Institut für Erd- und Umweltwissenschaften
Allgemeine Geophysik

EMPIRICAL GROUND-MOTION MODELS FOR PROBABILISTIC SEISMIC HAZARD ANALYSIS: A GRAPHICAL MODEL PERSPECTIVE

Kumulative Dissertation
zur Erlangung des akademischen Grades
"doctor rerum naturalium"
(Dr. rer. nat.)
in der Wissenschaftsdisziplin "Seismologie"

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
UNIVERSITÄT POTSDAM

von
Dipl. Geophysiker Nicolas M. Kuehn

Berlin, den 31.08.2010

To my Father

CONTENTS

List of Figures

List of Tables

ACKNOWLEDGMENTS

This thesis would not have been completed and finalized without the help of numerous people - Thanks to all of you!

First of all, I would like to thank my supervisor Frank Scherbaum. His guidance and knowledge were invaluable for my work. Every time I got stuck, Frank knew how to anchor my work in the greater picture of PSHA. He was always supportive towards me and my work. I am especially grateful that he gave me the opportunity to work for the PEGASOS Refinement Project, which provided me with a lot of insight into the practical side of PSHA.

I would also like to thank Carsten Riggelsen, my (unofficial) second supervisor. Carsten brought the fields of machine learning and artificial intelligence to my attention, which provided both philosophy and methods for most of the work in this thesis. Without Carsten, my thesis would have looked quite different. Carsten was always helpful when I had problems understanding hard machine learning stuff, and pointed me to new directions. His view from outside the seismological community always provided new views on different aspects of PSHA and ground motion models. I am also grateful that I was allowed to use his program for learning Bayesian networks.

I would like to thank the whole geophysics group at the University of Potsdam for providing an enjoyable working environment. Especially Frank Krüger and Matthias Ohrnberger always had good questions/comments whenever I presented my work during seminars.

Helmut Städtke was very helpful with the computation of the inversions of equivalent stochastic models, as well as with some Mathematica related problems. In particular his implementation of the SMSIM code that runs on a GPU sped things up quite a bit.

Thanks to my fellow PhD-students, who made my PhD-time enjoyable, both inside and outside the University. Here in particular Urs and Steffen are to mention. Thanks also to all my office mates over the years: Dietrich, Andrés, Mauricio, Manu, Lydia, Yanqiu, Anke, Veronica, Fred and Henry. Thanks also to Urs for providing the LaTeX-style file for this thesis.

I would like to thank the University of Potsdam for providing financial support via the graduate school “Natural disasters”. I also thank Swiss Nuclear (especially Phillipe Renault) for providing some financial support, but in particular for letting me work as part of the PEGASOS Refinement project. In the line of this work and the workshops I attended, I learned a lot about the difficulties

that arise when it comes to the practical application of PSHA.

Thanks also to my family, who were always supportive of me, even though they didn't understand what I was doing (which is totally my failure, as I didn't explain my work well enough).

A special thanks goes to Fausto Coppi for being a good companion when I needed time for myself, which I could use to both clear my head of my work as well as think deeply about it.

Last but not least, this thesis would not have been completed without the help of many (and I mean very many) cups of coffee.

1

INTRODUCTION

1.1 Probabilistic Seismic Hazard and Ground Motion Models

The goal of a seismic hazard analysis is to quantitatively determine the ground-shaking hazard at a particular site. In many countries, present state of practice is that such an analysis should be carried out as a *probabilistic seismic hazard analysis* (PSHA). The concept of PSHA was pioneered in the 60's by work of Cornell and Esteva (see McGuire (2008) for a history of PSHA). An important point in PSHA is that it makes it possible to include aleatory uncertainty in the earthquake source process as well as in ground motions. Especially the last part is what is driving the hazard (Bommer and Abrahamson, 2006).

The basic concept of PSHA is shown in Figure 1.1. It consists of four steps (Reiter, 1990):

(a) All seismically active zones that are of engineering significance for the site are identified. The source zones can be faults or areal zones. For each source zone, a probability distribution of possible rupture locations is defined (usually uniform). This leads to a probability distribution over source-site distances for each source zone i, $f_{R,i}(r)$.

(b) The seismicity of each source zone i is characterized by a recurrence relation, such as a Gutenberg-Richter distribution or a characteristic distribution. This results in a probability distribution for magnitudes, $f_{M,i}(m)$.

