrank` in a decreasing order. We observe that most Gaussian- and Poisson-distributed variables are highly ranked in relevance with respect to the outcome *Pain Rank log*. Both the matrix and the score densities show a rather symmetrical distribution of scores across variables, albeit highly ranked variables being the minority. A clear amount of high redundancy is evident in the lower-ranked variables.

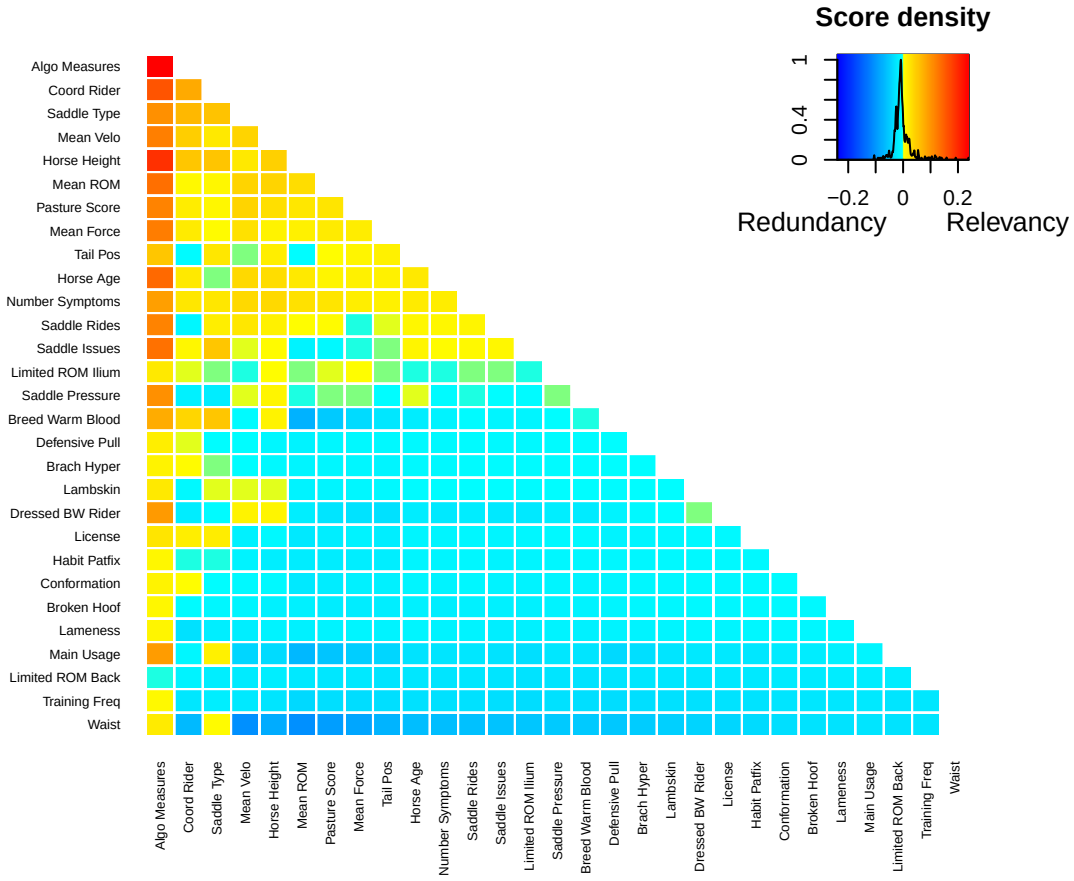

Figure 3.1: Varrank score matrix and density.

3.2 Random Forest Imputation Results

Table 3.1 lists the out-of-bag error rates for each variable of the random forest decision trees used to impute missing data. The mean squared error is particularly high in variables *Mean ROM*, *Mean Force*, *Mean Velo*, *Coord Rider*, *Dressed BW Rider* and *Saddle Pressure*, while it remains reasonably low and rather constant for the remaining continuous variables. On the other hand, the proportion of misclassified categorical variables is rather constant for the most, albeit quite high in certain variables, such as *Lameness*, *Tail Pos* and *Habit Patfix*.

Table 3.1: Mean squared error (MSE) and misclassified proportion (PFC) of variables from the random forest used to impute missing values.

Variable	Error Measure	Error Measure Type	%NA
Horse Height	0	MSE	0
Breed Warm Blood	0	PFC	0
Horse Age	0	MSE	0
Pasture Score	0	MSE	0
License	0	PFC	0
Main Usage	0	PFC	0
Saddle Rides	0	MSE	0
Training Freq	0	PFC	0
Saddle Type	0	PFC	0
Lambskin	0	PFC	0
Number Symptoms	0	MSE	0
Algo Measures	1.69	MSE	4.84
Conformation	0.36	PFC	4.84
Brach Hyper	0.31	PFC	4.84
Pain Rank log	0.7	MSE	6.45
Limited ROM Back	0.27	PFC	5.24
Limited ROM Ilium	0.18	PFC	5.65
Broken Hoof	0.31	PFC	4.44
Habit Patfix	0.41	PFC	4.44
Tail Pos	0.42	PFC	10.89
Lameness	0.48	PFC	5.24
Defensive Pull	0.29	PFC	4.44
Mean ROM	16.53	MSE	4.44
Mean Force	156.74	MSE	4.44
Mean Velo	16.78	MSE	4.44
Coord Rider	10.41	MSE	4.44
Dressed BW Rider	98.62	MSE	4.84
Saddle Pressure	15.6	MSE	4.84
Saddle Issues	1.76	MSE	12.5
Waist	0.34	PFC	22.58

3.3 Complete Case Analyses Results

The results stemming from the complete-case analysis of maximum number of parent nodes in each network are shown in Figure 3.2 and Figure 3.3 for the sets of high- and low-ranking variables, respectively. For high-ranking variables, we observe a nick in the log marginal likelihood at three maximum number of parent nodes, after which the likelihood remains constant. Subsequently, the maximum number of parents for high-ranking variables was set to four. In contrast, the log marginal likelihood of low-ranking variables caps at two maximum number of parent nodes and then remains constant, which is why the maximum number of parent nodes was set to three. These modelling choices ensure a complex enough network to capture potentially interesting associations in-between variables.

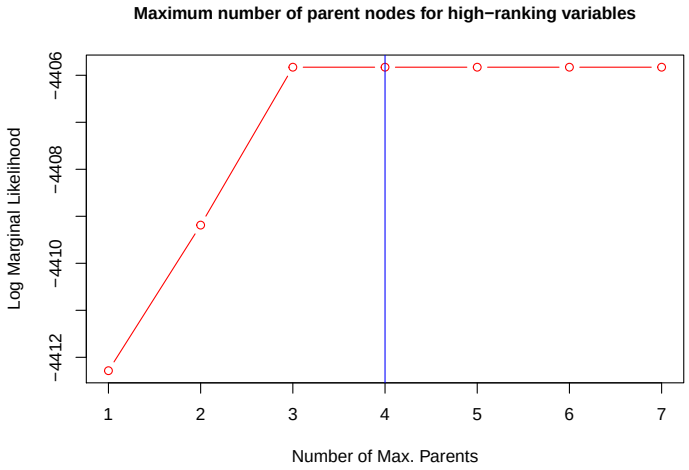

Figure 3.2: Log marginal likelihood of the BIC-scored networks of high-ranking variables with different number of maximum parent nodes (complete-case analysis).

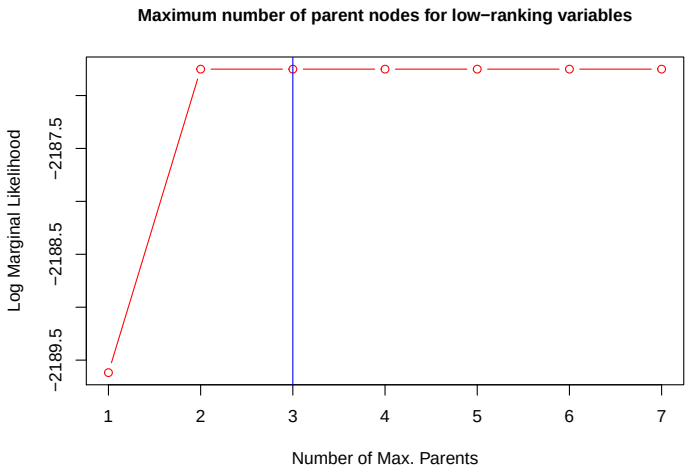

Figure 3.3: Log marginal likelihood of the BIC-scored networks of low-ranking variables with different number of maximum parent nodes (complete-case analysis).

The following texts describing the additive Bayesian networks are by no means an exhaustive description of all associations found in the network. It is merely an emphasis of particular relationships, that intuitively and well-describe the rider-horse system at hand. The coefficients depicted in figures are interpreted as follows: a one unit change in a variable/changing the category of a variable from the reference level to another causes a change in mean, mean log, log odds ratio or log rate ratios, in magnitude of the value of the variable's coefficient on the target variable, while holding all other explanatory variables constant.

The additive Bayesian network of the complete cases of high-ranking variables is depicted in Figure 3.4 as a directed acyclic graph showing parent-child relationships. We observe that variable *V1/Algo Measures*, *V3/Saddle Type*, *V7/Pasture Score* and *V10/Tail Pos* are all associated with a direct influence on *Pain/Pain Rank log*. The Markov blanket additionally includes variables *V6/Mean ROM*, *V8/Number Symptoms*, *V13/Saddle Issues*, *V14/Limited ROM Ilium* and *V15/Saddle Pressure*. All variables with a direct influence on *Pain/Pain Rank log* seem to be associated with it negatively. Variable *V12/Saddle Rides* and *V5/Horse Height* are both negatively associated with *V1/Algo Measures*, while *V11/Mean Force* and *V15/Saddle Pressure* are associated with a positive influence. *V3/Saddle Type* is associated negatively with *V5/Horse Height*, *V9/Horse Age* and *V11/Mean Force*, and positively with *V4/Mean Velo*. *V5/Horse Height* and *V11/Mean Force* are both positively associated with a positive influence on *V10/Tail Pos*, while *V9/Horse Age* and *V14/Limited ROM Ilium* are associated with a negative influence on it. Additionally, *V7/Pasture Score* is associated with a positive and negative influence by *V9/Horse Age* and *V12/Saddle Rides*, respectively.

In particular, we find many other interesting associations, such as positive associations between variables *V3/Saddle Type* and *V15/Saddle Pressure*, *V5/Horse Height* and *V8/Number Symptoms*, *V6/Mean ROM* and *V14/Limited ROM Ilium*, *V9/Horse Age* and *V7/Pasture Score*, *V12/Under Saddle Per Week* and *V13/Saddle Issues*, *V12/Under Saddle Per Week* and *V8/Number Symptoms*, *V11/Mean Force* and *V15/Saddle Pressure*, and *V15/Saddle Pressure* and *V1/Algo Measures*. We also find negative association between variables *V3/Saddle Type* and *V13/Saddle Issues*, *V5/Horse Height* and *V3/Saddle Type*, *V5/Horse Height* and *V13/Saddle Issues*, *V9/Horse Age* and *V3/Saddle Type*, *V9/Horse Age* and *V12/Under Saddle Per Week*, and *V7/Pasture Score* and *V14/Limited ROM Ilium*. Table 3.2 provides basic information measures describing the general network model.