(c) The ground motion resulting from an earthquake of a given size and distance is characterized by a predictive relationship - a so-called ground motion model (GMM). The GMM quantifies the conditional probability of a ground motion given magnitude and distance (and other parameters $\boldsymbol{X}$), $\Pr(Y|m, r, \boldsymbol{X})$. $\boldsymbol{X}$ contains parameters predictive of ground mo-

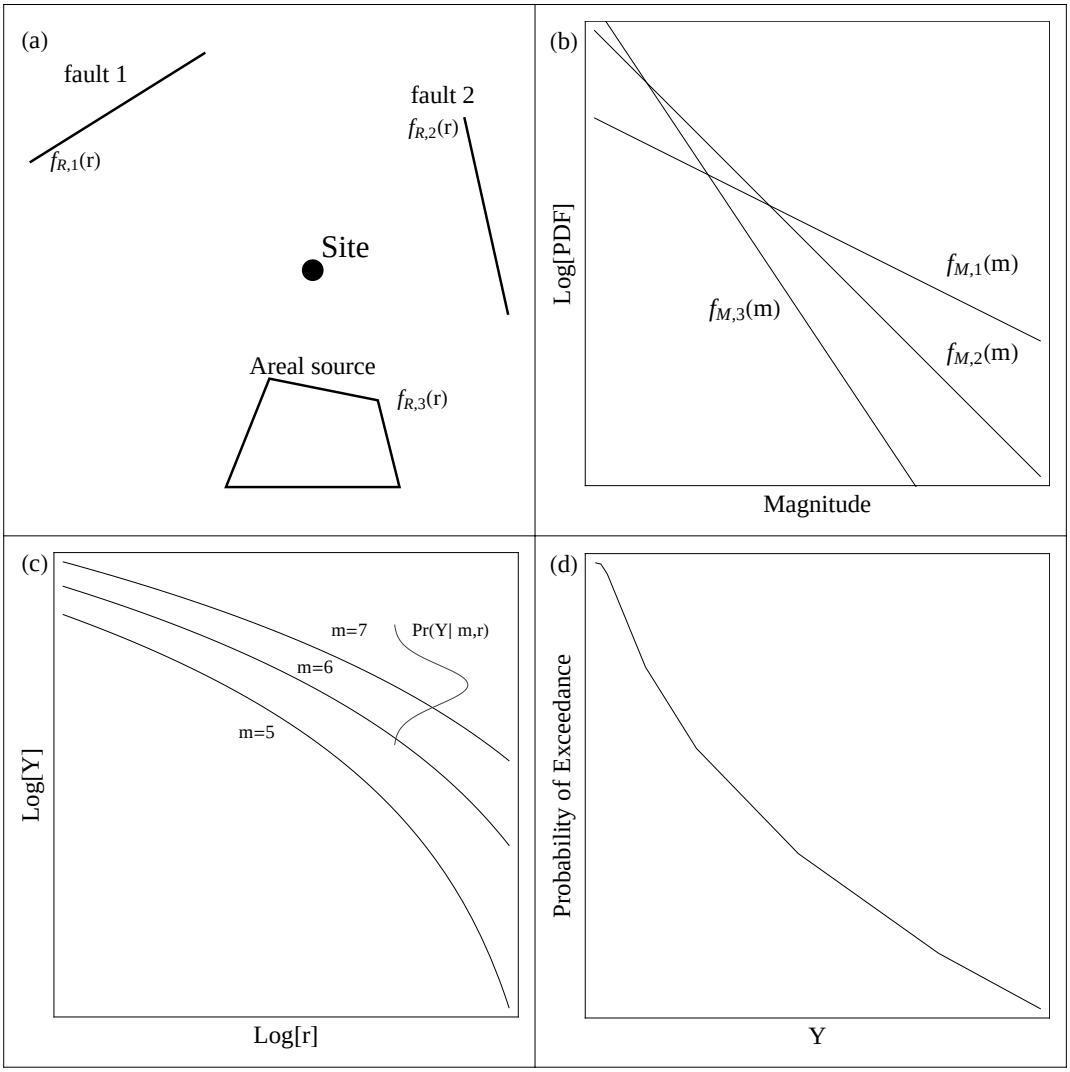

Figure 1.1: Basic concept of PSHA (after Reiter (1990)). See text for explanation.

tion besides magnitude and distance, such as style-of-faulting or a parameter describing site effects.

(d) The above considerations are combined to estimate the probability or rate that a certain ground motion level A will be exceeded at the site under consideration during a particular time period. This is done for different ground motion levels A, and the result is a so-called "hazard curve" which relates each ground motion level with its expected rate of exceedance.

The last step of PSHA is described by the following equation, which specifies how the expected

rate of exceedance for a given ground motion value A, $\nu(Y > A)$ can be computed:

$$\nu(Y > A) = \sum \nu_i \int\int\int f_{m_i}(m) f_{r_i}(r) \Pr(y|m, r, \boldsymbol{X})\, dy\, dr\, dm. \tag{1.1}$$

In eq. 1.1, f_{m_i} and f_{r_i} are the probability densities for magnitude and distance, respectively, for source region i. ν_i is the activity rate of region i, and the summation over i is over all source regions.