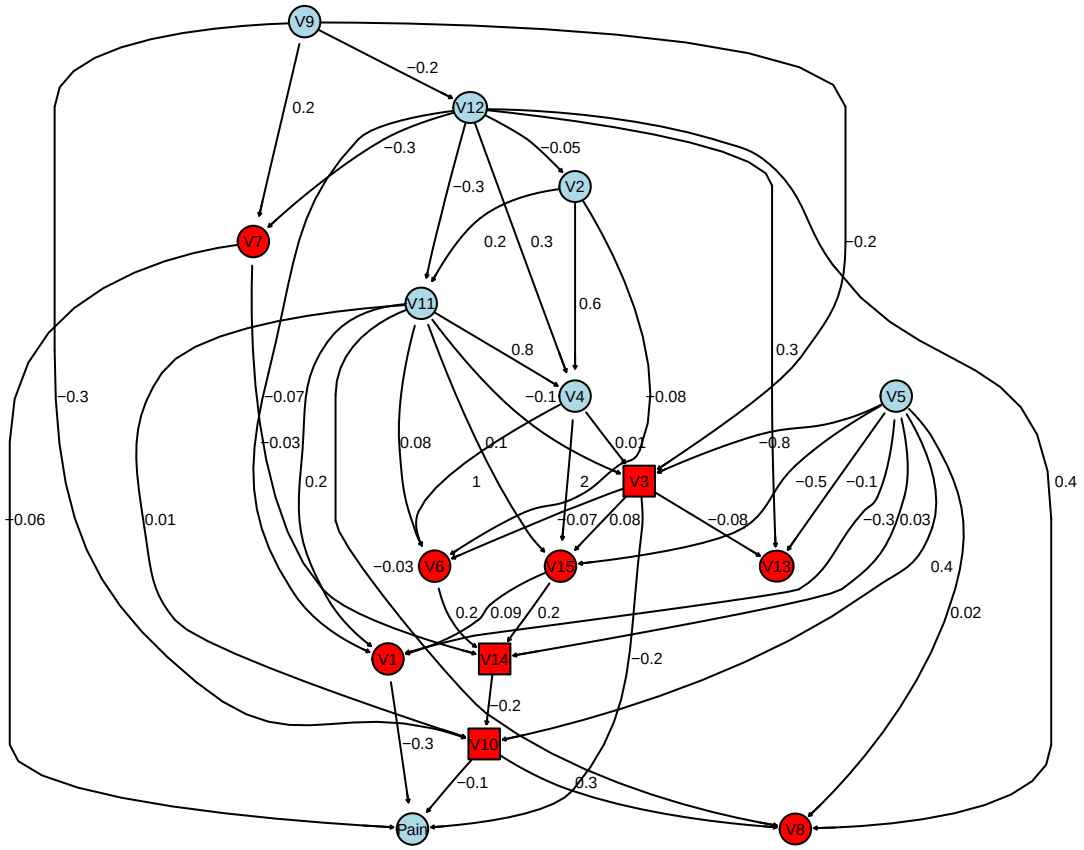

Figure 3.4: Additive Bayesian network directed acyclic graph of high-ranking variables, showing parent-child relationships (complete-case analysis). Variables in circles, ovals and rectangles represent Gaussian, Poisson and binomial variables, respectively. Variable coefficients are shown next to their respective edges (rounded to first significant digit). Markov blanket of *Pain* shown in red.

Table 3.2: General information about the complete-case analysis directed acyclic graph of high-ranking variables.

Network Measure	Value
Number of Nodes	16
Number of Arcs	44
Markov Blanket Average Set Size	10.88
Neighbour Average Set size	5.5
Parent Average Set Size	2.75
Children Average Set Size	2.75

Figure 3.5 is a similar representation of Figure 3.4. It shows the same relationships between variables, but reversed as child-parent relationships. As a result, the variable coefficients are reversed in terms of direction.

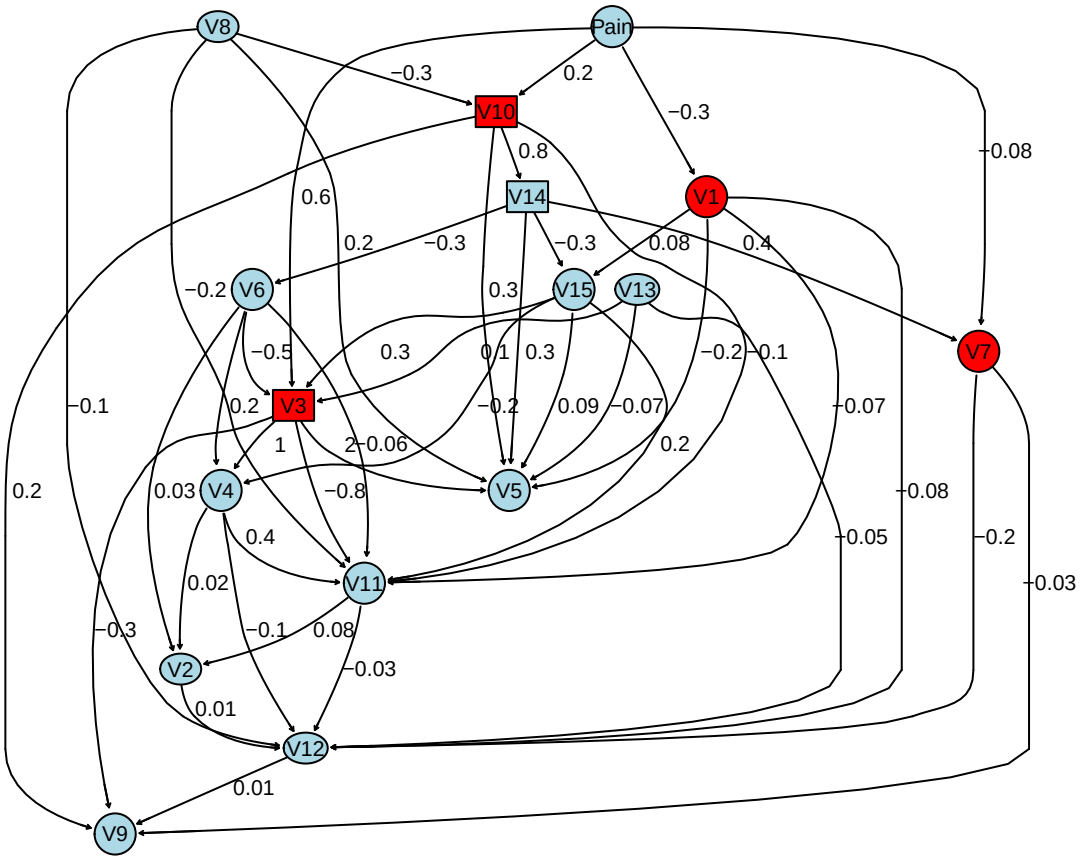

Figure 3.5: Additive Bayesian network directed acyclic graph of high-ranking variables, showing child-parent relationships (complete-case analysis). Variables in circles, ovals and rectangles represent Gaussian, Poisson and binomial variables, respectively. Variable coefficients are shown next to their respective edges (rounded to first significant digit). Markov blanket of *Pain* shown in red.

The additive Bayesian network of the complete-cases of low-ranking variables are depicted as directed acyclic graphs in Figure 3.6 (parent-child relations). We observe variables *V1/Breed Warm Blood*, *V3/Hyper Brach* and *V11/Main Usage* being associated with having a direct influence on *Pain/Pain Rank log*. The Markov blanket of *Pain/Pain Rank log* additionally includes variable *V2/Defensive Pull*, *V7/Habit Patfix*, *V8/Conformation* and *V9/Broken Hoof*. *V1/Breed Warm Blood* and *V11/Main Usage* seem to be associated with a positive influence on *Pain/Pain Rank log*, while *V3/Hyper Brach* seems to be associated with a negative influence. Both variables *V1/Breed Warm Blood* and *V11/Main Usage* are associated with a negative influence on *V2/Defensive Pull*, *V3/Hyper Brach* and *V8/Conformation*. *V1/Breed Warm Blood* and *V11/Main Usage* are additionally negatively associated with *V7/Habit Patfix* and *V9/Broken Hoof*, respectively. We find three variables not being associated with any other variable in the network (*V4/Lambskin*, *V13/Training Freq* and *V14/Waist*). Additionally, we find negative associations between variables *V3/Hyper Brach* and *V8/Conformation*, *V5/Dressed BW Rider* and *V9/Broken Hoof*, *V8/Conformation* and *V9/Broken Hoof*, and *V10/Lameness* and *V7/Habit Patfix*. Table 3.3 lists the general network measures.

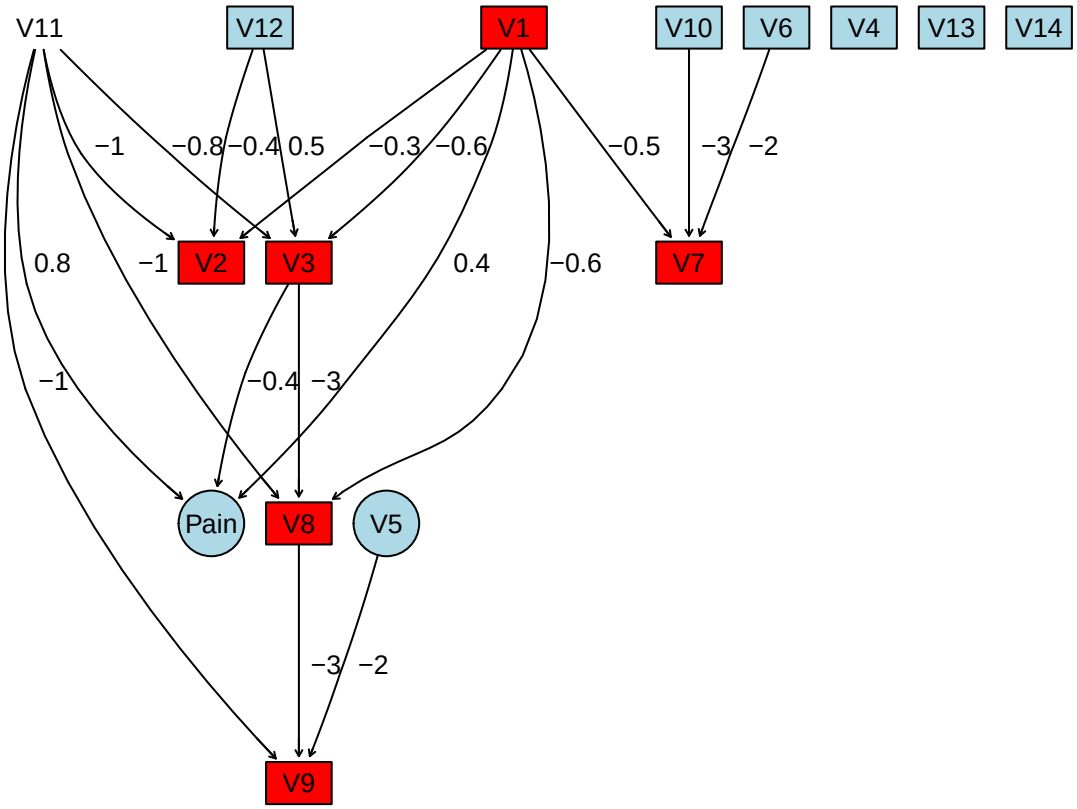

Figure 3.6: Additive Bayesian network directed acyclic graph of low-ranking variables, showing parent-child relationships (complete-case analysis). Variables in circles, ovals, rectangles and no enclosure represent Gaussian, Poisson, binomial and multinomial variables, respectively. Variable coefficients are shown next to their respective edges (rounded to first significant digit). Markov blanket of *Pain* shown in red ($+V11$).

Table 3.3: General information about the complete-case analysis directed acyclic graph of low-ranking variables.

Network Measure	Value
Number of Nodes	15
Number of Arcs	18
Markov Blanket Average Set Size	3.87
Neighbour Average Set size	2.4
Parent Average Set Size	1.2
Children Average Set Size	1.2

Figure 3.7 is the child-parent representation of the graph shown in Figure 3.6. It shows the same relationships between variables, but reversed as child-parent relationships. As a result, the variable coefficients are reversed as well.

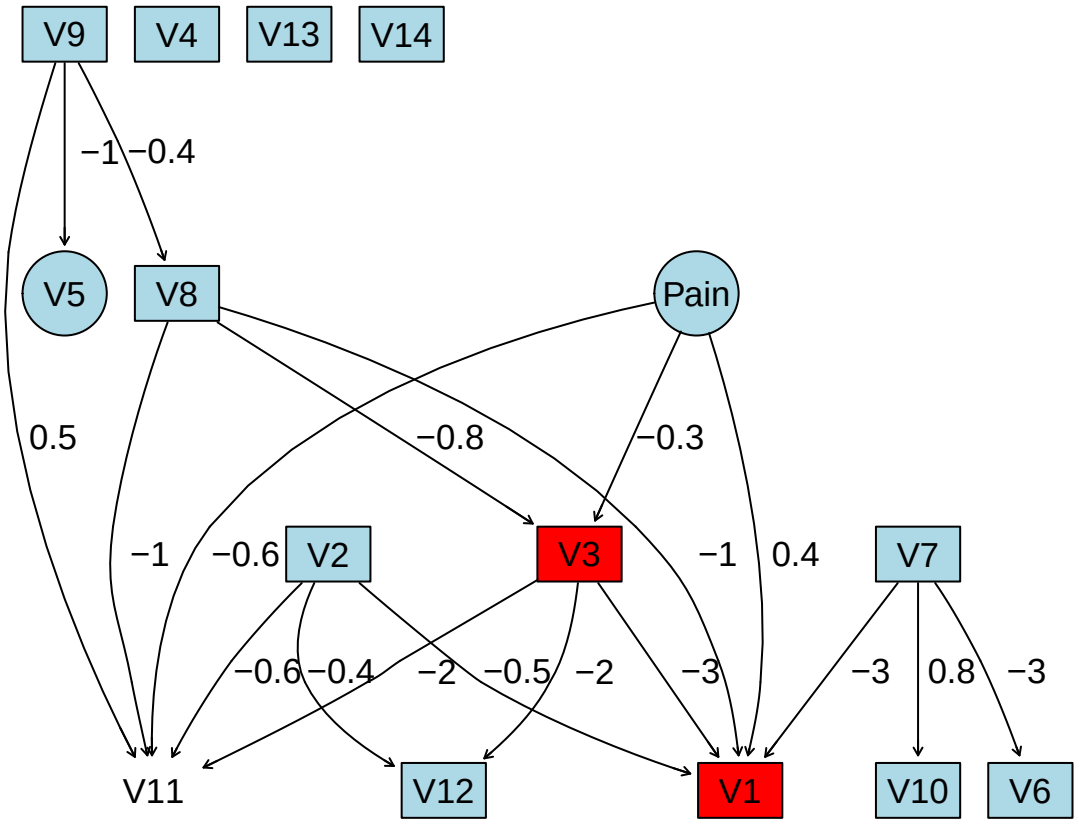

Figure 3.7: Additive Bayesian network directed acyclic graph of low-ranking variables, showing child-parent relationships (complete-case analysis). Variables in circles, ovals, rectangles and no enclosure represent Gaussian, Poisson, binomial and multinomial variables, respectively. Variable coefficients are shown next to their respective edges (rounded to first significant digit). Markov blanket of *Pain* shown in red ($+V11$).

3.4 Imputed Data Analyses Results

The results stemming from the imputed data set analysis of maximum number of parent nodes in each network are shown in Figure 3.8 and Figure 3.9 for the sets of high- and low-ranking variables, respectively. Correspondingly, the log marginal likelihood of high- and low-ranking variables caps at three and two maximum number of parent nodes and then remains constant. As with the complete-case analysis, as well as in order to ensure a complex enough network and comparability, the maximum number of parent nodes was set to four in the analysis of high-ranking variables and three in the analysis of low-ranking variables.

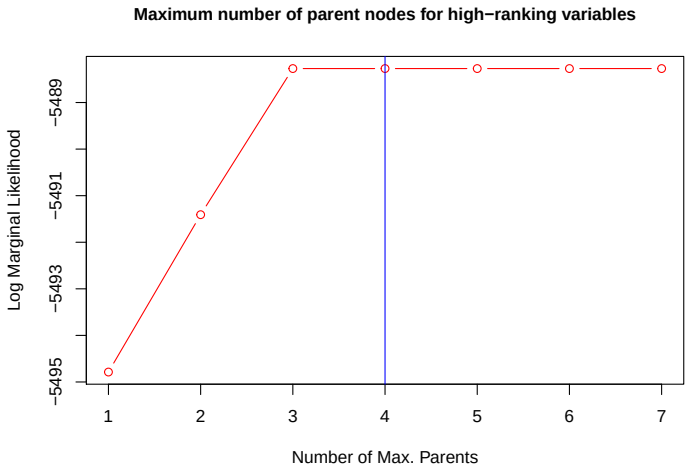

Figure 3.8: Log marginal likelihood of the BIC-scored networks of high-ranking variables with different number of maximum parent nodes (imputed data analysis).

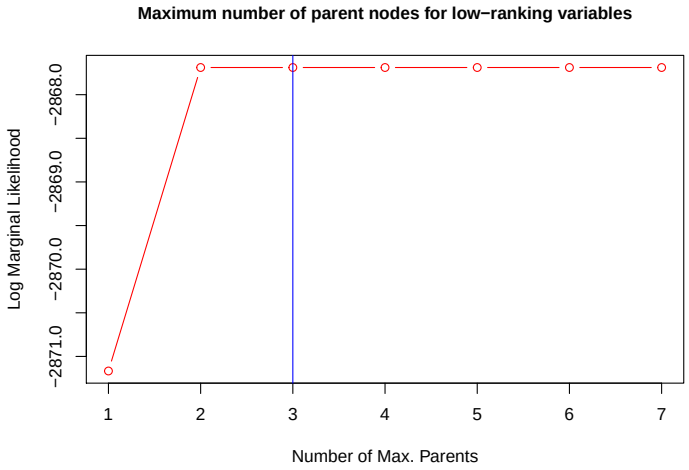

Figure 3.9: Log marginal likelihood of the BIC-scored networks of low-ranking variables with different number of maximum parent nodes (imputed data analysis).

The additive Bayesian networks of high-ranking variables stemming from an analysis of the imputed data set is shown in Figure 3.10 (parent-child relations). In comparison to the complete-case analysis, we now find a direct positive association between *V5/Horse Height*, *V7/Pasture Score* and *V11/Mean Force* with *Pain/Pain Rank log*, while *V3/Saddle Type* remains negatively associated with it. The Markov blanket is now comprised of all other variables present in the network, with the exception of *V9/Horse Age*. *V5/Horse Height* is a variable/node without any parents, while *V7/Pasture Score* is associated with a positive influence by variables *V12/Saddle Rides* and *V9/Horse Age*. *V12/Saddle Rides* and *V2/Coord Rider* are associated negatively with *V11/Mean Force*, while *V3/Saddle Type* is positively associated with it. *V3/Saddle Type* is associated with being influenced negatively by *V5/Horse Height* and positively by *V9/Horse Age*.

In comparison to the complete-case analysis, aforementioned associations between variables *V5/Horse Height* and *V3/Saddle Type*, *V5/Horse Height* and *V8/Number Symptoms*, *V5/Horse Height* and *V13/Saddle Issues*, *V9/Horse Age* and *V7/Pasture Score*, *V12/Saddle Rides* and *V13/Saddle Issues*, and *V11/Mean Force* and *V15/Saddle Pressure*, *V15/Saddle Pressure* and *V1/Algo Measures* all remained present. On the other hand, the network of high-ranking variables in the imputed data analysis did change to a certain degree. We find aforementioned associations in the complete-case analysis results between *V2/Coord Rider* and *V11/Mean Force*, *V3/Saddle Type* and *V13/Saddle Issues*, *V5/Horse Height* and *V1/Algo Measures*, *V5/Horse Height* and *V10/Tail Pos*, *V7/Pasture Score* and *V14/Limited ROM Ilium*, *V9/Horse Age* and *V3/Saddle Type*, *V9/Horse Age* and *V10/Tail Pos*, *V9/Horse Age* and *V12/Saddle Rides*, and *V12/Saddle Rides* and *V7/Pasture Score* all remaining present, but changing in coefficient sign. Furthermore, we find associations between *V6/Mean ROM* and *V14/Limited ROM Ilium*, *V11/Mean Force* and *V1/Algo Measures*, *V11/Mean Force* and *V10/Tail Pos*, *V12/Saddle Rides* and *V1/Algo Measures*, *V12/Saddle Rides* and *V8/Number Symptoms*, and *V14/Limited ROM Ilium* and *V10/Tail Pos* all being broken. We also find the in-between associations of *Pain/Pain Rank log* with *V1/Algo Measures* and *V10/Tail Pos* becoming reversed in comparison to the complete-case analysis. We observe from Table 3.4 that only the number of nodes remained unchanged in comparison to the complete-case analysis. The number of arcs, average Markov blanket set size, average neighbour set size, average parent set size and average children set size all increased.

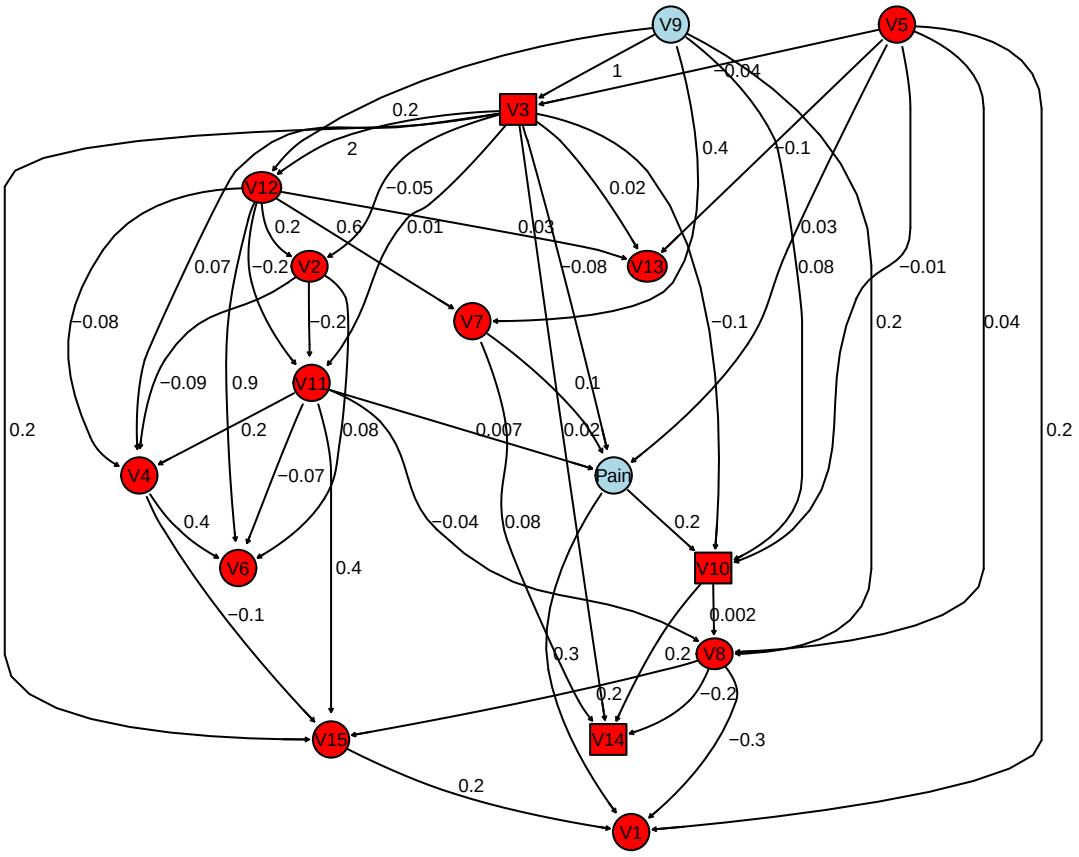

Figure 3.10: Additive Bayesian network directed acyclic graph of high-ranking variables, showing parent-child relationships (imputed data analysis). Variables in circles, ovals and rectangles represent Gaussian, Poisson and binomial variables, respectively. Variable coefficients are shown next to their respective edges (rounded to first significant digit). Markov blanket of *Pain* shown in red.

Table 3.4: General information about the imputed data analysis directed acyclic graph of high-ranking variables.