Eq. 1.1 shows how PSHA deals with aleatory uncertainty, i.e. uncertainty that is intrinsic to nature (e.g. time, size and location of an earthquake, which ground motion it will generate). It has also become standard practice to include epistemic uncertainty (i.e. uncertainty about our state of knowledge) in PSHA in the form of logic trees (e.g. Kulkarni et al., 1984; Coppersmith and Youngs, 1988; Reiter, 1990; Bommer et al., 2005; Bommer and Scherbaum, 2005).

The preceding paragraphs highlight an important aspect in PSHA - there are considerable uncertainties associated with it. These uncertainties are both aleatory as well as epistemic (i.e. which model to use, how certain can the parameters be estimated etc). Aleatory uncertainty controls the shape of the hazard curve, while epistemic uncertainty leads to alternative hazard curves (a hazard curve distribution) (Abrahamson and Bommer, 2005). From a statistical point of view, the difference is that aleatory uncertainty is integrated out, while epistemic uncertainty is retained in the calculation of hazard curves.

In this context, ground-motion uncertainties (both aleatory and epistemic) have the largest effect on the results of PSHA (Toro, 2006), especially for low annual exceedance frequencies, which are important for critical facilities such as nuclear power plants. Here, it is worth quoting Norm Abrahamson from his keynote lecture at the 2006 First Conference on Earthquake Engineering and Seismology: *Engineers don't design for earthquakes, they design for ground motion.*[1]

Both the above statement as well as the observation by Toro (2006) highlight the importance of the ground motion domain and GMMs in particular for PSHA. Therefore, it is important to investigate possible new directions for estimating GMMs and characterization of associated uncertainty. In this thesis, some new directions are pursued, which may shed some new light on the ground motion domain and the estimation of GMMs.

In technical terms, a ground motion model quantifies the conditional distribution of a ground motion intensity parameter Y (the target value), given a set of earthquake and site related parameters $\boldsymbol{X}$ (the predictors). $\boldsymbol{X}$ contains parameters that influence (or are thought to influence) ground motion, e.g. the magnitude, distance, style-of-faulting, or the average shear wave velocity in the upper 30 m, V_S30. Y is usually assumed to be log-normally distributed, with the median being a function f of the predictors $\boldsymbol{X}$. This leads to the following model

$$\log Y \sim \mathcal{N}(\mu = f(\boldsymbol{\theta}, \boldsymbol{X}), \sigma), \tag{1.2}$$

where σ describes the total ground motion variability, which is usually decomposed into between-event variability τ and within-event variability ϕ, $\sigma = \sqrt{\tau^2 + \phi^2}$ (however, see Al Atik et al., 2010). $\boldsymbol{\theta}$ denotes the parameters of f.

[1] Along the same lines, Frank Scherbaum repeats the following statement over and over in his university courses on PSHA: *Don't think about earthquakes, think about ground motion.*

There is considerable uncertainty about all aspects of eq. 1.2, notwithstanding σ. For example, different published GMMs employ quite different functional forms for $f(\boldsymbol{\theta}, \boldsymbol{X})$. An example is the NGA project (Next Generation of Attenuation relations, see Power et al. (2008)), which resulted in five models (Abrahamson and Silva, 2008; Boore and Atkinson, 2008; Campbell and Bozorgnia, 2008; Chiou and Youngs, 2008; Idriss, 2008) that have quite different functional forms, even though they are based on the same dataset (Chiou et al., 2008). Related to the uncertainty about the functional form is parameter uncertainty. Due to limited amount of data, the parameters $\boldsymbol{\theta}$ of a GMM can only be estimated with uncertainty.

Another source of uncertainty is the set of predictor variables, $\boldsymbol{X}$. All published models include a predictor for earthquake size (i.e. magnitude) and source-to-site attenuation (i.e. distance). Newer models also include the style-of-faulting as a predictor for ground motions (Bommer et al., 2003). Site effects are usually characterized by a simple categorical scheme (often "STIFF SOIL", "SOFT SOIL", "ROCK") based on V_S30 or V_S30 directly. The usefulness of V_S30 as a proxy for site amplification has been questioned lately (Castellaro et al., 2008; Gallipoli and Mucciarelli, 2009). There are several other potential predictor variables that are either used or investigated, such as depth-to-the-top-of-rupture (e.g. Abrahamson and Silva, 2008; Campbell and Bozorgnia, 2008; Chiou and Youngs, 2008) or directivity (Spudich and Chiou, 2008). The NGA dataset contains information on more than 20 potential predictor variables (some of them probably redundant). Overall, it is not entirely clear which parameters influence ground motion and should therefore be included in the set of predictors $\boldsymbol{X}$. This is tied to uncertainties about the functional form - if we do not know which parameters to include, we also do not know their exact relation to ground motion. Furthermore, the predictor variables carry uncertainties as well - they are not measured error-free.

More on the application side, there is uncertainty about which model to use in a particular PSHA. This pertains to the selection of the models (Cotton et al., 2006; Bommer et al., 2010), as well as to ranking them/deciding on logic tree weights (Bommer and Scherbaum, 2005; Bommer et al., 2005; Delavaud et al., 2009; Sabetta et al., 2005; Scherbaum et al., 2004a, 2009).

In this thesis, some aspects of these uncertainties are addressed, as well as some other interesting issues in the ground motion domain. Many of the approaches used are influenced by the fields of machine learning and artificial intelligence. In particular, graphical models (e.g. Koller and Friedman, 2009) are employed, which are a powerful and flexible tool for reasoning under uncertainty.

The main body of the thesis comprises five papers, all of which are published in or submitted to internationally peer reviewed journals. The first two papers are similar with respect to them both being about what can be learned from the data at hand - what features does the dataset support, where are its limitations. Thus, these papers are concerned with epistemic uncertainty about the functional form of f (cf. eq. (1.2)). The third and fourth paper both address parameter uncertainty of GMMs using a Bayesian approach and graphical models. They both use the same 'basic' model, which is extended in two ways to account for different complexities (correlations between different ground motion intensity parameters and regional differences), thus showing the flexibility of graphical models. The last paper is slightly different in focus and introduces Naive Bayes as a tool to convert instrumental ground motion parameters to seismic intensities, which can provide important information for the selection of GMMs in PSHA in regions where instrumental

data is sparse (Scherbaum et al., 2009; Delavaud et al., 2009). In the subsequent sections, a short overview of each paper is given. The full papers are reprinted in chapters 2 to 6.

1.2 Deriving Empirical Ground-Motion Models: Balancing Data Constraints and Physical Assumptions to Optimize Prediction Capability

> **Kuehn, N. M.**, F. Scherbaum, C. Riggelsen (2009). *Bull. Seism. Am.* **99**, 2335-2347.

This paper presents an investigation into the predictive capabilities of GMMs. Therefore, generalization error as a measure for predictive potential, as well as cross-validation as a tool to estimate it, is introduced. Based on cross-validation and the most comprehensive strong motion dataset currently available, the NGA dataset, a new GMM is learned. This new GMM employs polynomials of the predictor variables, whose degrees are optimized to yield low generalization error. Low generalization error prevents the model from being overfit, and the flexibility of the polynomials allows to model the characteristics of the data. This procedure reveals features that are inherent in the data, as well as data ranges that are only poorly represented in the dataset. Thus, it can be used to investigate uncertainty about the functional form of a GMM.

To ensure that the model, which is not based on physical considerations, does not violate our understandings of the physical foundations of the ground motion domain, an equivalent stochastic model (see Boore, 2003) is calculated using the method of Scherbaum et al. (2004b). The inverted stochastic model is similar to published stochastic models of Western North America.

1.3 Modeling the Joint Probability of Earthquake, Site, and Ground-Motion Parameters Using Bayesian Networks

> **Kuehn, N. M.**, C. Riggelsen, and F. Scherbaum (2010). *Bull. Seism. Am., in press*

In this paper, the possibility of learning Bayesian networks from data is investigated. Bayesian networks are powerful tools for reasoning under uncertainty and have been used in several decision support systems, also in PSHA (Bayraktarli et al., 2006). First, findings about learning from synthetic datasets are presented. Subsequently, a Bayesian network is learned on the NGA dataset. Both the structure and the parameters of the network are learned. Learning the structure allows to assess all (in)dependencies between variables in a dataset, thus addressing the question of which parameters directly influence ground motion. In this paper, the network is learned with minimum prior assumptions. Hence, the result is a representation of which (in)dependences are currently supported by the data.

1.4 A Bayesian Ground-Motion Model with Correlation of Ground Motion Intensity Parameters

Kuehn, N. M., C. Riggelsen, F. Scherbaum, and T. I. Allen (2010). *submitted to Bull. Seism. Am.*

In this paper, a Bayesian GMM is developed. The Bayesian approach allows to assess the uncertainty in the parameters of a GMM ($\boldsymbol{\theta}$ in eq. 1.2) in a probabilistic way. It also makes it easy to update the model once new data is available via Bayes' rule. The model is developed as a probabilistic graphical model, which makes it easy to assess conditional (in)dependence between parameters, and as a concept of the data generating process allows intuitive insight into the model. The model directly estimates the covariance structure between different ground motion intensity parameters during the learning phase, which are commonly assumed independent.

1.5 A Bayesian Hierarchical Global Ground-Motion Model To Take into Account Regional Differences

Kuehn, N. M., F. Scherbaum, C. Riggelsen, and T. I. Allen (2010). *submitted to Bull. Seism. Am.*

In this paper another Bayesian GMM is presented. The core of the model is the same as for the one of the previous section, but is enhanced in a different way. In the model, an individual GMM is learned for different regions, but the parameters of each model are not independent, but connected by global hyperparameters. That way, data from one region helps to estimate the parameters in the other regions. All parameters are estimated using Bayesian inference to account for their respective uncertainty.