Network Measure	Value
Number of Nodes	16
Number of Arcs	46
Markov Blanket Average Set Size	11.38
Neighbour Average Set size	5.75
Parent Average Set Size	2.88
Children Average Set Size	2.88

Figure 3.11 is the child-parent representation of the graph shown in Figure 3.10. As a result, relationships and coefficients are reversed.

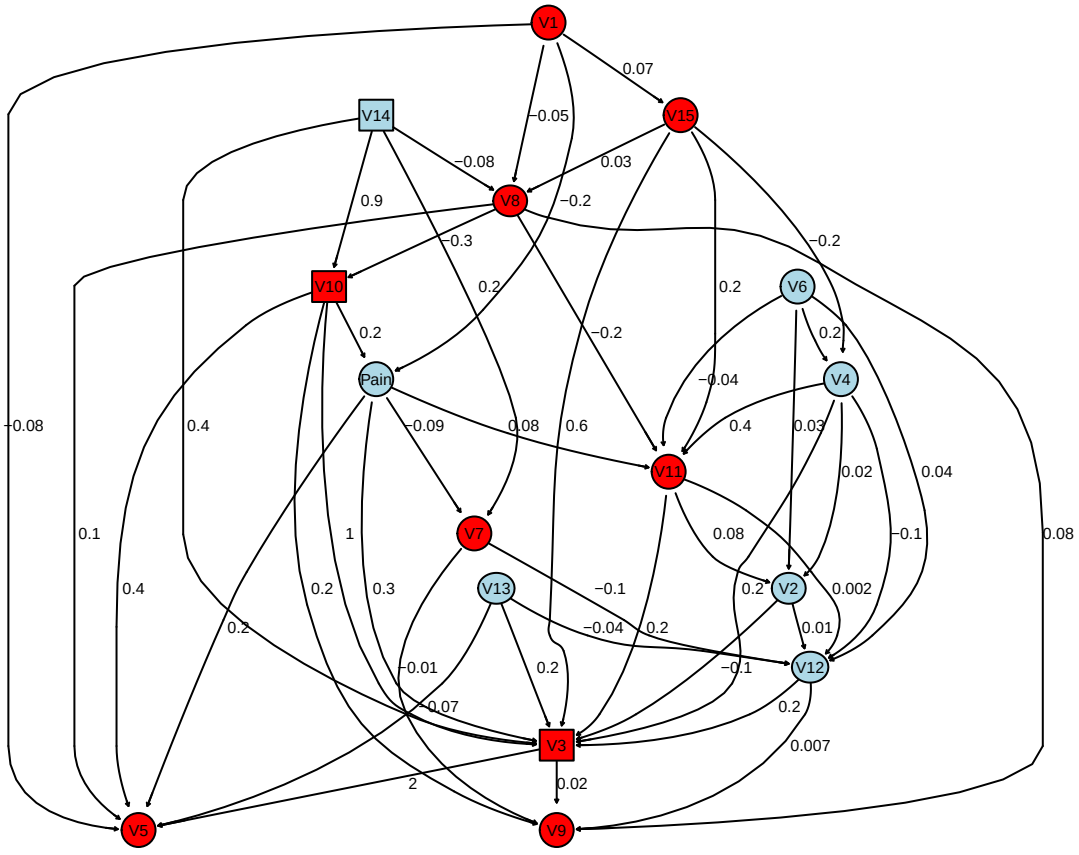

Figure 3.11: Additive Bayesian network directed acyclic graph of high-ranking variables, showing child-parent relationships (imputed data analysis). Variables in circles, ovals and rectangles represent Gaussian, Poisson and binomial variables, respectively. Variable coefficients are shown next to their respective edges (rounded to first significant digit). Markov blanket of *Pain* shown in red.

The additive Bayesian network of low-ranking variables stemming from an analysis of the imputed data set is shown in Figure 3.12 (parent-child). We observe variables *V1/Breed Warm Blood*, *V3/Hyper Brach* and *V11/Main Usage* being associated directly with *Pain/Pain Rank log*, just as is the case in the complete-case analysis. Variable *V3/Hyper Brach* and *V11/Main Usage* seem to be negatively associated with *Pain/Pain Rank log*, while *V1/Breed Warm Blood* seems to be positively associated with it. The Markov blanket additionally includes variable *V2/Defensive Pull* and *V8/Conformation*.

We find that associations between variables *V1/Breed Warm Blood* and *V3/Hyper Brach*, *V1/Breed Warm Blood* and *V7/Habit Patfix*, *V1/Breed Warm Blood* and *V8/Conformation*, *V5/Dressed BW Rider* and *V9/Broken Hoof*, *V10/Lameness* and *V7/Habit Patfix*, *V11/Main Usage* and *V8/Conformation*, *V11/Main Usage* and *V2/Defensive Pull*, and *V11/Main Usage* and *V3/Hyper Brach* all remain present, albeit the latter two associations changing in sign. In comparison to the complete-case analysis, we also find that *V11/Main Usage* is now additionally negatively associated with *V7/Habit Patfix*. Also in this network we find three variables not being associated with any other variable in the network (*V4/Lambskin*, *V6/License* and *V13/Training Freq*). The association of *V8/Conformation* and *V3/Hyper Brach*, as well as *V8/Conformation* and *V9/Broken Hoof* have been reversed in contrast to the direction found in the complete-case analysis. Furthermore, the associations between *V1/Breed Warm Blood* and *V2/Defensive Pull*,

as well as the association between *V11/Main Usage* and *V9/Broken Hoof* have been broken in the analysis of the imputed data set. As is the case with the complete-case and imputed data analysis of high-ranking variables, general network measurements listed in Table 3.5 remain unchanged for the network of imputed low-raking variables, the only exception being a slightly reduced average Markov blanket set size.

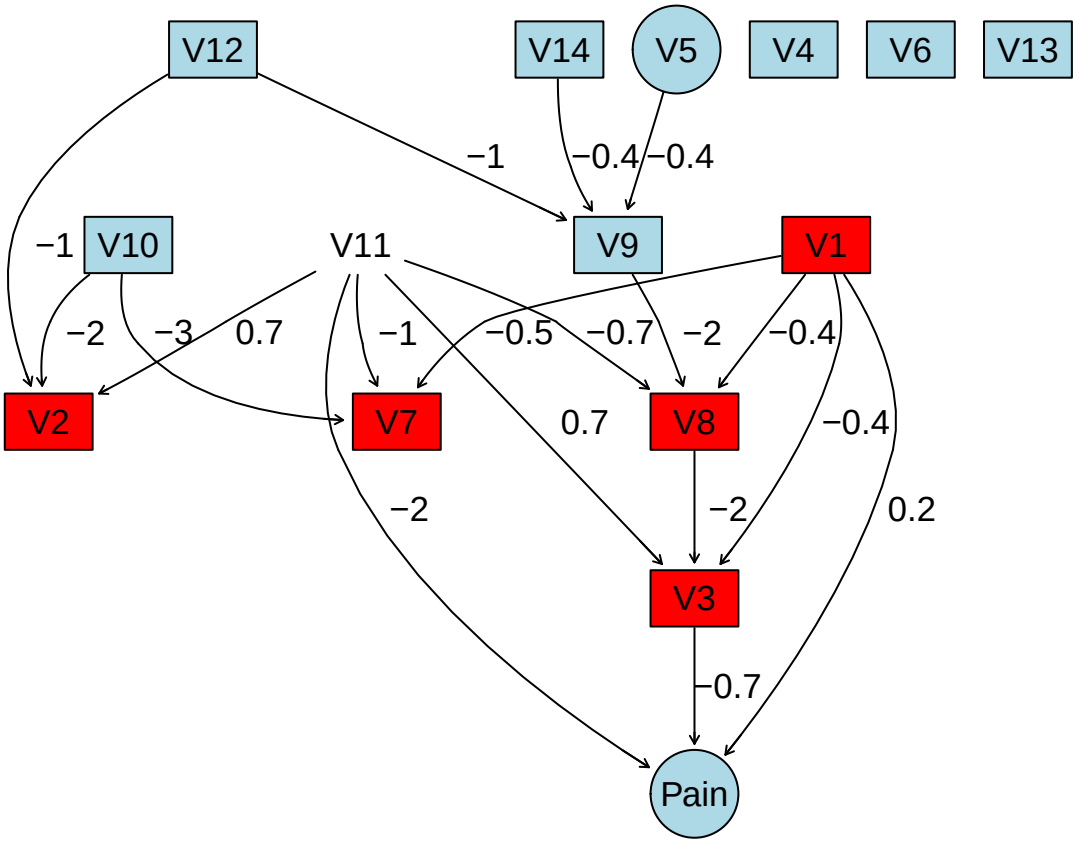

Figure 3.12: Additive Bayesian network directed acyclic graph of low-ranking variables, showing parent-child relationships (imputed data analysis). Variables in circles, ovals, rectangles and no enclosure represent Gaussian, Poisson, binomial and multinomial variables, respectively. Variable coefficients are shown next to their respective edges (rounded to first significant digit). Markov blanket of *Pain* shown in red ($+V11$).

Table 3.5: General information about the imputed data analysis directed acyclic graph of low-ranking variables.

Network Measure	Value
Number of Nodes	15
Number of Arcs	18
Markov Blanket Average Set Size	3.6
Neighbour Average Set size	2.4
Parent Average Set Size	1.2
Children Average Set Size	1.2

Figure 3.13 is the child-parent representation of the graph shown in Figure 3.12. As a result, relationships and coefficients are reversed.

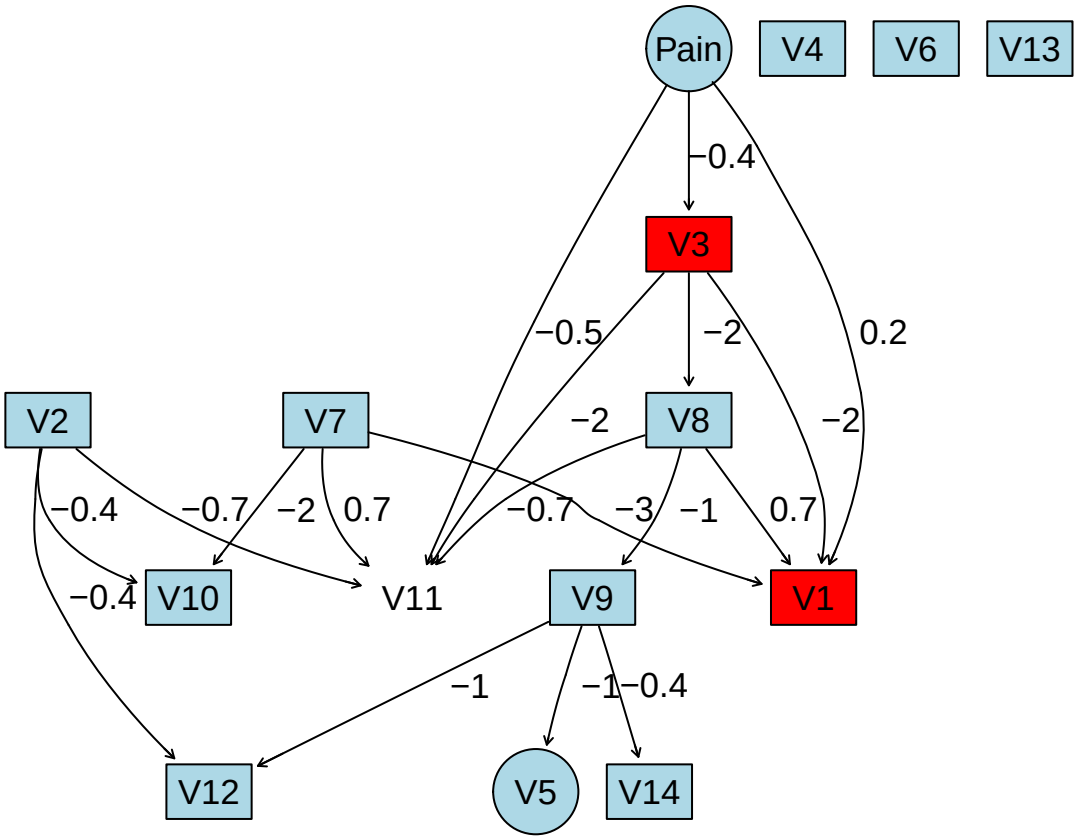

Figure 3.13: Additive Bayesian network directed acyclic graph of low-ranking variables, showing child-parent relationships (imputed data analysis). Variables in circles, ovals, rectangles and no enclosure represent Gaussian, Poisson, binomial and multinomial variables, respectively. Variable coefficients are shown next to their respective edges (rounded to first significant digit). Markov blanket of *Pain* shown in red ($+V11$).