1.6 A Naive Bayes Classifier for Intensities Using Peak Ground Velocity and Acceleration

Kuehn, N. M., and F. Scherbaum (2010). *Bull. Seism. Am., in press*.

In this paper, a naive Bayes classifier to convert instrumental ground motion parameters to seismic intensities is learned. This approach is an alternative to traditionally employed regression models, which neglect the ordinal nature (discrete states) of seismic intensities. By contrast, the naive Bayes classifier estimates a discrete probability distribution of intensities given the instrumental ground motion parameters. This is important to proper account for uncertainty when a conversion from PGA or PGV to seismic intensities is needed.

1.7 Other Publications

In addition to the abovementioned publications, I also took part in the following publications, which are concerned with related topics, but are not part of the thesis:

- Delavaud, E., F. Scherbaum, **N. Kuehn**, and C. Riggelsen (2009). Information-Theoretic Ground-Motion Model Selection for Seismic Hazard Analysis: An Applicability Study Using Californian Data, *Bull. Seism. Soc. Am.* **99**, 3248-3263.

  This paper is an application of the method of Scherbaum et al. (2009) to select GMMs to be used in a PSHA based on available instrumental and intensity data. My contribution was to implement the GMMs, and in the discussion of the results.

- **Kuehn, N. M.**, C. Riggelsen, and F. Scherbaum (2009). Facilitating Probabilistic Seismic Hazard Analysis Using Bayesian Networks, *7th Workshop on Bayes Applications, UAI/ICML/COLT 2009*.

  This conference paper is a short version of Kuehn et al. (2010), which is enhanced by an investigation about using Bayesian networks as a tool to perform sensitivity analysis of PSHA.

- Scherbaum, F., **N. M. Kuehn**, M. Ohrnberger, and A. Koehler (2010). Exploring the Proximity of Ground-Motion Models Using High-Dimensional Visualization Techniques, *Earthquake Spectra, in press.*

  In this paper, the similarity of GMMs is investigated using high-dimensional visualization techniques such as Sammon's Maps (Sammon, 1969) or self-organizing maps (Kohonen, 2001). This can be used as guidance/prior information when setting weights on the logic tree in PSHA. My contribution was to implement the GMMs and in discussing/interpreting the resulting maps/networks.

- Al Atik, L., N. Abrahamson, F. Cotton, F. Scherbaum, J. Bommer, and **N. Kuehn** (2010). The Variability of Ground-Motion Prediction Models and its Components, *submitted to Seism. Res. Let.*

  This paper presents a unifying notation for different components of ground motion variability – this is necessary in particular when the ergodic assumption is removed (Anderson and Brune, 1999; Walling, 2009). In this case, the total variability can be partitioned into several components, both aleatory and epistemic. At the moment, there exist several different notations for these components, which makes communication difficult and confusing. Hence, a unifying notation is warranted. My contribution was to take part in discussion about this notation.

2

DERIVING EMPIRICAL GROUND-MOTION MODELS: BALANCING DATA CONSTRAINTS AND PHYSICAL ASSUMPTIONS TO OPTIMIZE PREDICTION CAPABILITY

Kuehn, N. M., F. Scherbaum, and C. Riggelsen
Bulletin of the Seismological Society of America **99**, 2335-2347.

Empirical ground-motion models used in seismic hazard analysis are commonly derived by regression of observed ground motions against a chosen set of predictor variables. Commonly, the model building process is based on residual analysis and/or expert knowledge/opinion, while the quality of the model is assessed by the goodness-of-fit to the data. Such an approach, however, bears no immediate relation to the predictive power of the model, and with increasing complexity of the models is increasingly susceptible to the danger of overfitting. Here, a different, primarily data-driven method for the development of ground-motion models is proposed, which makes use of the notion of generalization error to counteract the problem of overfitting. Generalization error directly estimates the average prediction error on data not used for the model generation and is thus a good criterion to assess the predictive capabilities of a model. The approach taken here makes only few a-priori assumptions. At first, PGA and response spectrum values are modeled by flexible, non-physical functions - polynomials - of the predictor variables. The inclusion of a particular predictor and the order of the polynomials are based on minimizing generalization error. The approach is illustrated for the NGA dataset. The resulting model is rather complex, comprising 48 parameters, but has considerably lower generalization error than functional forms commonly used in ground-motion models. The model parameters have no physical meaning, but a visual interpretation is possible and can reveal relevant characteristics of the data, for example the

Moho bounce in the distance scaling. In a second step, the regression model is approximated by an equivalent stochastic model, making it physically interpretable. The resulting resolvable stochastic model parameters are comparable to published models for Western North America. In general, for large datasets generalization error minimization provides a viable method for the development of empirical ground-motion models.

2.1 Introduction

The accurate prediction of ground motion for future earthquake scenarios is a key issue in any seismic hazard analysis. Most popular in this context is the use of empirical ground-motion attenuation relations derived by regression analysis from observed ground motion values such as peak ground acceleration (PGA) or response spectral values against a chosen set of predictor variables. In recent years there has been a growing trend to ever more complicated models, both in the number of predictor variables and the functional forms employed. Whereas earlier empirical ground-motion models typically included only magnitude and distance as predictors, most current models also take into account site effects (either by local geology or local shear wave velocity V_S30) and fault mechanisms (e.g. Bommer et al., 2003) (for a compilation and description of early and recent ground-motion models, see Douglas (2003)). In addition, the functional forms have become more complex as well, for example to account for a magnitude-dependent distance decay of ground motions (e.g. Anderson, 2000). Recent ground motion models now incorporate more than 8 predictor variables (Abrahamson and Silva, 2008; Campbell and Bozorgnia, 2008; Chiou and Youngs, 2008), including parameters such as the depth-to-the-top-of-the-rupture and terms for nonlinear soil-amplification (Choi and Stewart, 2005). Furthermore, the influence of additional effects such as directivity is investigated (Spudich and Chiou, 2008).

Exploring new potential predictor variables for ground motion is very important, since including previously unmodeled effects can help to increase the quality of ground-motion prediction for seismic hazard analysis. In this light, the goal of a more complicated model is to improve its predictive power over a more simple one. However, there are two problems adhering to the trend to more complicated models. First, more complex models are more susceptible to the danger of overfitting, i.e. modeling more spurious details of the sample than are supported by the data generating process. As a second, related problem, the exact physical relationships between many of the predictor variables and the target ground motion parameters are unknown. These issues are often treated quite informally in engineering seismology. Here, the quality of a model and the appraisal of a new predictor variable are typically based on minimized residuals and residual plots. However, residuals are not a good indicator to assess the predictive performance of a model, i.e. how well it is able to generalize its underlying dataset and predict new, independent data.

In this paper, we present a strategy for the development of ground-motion models that is directed towards the optimization of the model's predictive power, which could be considered the most important quality criterion for ground-motion models in the context of seismic hazard analysis. This approach follows the "intelligent data analysis" philosophy commonly advocated in the fields of machine learning and artificial intelligence. It makes relatively few physical assumptions but rather "lets the data speak". This makes it quite flexible, and it can be used to explore potentially new predictor variables as well as to gain insight into the relationships between the predictor and

target variables, thus helping to select functional forms that capture the characteristics of the data. Hence, this paper is in line with other works that explore new approaches for the development of ground-motion models (e.g. Ahmad et al., 2008; Anderson and Lei, 1994; Graizer and Kalkan, 2007; Tavakoli and Pezeshk, 2007).

To estimate the predictive power of a model, which is the quality criterion we aim to optimize, we use the notion of generalization error, a tool from the field of machine learning that represents the average prediction error of a model on an independent test dataset. This makes it a better criterion than minimized residuals to assess the quality of a model that is used for predictive purposes. Generalization error can be estimated by cross-validation (Stone, 1974), an easily understandable and widely used method in machine learning/artificial intelligence.

The proposed model building strategy optimizes generalization error to develop a model with good predictive power. In this context, we make only very few a-priori assumptions about the functional forms of the models considered, since these can impose strong constraints. By contrast, we rather model the dependence of the ground motion parameter on the predictors with simple but flexible functions, using cross-validation as a correction tool to avoid overfitting. Therefore, we start out with a simple model and add terms based on their ability to reduce generalization error. This approach results in a model that is neither underfit nor overfit to the data, thus emphasizing those characteristics that are actually supported by the data. However, this comes at the cost of at first having no physical interpretable parameters, due to the use of flexible, non-physical functions in the development of the model. In the machine learning literature, such a model is called "subsymbolic". However, a visual interpretation of the model can give insight into the relationship s between the target and the predictors, while the control of the model building process by the principle of generalization error minimization ensures good predictive power.

In a second step, we aim at achieving physical consistency of the regression model. Therefore, it is approximated by a so-called stochastic model (Boore, 2003), using the method of Scherbaum et al. (2006). This way, the regression model is made physically interpretable - the step corresponds to a transition from a "subsymbolic" to a "symbolic" interpretation.

We illustrate our approach on the dataset compiled for the next generation of ground motion attenuation models (NGA) project (Power et al., 2008), which is the best strong motion dataset currently at hand. For the development of the non-physical, subsymbolic model we use polynomials as basis functions, since these are simple, flexible and easy to understand. It is not our objective in this paper to develop a full-fledged alternative NGA model, but rather to present an extension to traditional modeling strategies that pursues the philosophy of "intelligent data analysis" advocated in the fields of machine learning and artificial intelligence.