3.5 Robustness Analysis Results

The results of the additive Bayesian network model robustness analysis of complete cases performed on the of the eleven highest-ranking variables are shown as a matrix representation in Table 3.6. The table shows the sum of all associations present after iteratively resampling the data and constructing a network model. We observe few associations being present as little as once or twice in the network. We also observe two associations being present throughout all network models. Most importantly, we notice that most associations that were not present at all are restricted by the banned arc matrix/restriction matrix in Table 2.4 set up to model the variables.

Table 3.6: Matrix representation of the absolute frequency and percentage with which associations are present in the robustness analysis of the complete-case data set. An 'X' represents a banned association in the banned arc/restriction matrix, while 'O' represents a non-restricted association.

	Pain	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
Pain	0|X	32|O	0|X	79|O	0|X	63|O	0|X	74|O	6|O	64|O	13|O	69|O
V1	68|O	0|X	0|X	54|O	0|X	80|O	0|X	52|O	28|O	53|O	24|O	41|O
V2	0|X	0|X	0|X	14|O	5|O	0|X	2|O	0|X	0|X	0|X	0|X	1|O
V3	0|X	0|X	75|O	0|X	62|O	100|O	22|O	0|X	0|X	68|O	0|X	53|O
V4	0|X	0|X	95|O	38|O	0|X	0|X	35|O	0|X	0|X	0|X	0|X	65|O
V5	0|X	0|X	0|X	0|X	0|X	0|X	0|X	0|X	0|X	0|X	0|X	0|X
V6	0|X	0|X	98|O	78|O	65|O	0|X	0|X	0|X	0|X	0|X	0|X	79|O
V7	0|X	0|X	0|X	0|X	0|X	0|X	0|X	0|X	37|O	100|O	0|X	0|X
V8	18|O	22|O	0|X	63|O	0|X	81|O	0|X	25|O	0|X	33|O	60|O	98|O
V9	0|X	0|X	0|X	0|X	0|X	0|X	0|X	0|X	0|X	0|X	0|X	0|X
V10	58|O	22|O	0|X	80|O	0|X	74|O	0|X	0|X	29|O	71|O	0|X	65|O
V11	0|X	0|X	99|O	47|O	35|O	0|X	21|O	0|X	0|X	0|X	0|X	0|X

Shown in Figure 3.14 and Figure 3.15 is the directed acyclic graph of associations shown in Table 3.6 which are present at least 50% of the time during the simulation, e.g. a pruned network. The thickness of edges in Figure 3.14 is representative of the frequency with which the association appeared during the simulation. Figure 3.15 shows the respective network model along with the coefficients. We observe all variables with the exception of *V2/Coord Rider* being included in the Markov blanket of *Pain/Pain Rank log*. Associations between variables *V2/Coord Rider* and *V6/Mean ROM*, *V2/Coord Rider* and *V4/Mean Velo*, and *V2/Coord Rider* and *V11/Mean Force* were present in the vast majority of resampled network models, which is consistent with the fact that all of these variables step from the rider itself. Additionally, associations between *V3/Saddle Type* and *Pain/Pain Rank log*, and *V7/Pasture Score* and *Pain/Pain Rank log*, and *V11/Mean Force* and *Pain/Pain Rank log* were also present in the vast majority of cases. Furthermore, associations between variables *V5/Horse Height* and *V3/Saddle Type*, *V9/Horse Age* and *V7/Score Meadow*, and *V11/Mean Force* and *V8/Number Symptoms* were also present in the vast majority of networks.

We now find variables *V3/Saddle Type*, *V5/Horse Height*, *V7/Pasture Score* and *V9/Horse Age* being directly associated with *Pain/Pain Rank log*. There is now a positive association between *V3/Saddle Type* and *Pain/Pain Rank log*, as well as *V7/Pasture Score* and *Pain/Pain Rank log*, while both *V5/Horse Height* and *V9/Horse Age* are both negatively associated with it. Particularly, we find positive associations between variables *Pain/Pain Rank log* and *V1/Algo Measures*, *Pain/Pain Rank log* and *V10/Tail Pos*, *V3/Saddle Type* and *V1/Algo Measures*, and *V11/Mean Force* and *V3/Saddle Type*. We also find negative associations between variables *V10/Tail Pos* and *V8/Number Symptoms*, *V3/Saddle Type* and *V10/Tail Pos*, and *V9/Horse Age* and *V3/Saddle Type*.

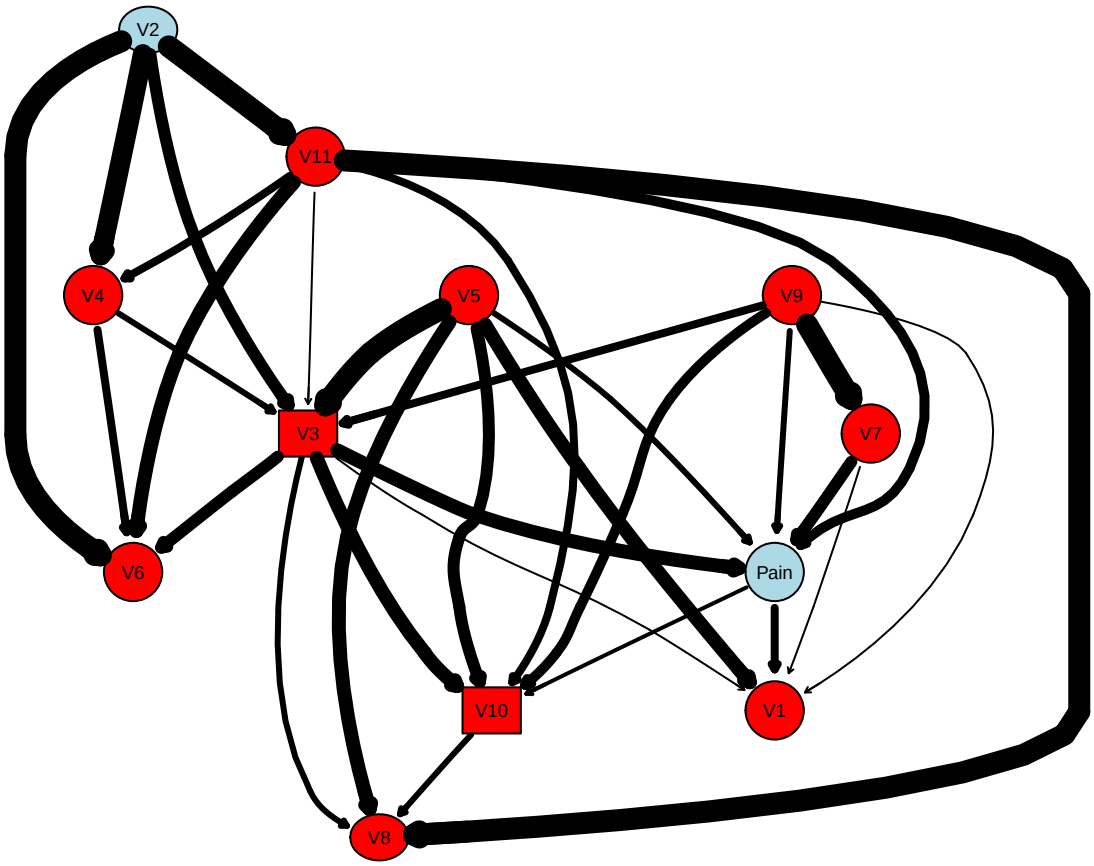

Figure 3.14: Additive Bayesian network directed acyclic graph of the 11 highest-ranking variables, showing only child-parent associations present in at least 50 percent of graphs in the robustness analysis of the complete-case and random-forest-imputed data set. Variables in circles, ovals and rectangles represent Gaussian, Poisson and binomial variables, respectively. Markov blanket of *Pain* shown in red. The thickness of edges are a representation of the frequency with which an association was present in the robustness analysis.

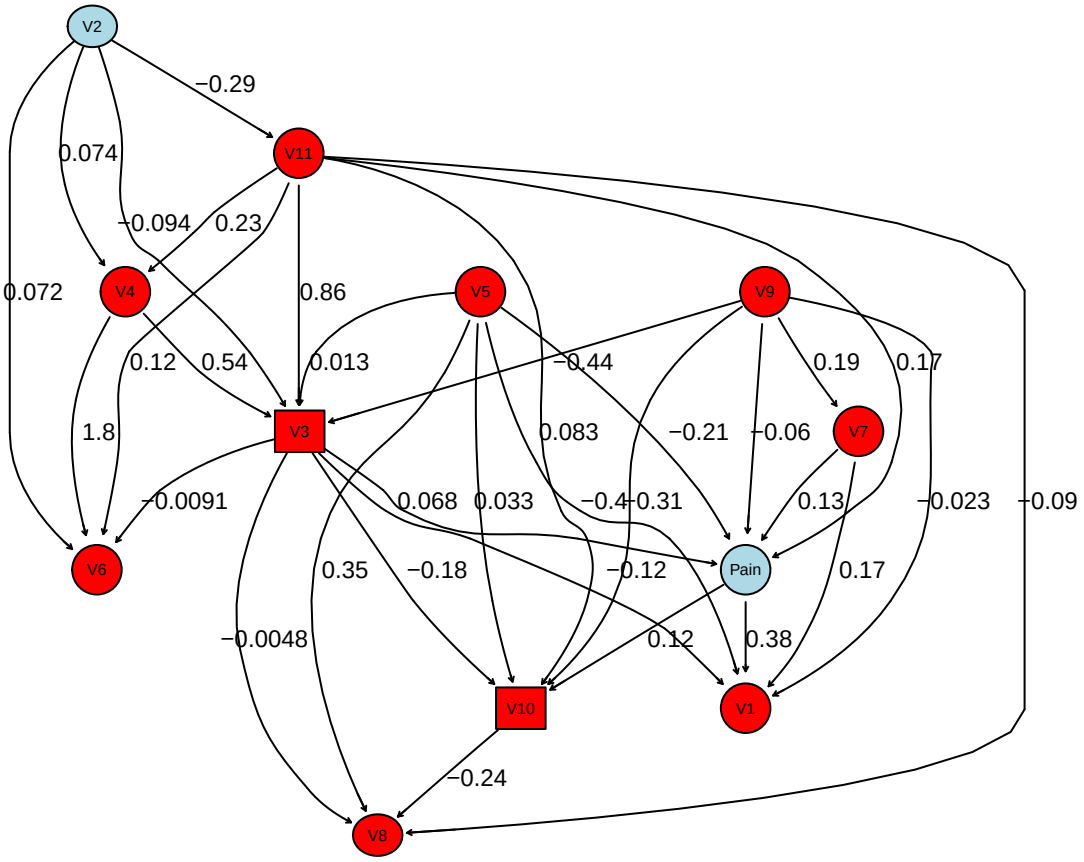

Figure 3.15: Additive Bayesian network directed acyclic graph of the 11 highest-ranking variables, showing only child-parent associations present in at least 50 percent of graphs in the robustness analysis of the complete-case data set. Variables in circles, ovals and rectangles represent Gaussian, Poisson and binomial variables, respectively. Variable coefficients are shown next to their respective edges (rounded to first significant digit). Markov Blanket of *Pain* shown in red.

Shown in Table 3.7 is a matrix representation of the number of times an association was found during the resampling simulation of the imputed set of the eleven highest-ranking variables after the set was randomly imposed with missing values. We notice that it is nearly identical to the matrix representation of the complete-case simulation shown in Table 3.6. Table 3.8 is a matrix representation of the differences between Table 3.6 and Table 3.7.

Table 3.7: Matrix representation of the absolute frequency and percentage with which associations are present in the robustness analysis of the random-forest-imputed, missing-values-imposed complete-case data set.

	Pain	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
Pain	0	32	0	79	0	63	0	74	6	64	12	70
V1	68	0	0	53	0	79	0	53	29	53	24	41
V2	0	0	0	14	5	0	2	0	0	0	0	1
V3	0	0	76	0	62	100	22	0	0	67	0	53
V4	0	0	95	38	0	0	35	0	0	0	0	64
V5	0	0	0	0	0	0	0	0	0	0	0	0
V6	0	0	98	78	65	0	0	0	0	0	0	79
V7	0	0	0	0	0	0	0	0	36	100	0	0
V8	18	21	0	64	0	80	0	26	0	33	60	98
V9	0	0	0	0	0	0	0	0	0	0	0	0
V10	59	22	0	80	0	74	0	0	29	70	0	65
V11	0	0	99	47	36	0	21	0	0	0	0	0

Table 3.8: Matrix representation of the absolute difference between associations present in the robustness analysis of the random-forest-imputed and the complete-case data set.

	Pain	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
Pain	0	0	0	0	0	0	0	0	0	0	1	-1
V1	0	0	0	1	0	1	0	-1	-1	0	0	0
V2	0	0	0	0	0	0	0	0	0	0	0	0
V3	0	0	-1	0	0	0	0	0	0	1	0	0
V4	0	0	0	0	0	0	0	0	0	0	0	1
V5	0	0	0	0	0	0	0	0	0	0	0	0
V6	0	0	0	0	0	0	0	0	0	0	0	0
V7	0	0	0	0	0	0	0	0	1	0	0	0
V8	0	1	0	-1	0	1	0	-1	0	0	0	0
V9	0	0	0	0	0	0	0	0	0	0	0	0
V10	-1	0	0	0	0	0	0	0	0	1	0	0
V11	0	0	0	0	-1	0	0	0	0	0	0	0

Figure 3.14 and Figure 3.16 show the pruned and robust additive Bayesian network models of the imputed set of variables after imposing missing values upon the complete case data set. Due to the fact that both the complete-case robustness analysis and the imputed set robustness analysis resulted in almost exactly the same association matrices, Figure 3.14 also represents the pruned acyclic graph of the robustness analysis of the imputed data set, showcasing the frequency with which associations were present during the simulation. Structurally, both network models are virtually identical. We only find differences in the coefficients describing the network associations. Most of the coefficients in Figure 3.16 are very similar to the those shown in

Figure 3.15. We now find that almost all coefficients retained their respective sign. Table 3.9 provides general information measures about the directed acyclic graph of both the complete cases and the random-forest-imputed, missing-value-imposed data set robustness analysis.

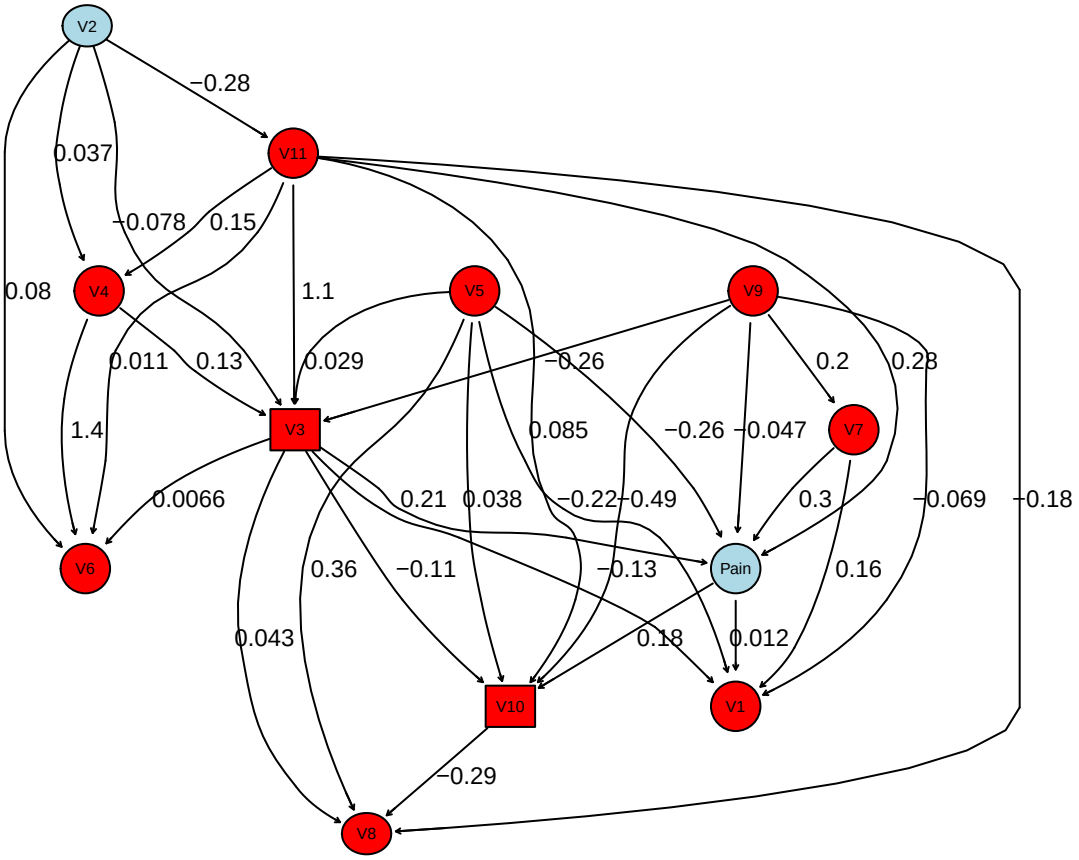

Figure 3.16: Additive Bayesian network directed acyclic graph of the 11 highest-ranking variables, showing only child-parent associations present in at least 50 percent graphs in the robustness analysis of random-forest-imputed, missing-values-imposed complete-case data set. Variables in circles, ovals and rectangles represent Gaussian, Poisson and binomial variables, respectively. Variable coefficients are shown next to their respective edges (rounded to first significant digit). Markov Blanket of *Pain* shown in red.

Table 3.9: General information about the directed acyclic graph of complete cases and the random-forest-imputed robustness analysis of the 11 highest-ranking variables.

Network Measure	Value
Number of Nodes	12
Number of Arcs	32
Markov Blanket Average Set Size	7.33
Neighbour Average Set size	5.33
Parent Average Set Size	2.67
Children Average Set Size	2.67

The results of the robustness analysis of the complete cases of missing-value-imposed complete case set of the eleven highest-ranking variables are shown as a matrix representation in Table 3.10. The table shows the sum of all associations present in a network model after iteratively resampling the data and creating a network model.

Table 3.10: Matrix representation of the absolute frequency and percentage with which associations are present in the robustness analysis of the complete-cases of the missing-values-imposed complete-case data set.

	Pain	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
Pain	0	25	0	86	0	43	0	94	15	57	12	68
V1	41	0	0	72	0	67	0	63	26	59	19	53
V2	0	0	0	6	0	0	1	0	0	0	0	0
V3	0	0	83	0	42	100	30	0	0	72	0	50
V4	0	0	100	58	0	0	58	0	0	0	0	70
V5	0	0	0	0	0	0	0	0	0	0	0	0
V6	0	0	99	70	42	0	0	0	0	0	0	77
V7	0	0	0	0	0	0	0	0	18	100	0	0
V8	24	22	0	37	0	59	0	32	0	60	80	86
V9	0	0	0	0	0	0	0	0	0	0	0	0
V10	71	38	0	66	0	79	0	0	18	69	0	59
V11	0	0	100	50	30	0	23	0	0	0	0	0

Shown in Figure 3.17 and 3.18 is the pruned directed acyclic graph of associations shown in Table 3.10. There are quite substantial differences to the previous models in both structural terms and in the coefficients that describe the found associations. All variables with the exception of *V2/Coord Rider* and *V5/Horse Height* are included in the Markov blanket of *Pain/Pain Rank log*. Associations between variables *V2/Coord Rider* and *V6/Mean ROM*, *V2/Coord Rider* and *V4/Mean Velo*, and *V2/Coord Rider* and *V11/Mean Force* are once again present in the vast majority of resampled network models. Associations between *V5/Horse Height* and *V3/Saddle Type*, *V9/Horse Age* and *V7/Score Meadow* were also present in the vast majority of networks. Variables *V3/Saddle Type*, *V7/Pasture Score* and *V9/Horse Age* and *V11/Mean Force* are now directly associated with *Pain/Pain Rank log*. There is now a negative association between *V3/Saddle Type* and *Pain/Pain Rank log*, as well as a positive association between *V7/Pasture Score* and *Pain/Pain Rank log*. Both *V9/Horse Age* and *V11/Mean Force* are positively associated with *Pain/Pain Rank log*. We find that the direct association between *V1/Algo Measures* and *Pain/Pain Rank log*, and *V11/Mean Force* and *V3/Saddle Type* are not present anymore. Additionally, the association between *V9/Horse Age* and *V3/Saddle Type* changed in sign. However, we still find the same associations between variables, *Pain/Pain Rank log* and *V10/Tail Pos*, *V3/Saddle Type* and *V1/Algo Measures*, *V10/Tail Pos* and *V8/Number Symptoms*, *V3/Saddle Type* and *V10/Tail Pos*. Table 3.11 provides general information measures of graphs shown in Figure 3.17 and Figure 3.18. We find a small decrease in the number of arcs, average neighbour, parent and child set size, and a slight increase in the average Markov blanket set size.

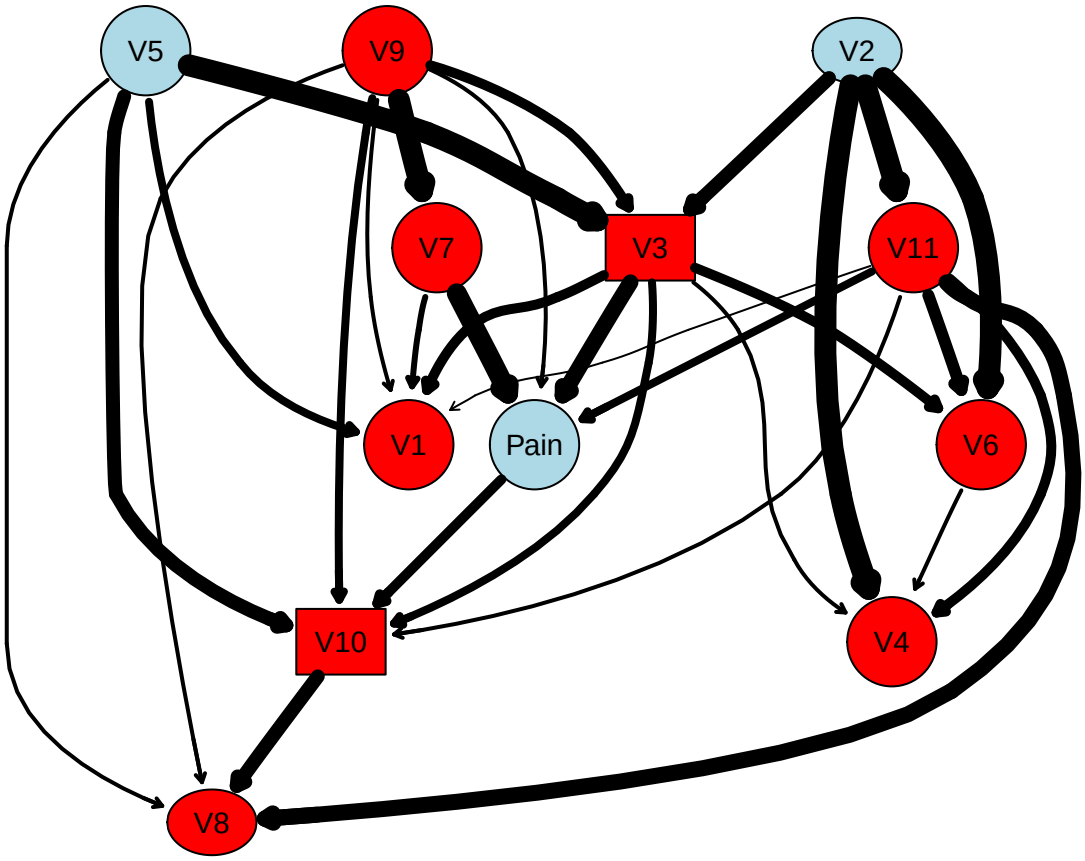

Figure 3.17: Additive Bayesian network directed acyclic graph of the 11 highest-ranking variables, showing only child-parent associations present in at least 50 percent graphs in the robustness analysis of the complete-cases of the missing-values-imposed complete-case data set. Variables in circles, ovals and rectangles represent Gaussian, Poisson and binomial variables, respectively. Markov Blanket of *Pain* shown in red. The thickness of edges is a representation of the frequency an association was present in the robustness analysis.