The paper is organized as follows: In the next section, we briefly outline the task of model selection and introduce the concept of generalization as well as cross-validation. The subsymbolic model is developed in section 2.4.1 and subsequently physically interpreted in section 2.4.2 by means of the stochastic model. The results are discussed in section 2.5. Finally, we draw conclusions and sketch ideas for future research.

2.2 Model Selection, Generalization and Cross-validation

Model selection constitutes a very important step in many data analyzing processes. Typically, data is gathered in an experiment, and from this data one wants to draw inferences about the properties of the system under study (e.g., nature) and make predictions of future observations (e.g. Breiman, 2001a). Hence, a model is estimated from the data yielding an approximation of the underlying data generating process. This poses the question of what exactly makes a model good with respect to the data.

There are several criteria by which the quality of a model may be judged, and which one is chosen depends on the nature of the application. For example, if the goal is to get information about the system under study, one will seek physical interpretability of the model. In other situations, the goal may be to obtain a very simple model or one that is well suited to predict new data. Hence, the quality criteria in these cases would be parsimony and predictive power, respectively. If a good model is used, the information we infer will be reliable with respect to the quality-criteria that may apply. On the other hand, a bad or incorrectly specified model will almost certainly lead to wrong conclusions about the system under study. In this light, model selection can be described as "estimating the performance of different models in order to choose the (approximate) best one" (Hastie et al., 2001). However, the performance of a model is subject to the quality criterion we apply. A model with high predictive power is not necessarily physical interpretable or vice versa. In the context of ground motion and seismic hazard analysis, the issue of prediction is of primary concern: a bad prediction renders a ground motion model close to useless, no matter whether parsimony or interpretability is well-catered for in the model. Hence, we judge the quality of our ground motion model primarily in terms of its predictive power.

Model selection with the aim of optimizing the predictive performance has several consequences: First, the model needs to generalize the data. This means that spurious details or noise are not modeled, since this will degrade the predictive performance of a model (the very definition of noise is that it is not part of the underlying data generating process and varies on a per-observable basis). Hence, the model cannot be overly complex. Such a model that is too adapted to the dataset is referred to as an overfit model in the literature. On the other hand, a model with good predictive power cannot be too sparse, i.e. too simple. In this case, the model becomes too inflexible and is not able to capture the important characteristics of the data, and the predictive performance declines. This is referred to as underfitting. Hence, a model that is optimized with respect to its predictive performance is carefully balanced between being too simple and too complex, given the underlying dataset.

Existing ground-motion models are usually assessed using the goodness-of-fit to the data (minimized residuals). However, this may not be the preferred approach, because this methodology does not take explicit care of the predictive performance. We need a principled and systematic model selection technique where the predictive power of the model is the driving force; in other words: we want to optimize the generalization performance of a model. The concept of generalization directly relates to the predictive power of a model and describes how well a model can predict new data that was not used in the derivation process.

For example, suppose we have a target variable Y, a vector of predictors X and a predictive model $f^*(X)$ that was learned from a training sample (x_i, y_i) with $i = 1 \dots N$. Differences

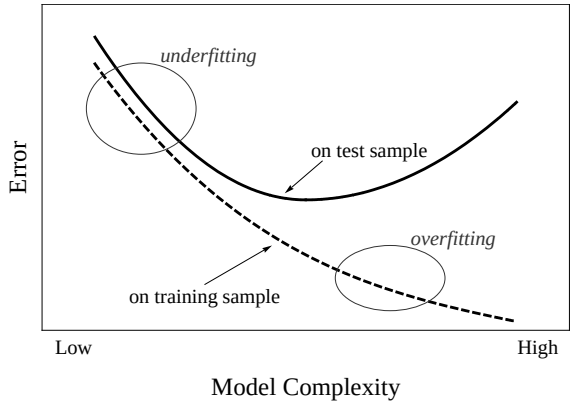

Figure 2.1: Typical behavior of the prediction error on the test sample and training sample error as the model complexity is varied.

between $f^*(X)$ and Y are measured by a loss function $L(Y, f^*(X))$, which is typically either absolute or squared error. Then the training error is defined as the average prediction error over the training sample

$$TE = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f^*(x_i)) \tag{2.1}$$

and is a measure of goodness-of-fit of the model to the data. By contrast, the generalization error is defined as the average prediction error over an independent test sample, $j = 1 \dots M$

$$GE = \frac{1}{M} \sum_{j=1}^{M} L(y'_j, f^*(x'_j)). \tag{2.2}$$

Hence, while training error measures how well a model is adapted to the data, generalization error directly estimates the error one is going to make when trying to predict unknown (future) data. Figure 2.1 shows schematically the typical characteristics of training and generalization error, dependent on the model complexity.