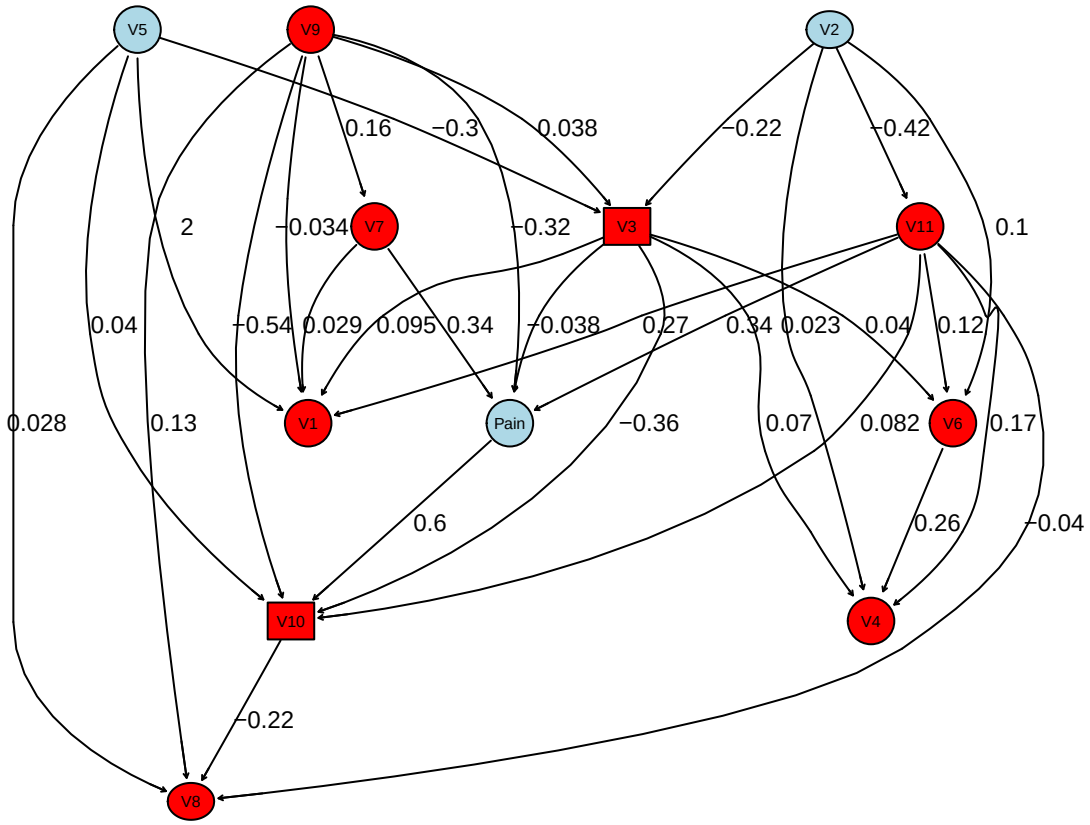

Figure 3.18: Additive Bayesian network directed acyclic graph of the 11 highest-ranking variables, showing only child-parent associations present in at least 50 percent graphs in the robustness analysis of the complete-cases of the missing-values-imposed complete-case data set. Variables in circles, ovals and rectangles represent Gaussian, Poisson and binomial variables, respectively. Variable coefficients are shown next to their respective edges (rounded to first significant digit). Markov Blanket of *Pain* shown in red.

Table 3.11: General information about the directed acyclic graph of the complete-cases of the missing-values-imposed complete-case data set robustness analysis of the 11 highest-ranking variables.

Network Measure	Value
Number of Nodes	12
Number of Arcs	30
Markov Blanket Average Set Size	7.83
Neighbour Average Set size	5
Parent Average Set Size	2.5
Children Average Set Size	2.5

Chapter 4

Discussion

Variable ranking analysis on the outcome *Pain Rank log* clearly shows a sizable amount of redundant variables along the selected variables. The analysis using `varrank` strongly underlines the importance of variable selection for analyses such as additive Bayesian network models. The ability to select variables without the explicit need of either a model or expert knowledge can be a great help to epidemiologists, especially when dealing with high-dimensional and mixed-type data. Furthermore, the lack of necessity for a model works around step-wise model selection processes that may induce bias and overfit the true effect of variables. During this analysis we found that the package `abn` could easily and reliably model around 15 variables, but struggled computationally once the amount exceeded. Additionally, the interpretation and visualization of an additive Bayesian model significantly decreases once too many variables are included in the network. This further encourages variable selection processes an possibly splitting an analysis in several parts, as done with these analyses.

The proportion of misclassified variables and the mean square error of variables for the imputed data set seem to be acceptable as a whole, albeit slightly high in few cases, especially for categorical variables. Random forest imputation proves to change the results of additive Bayesian networks to some degree, when comparing the results of a complete-case analysis to the analysis of a random forest imputed data set while not controlling for overfitting. Not only did the presence of edges change between nodes (i.e. the network structure), but the coefficients changed in a multitude of cases as well. The instability of random forest imputation heavily depends on the amount of missing data. The higher the amount of missing data, the less accurate random forest imputation will be. Nevertheless, the flexibility of random forest imputation makes its use strongly advantageous when dealing with high-dimensional, mixed-type data.

The analysis of maximum number of parents for any additive Bayesian network model yielded some rather surprising results. In each case, we found a constant and a rapid change in marginal likelihood at an already low number of maximum parent nodes. Much rather, we expected the marginal likelihood to remain constant at a higher amount of maximum number of parent nodes allowed. We argue that this is due to the BIC score penalization getting increasingly heavier as the number of parents rise. This ultimately results in the very same network being fitted repeatedly, thus resulting in the exact same marginal likelihood value. Choosing the maximum number of parent nodes allowed in a network model is an important modelling decision. Increasing the number of parent nodes present in an additive Bayesian network model heavily increases its complexity and decreases readability and interpretability. A higher model complexity possibly allows for the identification of more associations between variables, whereas model complexity is also problematic in the sense of overfitting. We reason that choosing a rather high number of maximum parent nodes allowed may be beneficial towards capturing possible associations, but may prove detrimental in the sense of overfitting and thus, generalization.

Additive Bayesian network models are indeed a very elegant model choice for holistic analyses. However, analyses of epidemiological data sets will often require specification of relations that should be retained or banned. This is done by creating a matrix specifying edges that should be either retained or banned when searching the model's optimal structure. We find that the slightest change in such a matrix may cause the results of an additive Bayesian network model to be widely different from before. Not only may the structure of the model change, but the coefficients describing relationships between variables may do so as well. This proves to be problematic when trying to draw quantitative conclusions from an additive Bayesian model. In the same sense, it is also difficult to draw qualitative conclusions, such as the direction of associations. The chosen restriction/retention matrix is not necessarily unique and may change depending on the reasoning behind possible causal relations between variables. This is why the construction of a restriction/retention matrix should be constructed along with expert opinion. However, a certain amount of subjectivity will always be present when constructing such a matrix, even when expert opinion is available. Considering the possible changes in coefficients and structure of a network when changing the restriction/retention matrix, we encourage care whenever drawing stronger conclusions from an additive Bayesian network model. We only show the results of an additive Bayesian network model that uses a single optimal restriction matrix (for high- and low-ranking variables, respectively), as it would be beyond the scope of this document to list the resulting networks drawn from using different restriction matrices that were built iteratively during this analysis.

We were able to identify several observations from the complete-case analysis of high-ranking variables that were consistent with our expectations previous to the statistical analysis. We found that algometry measures (a measure of pain/pressure resistance) are directly associated with back pain. The findings suggest that horses with back pain tolerate lower algometric pressures than horses with back pain. These observations are consistent with our beliefs that algometry measures may serve as a surrogate for equine back pain. The height and frequency of a horse's use were negatively associated with algometry values. The bigger the horse and the more often it was ridden, the lower the algometry value. Assuming that algometry value is a surrogate for back pain, this would indicate higher levels of back pain in tall, frequently ridden horses. On the other hand, overall saddle pressure and a rider's strength seem to increase algometry values, indicating lower back pain. This is surprising, as one would expect higher pressures to result in more pain. However, the higher algometry values could also indicate a higher resistance to pressure. Additionally, the more time a horse spends freely on a pasture, the lower the back pain score is and the smaller the chances of it having a chiropractic findings in the ilium. Moreover, the older a horse is, the more time it spends on the pasture and the less frequently it is ridden. This may be a reflection of the fact that rest and a natural environment are beneficial to a horse's health and that older horses are spared more frequently from straining activities. We also found that English saddle types are presented with fewer fit problems and were more frequently used on slightly smaller, younger horses. English saddle types are also associated with a higher mean saddle pressure, which may also directly cause back pain. English saddles are often used by riders with athletic ambitions, as most competitions in Switzerland focus on dressage and jumping. The rider's force, coordination, and reaction speed also appeared to influence the type of saddle being used. As saddle pressure was also influenced by the rider's force and reaction speed/coordination, this may indicate that not only the saddle, but also the riders technique may have impacted saddle pressure. Western saddle types on the other hand are more often used by less competitive, leisure riders, which may be why we observe an increased chance of Western saddle types in older and taller horses.

Also regarding the complete-case analysis of high-ranking variables, it is important to mention that a horse may build up a certain resistance to pain with time. We argue that stronger riders may exert greater force on the saddle, thus applying greater pressure and thereby increase a horse's resistance to pressure, resulting in higher algometry values. Additionally, higher saddle pressure was associated with chiropractic findings in the pelvis/ilium, which could be a direct result of riding. It is also apparent that tall, frequently ridden horses showed a higher number of (owner-reported) symptoms for discomfort, pain and/or lameness. Additionally, we argue that riding frequency directly influences the integrity of a horse's saddle, thus increasing the number of saddle problems. This may be a result of wear and tear and the overall quality of a saddle. Alternatively, riders who ride their horse more often may be less attentive of saddle fit issues. More importantly, we found that a horse's height and a rider's overall strength increase the chances of an abnormal tail position, which is a possible reflection of an underlying orthopedic issue. These observations thoroughly underline the assumption that horse-riding is straining for a horse's back. Contrary to our expectations however, and despite higher saddle pressure values, we found that English saddle types were associated with a lower back pain. On the other hand, this is a result that changed after controlling for overfitting. Also contrary to previous expectations, we found that an increased horse age lowers the chances of having an abnormal tail position. These findings could indicate that younger horses are more prone to orthopedic problems in the pelvis, possibly due to a lack of strength and balance at the beginning of their career as a riding horse. However, there was a negative association between chiropractic findings in the pelvis/ilium and an abnormal tail position, which would mean that an abnormal tail position is not indicative of orthopedic problems in the pelvis. We argue that the direction, sign and magnitude of some findings may not truly reflect the causal pathway behind the system and that they may have come to arise due to chance or unaccounted factors.

In contrast to high-ranking variables, the complete-case analysis of low-ranking variables resulted in many associations that were unexpected. Contrary to our previous expectations, hypertrophy in the brachiocephalic muscle was associated with a lower back pain score and a lower chance of suffering from a special back conformation, such as a longer or shorter back than usual. Additionally, special back conformation was associated with a decreased chance of a backwards broken hoof pastern axis and lame horses were associated with having an increased chance of having no issues with kneecap fixation. The negative association between a rider's body weight and the chances of a backwards broken hoof pastern axis were also rather surprising. Interestingly, we also found that warm blood breeds have a decreased chance of having issues with the habitual fixation of the kneecap, a common orthopedic issue in horses, and with a backwards broken hoof pastern axis. This is surprising, as warm blood horses are typically bred for athletic purposes and thus are more prone to having orthopedic issues. Furthermore, we find the chances of a backwards broken hoof pastern axis being decreased in horses that are used for dressage, military and jumping disciplines. Also here we argue that these finding may not necessarily reflect the causal pathway behind the system, or that the associations arose due to chance or unaccounted factors. Moreover, we argue that low-ranking variables may not carry the highest quality information to describe the system at hand, which may be due to the splitting of the analysis in higher- and lower-ranking variables with respect to the relevance towards back pain. Given the systemic nature of a biological living creature, we much rather expected to find the inverse for some of the above relations, as a problem or issue in one area may have systemic effects on other related areas. However, it is interesting to notice how many variables were "correctly" associated with each other in a semi-supervised model search, albeit the coefficients describing their relations pointing in the opposite direction of what was expected. On the other hand, we do find that dressage, military and jumping horse riding disciplines all increase back pain, which is consistent with our notion that straining activities are detrimental to a equine back health. Additionally, warm blooded breeds were associated with having higher back pain scores.

In general, the complete-case analyses of high-ranking variables resulted in a network that included many intuitive and expected associations. We recognize that the saddle, rider fitness, horse exercise and natural husbandry all play a crucial role in the complex system resulting in equine back pain. Conversely, there were less explicable relations in most associations of low-ranking variables.

Data imputation caused changes in the results of high-ranking variables to a certain degree. A multitude of associations either ceased to exist, were reversed in direction or changed in coefficient sign. However, we still do find many associations consistent with our conclusions drawn from the complete-case analysis, since husbandry, saddle type and rider strength continued playing a central role in equine back pain. We also noticed that, similar to the complete-case analysis, horse age and height were closely associated with the outcome back pain. Given that the complete-case analysis resulted in so many intuitive interpretations of model coefficients, we are now unable to explain some of the model's coefficients in a medical-biological sense. It is however encouraging to see that many expected associations remained. The presence of associations alone serve in the investigation of closely related variables and are a reason enough to see the complex interplay of variables in this system. An interesting observation is the large size of the Markov blanket for high-ranking variables. The Markov blanket is the set of variables that are needed in order to fully infer on a variable of interest. We argue that the size of the Markov blanket may thus be a good reflection of the complexity of the system that leads to equine back pain. However, we find the Markov blanket employed by the package `abn` being slightly different than previously defined. Rather than being the set of parents, children and those children parents of a target node, it seems as if it was defined as the set of parents, those parent's children and children. The results of the imputed data analysis of low-ranking variables was on the other hand fairly similar to the complete-case analysis. As a result, no new particularly striking associations could additionally be found. Since both analyses of low-ranking variables resulted in similar associations, it is difficult to explain the model's coefficients in a sensible medical-biological manner.

Additive Bayesian network models may change their results given a certain procedure, such as setting a restriction/retention matrix or imputation. We argue that in some cases, these analyses may be quite delicate in nature. Subsequently, we encourage interpreting the models' coefficients with care and rather encourage drawing attention to the presence or absence of associations. Given this fragility and taking into account that causality is not necessarily reflected by a certain network model, it may sometimes be important to depict not only parent-child relationships, but also the reverse child-parent relationships. Thus, additive Bayesian models may be a suitable tool to detect and explore possible associations, which may later be tested under more strict experimental conditions. In order to explore the instability of results after imputation, a random forest robustness analysis was conducted. In addition to analyzing the robustness of random forest imputation, the conducted simulation analysis also controlled for overfitting. Furthermore, the robustness analysis simulated a case where "true" data is available and compared the results with a complete case analysis after randomly imposing missing values. After controlling for robust networks, it is clear that random forest imputation indeed produce very stable and usable results in regards of additive Bayesian network modelling. Thus, additive Bayesian network models prove to be more much more robust against random forest imputation when controlling for overfitting. We find that a simple algorithm that counts and prunes the associations from a series of network models after several rounds of resampling is enough to produce robust results. Controlling for overfitting also reduces the necessity of investigating both parent-child and child-parent networks, as the direction of associations is also controlled for.

After overfitting correction, the resulting structure of an additive Bayesian network models after a single round of random forest imputation did not change at all, while there was also a very strong increase in the robustness of the coefficients. On the other hand, after performing a robust analysis of the complete cases of the missing-value-imposed complete case data set, both the structure and coefficients changed rather strongly. Thus, random forest imputation may change the results of an additive Bayesian network analysis more that when accounting for overfitting. In other words you may argue, that it is not the imputation in itself that causes a change in the network, but rather the overfitting combined with imputation, while a complete-case analysis may produce results that differ strongly, regardless whether overfitting was controlled or not. Although overfitting and complete-case analyses may indeed strongly dictate the structure and coefficients of additive Bayesian network models, we proclaim that complete-case analysis without controlling for overfitting may also serve as an exploratory tool. Additionally, we propose that the discrepancies between the complete-case robustness analysis and the remaining analyses in the simulation may be due to a relatively high number of lost cases after randomly introducing missing values. Even though there were some changes in the network models after accounting for overfitting, we still find several associations that were present throughout all analyses. Despite strong structural similarities, most coefficients of robust network models were equally difficult to explain in a medical-biological sense.

Throughout all results we found that algometry measures were closely related to the outcome of equine back pain measure, which was determined solely through palpitation examinations of experienced veterinarians. Initially, we hypothesized that pain resistance/algometry values should have an inverse relation with pain, e.g. the more pain resistance/higher algometry values, the less back pain and vice-versa. However, most results (throughout all analyses) point in the direction of a positive association between the two measures, e.g. the more pain resistance/higher algometry values a horse has, the more it suffers from back pain, which speaks for a different mechanism than previously hypothesized. We thus proclaim that a horse with a high pain tolerance may actually be the result of equine back pain in a number of cases. Nonetheless, it is clearly evident that both measures are strongly associated with each other. In general terms, we also consistently found that next to algometry measures, the saddle type, pasture score, tail position, rider's mean force during riding and horse height and age were always very closely or directly related to equine back pain throughout all results. In that sense, since we find the same variables either directly influencing, or being closely related to equine back pain measure throughout all high-ranking variables analytic results, we propose that the central variables to the surge of back pain would have been well probably the same after a robustness analysis of low-ranking variables. This is even further underlined by the fact that the overfitted model of low-ranking variables after random forest imputation was very similar to the overfitted model using only the complete cases. Furthermore, we propose that low-ranking variables may carry too much redundancy between variables, thus rendering some results difficult to interpret, which is why a robustness analysis for low-ranking variables was not conducted. Nonetheless, the analyses of low-ranking variables also determined that straining usage of a horse may prove detrimental to its back health. Additionally, the analysis of low-ranking variables also showed that deficiencies in brachiocephalic muscular development and a horse's breed may also play a direct role in the surge of equine back pain. The results found throughout this thesis support many hypotheses and suspicions for the surge of equine back pain. However, some results underline the necessity of further investigations under more strictly controlled experimental conditions, which would increase the ability to detect the presence, direction, magnitude and sign of relationships.

Chapter 5

Conclusions

Even though there were some discrepancies between the results' coefficients, there was still a fair amount of structural overlap and common factors that always played a central role to the surge of equine back pain throughout all networks. We conclude that saddle type choice is an important factor to the surge of equine back pain. We propose that the found association is a result of English saddles being typically used by more aggressive and ambitious riders and that this may have a detrimental effect on horse back health. Algometry measurements indeed show a very close relation to equine back pain levels, such that it may serve for a simpler surrogate measure in future studies. We also propose that a natural husbandry, abundant time on a pasture and sparing straining activities are beneficial to a horses back health. We additionally find close associations between possible orthopedic and muscular developmental issues and equine back pain. Even though several associations were in synchronization with our expectations, we find it a challenge explaining the coefficients of some of the aforementioned associations in a biological-medical way, and rather encourage focusing on the presence or absence of associations. Nonetheless, we recognize the increasing potential and usefulness of additive Bayesian network modelling as a statistical tool for holistic analyses. At the same time, we recognize the large computational and time-effort required for these sort of models, derived through the necessity of careful modelling decisions, variable selection, imputation methods and further methodologies to deal with overfitting. It is also evident that it becomes increasingly difficult to dissect the vast amounts of information present in an additive Bayesian network model as the number of variables to model increase. We conclude that some of the findings throughout this thesis indeed highlight and support certain proposed mechanisms leading to equine back pain. However, they also highlight the necessity for further studies under more strict experimental settings, should the presence and quantification of factors influencing back pain be studied with greater certainty.

Bibliography

Breinman, L. (2001). Random Forests. *Machine Learning*, **45**, 1–2. 8

Dittmann, M., Latif, S., Hefti, R., Hartnack, S., Hungerbuehler, V., and Weishaupt, M. (2020a). Husbandry, Use, and Orthopeadic Health of Horses Owned by Competitive and Leisure Riders in Switzerland. *Journal of Equine Veterinary Science*, **91**, 1–5. 1

Dittmann, M., Latif, S., Hungerbuehler, V., Weishaupt, M., and Arpagaus, S. (2021). Feel the Force: Prevalence of Subjectivity Assessed Saddle Fit Problems in Swiss Riding Horses and Their Association with Saddle Pressure Measurements and Back Pain. *Journal of Equine Veterinary Science*, **78**, 1–2. 3

Dittmann, M., Latif, S., Weishaupt, M., Arpagaus, S., Gunst, S., Roepstorff, C., Klaassen, B., Pauli, C., and Bauer, C. (2019). Influence of Functional Rider and Horse Asymmetries on Saddle Force Distribution During Stance and in Sitting Trot. *Journal of Equine Veterinary Science*, **78**, 1–2. 3

Dittmann, M., Latif, S., Weishaupt, M., Arpagaus, S., Roepstorff, C., and Mueller-Quirin, J. (2020b). Riding Soundless: Comparison of Subjective with Objective Lameness Assessment of Owner-Sound Horses at Trot on a Treadmill. *Journal of Equine Veterinary Science*, **95**, 1–2. 3

Henson, F. (2013). *Equine Neck and Back Pathology: Diagnosis and Treatment*, volume 1. John Wiley and Sons. 1

Kratzer, G. and Furrer, R. (2018). Varrank: an R Package for Variable Ranking Based on Mutual Information with Applications to Observed Systemic Datasets. *arXiv*, **1804.07134**, 1–4. 6, 7

Kratzer, G. and Furrer, R. (2020). Varrank: an R Package for Variable Ranking Based on Mutual Information with Applications to Systems Epidemiology. `https://www.math.uzh.ch/pages/varrank/articles/varrank.html`. Online accessed 18 September 2020. 7

Kratzer, G., Furrer, R., Lewis, F., Comin, A., and Pittavino, M. (2019). Additive Bayesian Network Modelling with the R Package abn. *arXiv*, **1911.09006**, 1–20. 1, 11, 12

Kratzer, G., Furrer, R., Lewis, F., Willi, B., Meli, M., Boretti, F., Hofmann-Lehrmann, R., Togerson, P., and Hartnack, S. (2020). Bayesian Network Modelling Applied to Feline Calcivirus Infection Among Cats in Switzerland. *Frontiers in Veterinary Science*, **7**, 1–15. 9, 10

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 3

Stekhofen, D. (2012). MissForest: Non-Parametric Missing Value Imputation for Mixed-Type Data. *Bioinformatics*, **28**, 1–6. 8