Often, however, data is scarce and valuable, and splitting the dataset is not an option. Several other methods exist that provide estimations to the generalization error, both theoretical, e.g. information-theory (Akaike, 1974; Burnham & Anderson, 2002), Bayesian methods (Schwarz, 1978), and empirical, e.g. bootstrap (Efron, 1979). A very simple, easy-to-use empirical method to estimate the generalization error is cross-validation, which we use in this work. Cross-validation (Stone, 1974; Mosteller & Tukey, 1977) is an easily understandable, generic and widely-used method that directly estimates generalization error. Here, the whole dataset is used for learning and testing the model via efficient sample re-use.

For cross-validation, in a first step the dataset is split into k roughly equal-sized subsets. Then, the model is fit to $k - 1$ subsets and the prediction error is calculated for the subset that was not used for fitting the model. This procedure is then repeated for all k subsets, and the k prediction error estimates are combined to give an estimate of the generalization error (this is known as k-fold

cross-validation).

In detail, the cross-validation estimate of prediction error can be estimated as follows: If $f^{-k}(X)$ is the fitted model with the kth subset left out, then the prediction error estimated by cross-validation can be calculated by

$$GE_{CV} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f^{k(i)}(x_i)). \tag{2.3}$$

Hence, to select the best model, GE_{CV} is calculated for all models under consideration via equation (2.3), and the one with the lowest estimate is chosen. Cross-validation requires the specification of one parameter, namely k. In general $k = 5$ or $k = 10$ are considered to be a good compromise (Hastie et al., 2001). The choice $k = N$, which is known as leave-one-out cross-validation, would lead to an unbiased estimate of prediction error. Depending on the size of the dataset, however, it can be computationally quite intensive, since the model has to be learned N times. In this work, we use a value of $k = 10$.

It is important to note that we use generalization error as our only criterion to judge the quality of a model. This means that even if a parameter might appear statistically insignificant, it is still retained in the model if it results in a lower generalization error.

2.3 Data Selection

The dataset underlying our analysis is the one compiled for the NGA project (Power et al., 2006). It consists of 3,551 recordings from 173 earthquakes from regions including California, Taiwan, Alaska and Europe and is the best strong-motion dataset currently at hand for seismically active regions. The dataset includes numerous meta-data for each recording such as finite fault models, earthquake source parameters, local geology and much more. In the following, we describe the criteria by which we select the records for subsequent analysis.

Despite it being the most comprehensive strong motion dataset that is currently available, a significant amount of meta-data is missing in the NGA dataset. For example, only 63 of the earthquakes in the NGA dataset have a finite fault model. Hence, records for earthquakes without a finite-fault model are lacking information e.g. on Joyner-Boore distance or depth-to-the-top-of-the-rupture.

Incomplete data poses a great challenge to any kind of principled statistical analysis. Incompleteness complicates the functional form of various (e.g., likelihood-based) statistical estimates and distributions, and the usual statistical analysis methods become non-trivial. Furthermore (Rubin, 1976) the nature of the process that gave rise to incomplete data needs to be analyzed before commencing, since this is very key to the question whether we can perform valid unbiased statistical inference based on the incomplete data at all. In a nutshell, when the presence of an unobserved/missing item is unrelated to its own or other missing items would-have-been values, then statistical inferences based on the incomplete data are unbiased, i.e., the estimates one makes based on the incomplete data pertain to the underlying data generating process. Formally speaking, we can distinguish two kinds of missing data mechanisms enabling us to perform unbiased statistical analysis: The data is missing at random (MAR) if an item that is missing depends on

observed values of other variables (and not on any missing variables). The data is missing completely at random (MCAR) if an item is independent of observed variables (and independent on any missing variables). Unbiasedness does not mean that the analysis becomes easy, since the functional complications and non-trivial solutions remain.

In e.g., Schafer (1997) or Allison (2002) various approximations and methods are introduced that try to ease the burden of performing statistical analysis when the missing data mechanism is MAR or MCAR. An important point to note is that it is not possible to test whether the missing data mechanism is MAR/MCAR or not using the incomplete dataset alone. It is thus necessary to have knowledge about how and why incompleteness came about in the first place.

We assume that the missing data mechanism for the NGA dataset is MAR. The reason that it is not MCAR is that a larger earthquake is more likely to have a finite rupture model. Hence, the probability that there is missing data regarding the depth-to-the-top-of-the-rupture or Joyner-Boore distance is higher for earthquakes with smaller magnitudes. The MAR assumption enables implies that it is indeed possible to perform unbiased statistical inference. However, there are still functional and computational complexities to be solved that come about with missing data. We evade these complexities by choosing a very simple way to deal with missing data: we include only those records that have complete information on all the predictors that we want to consider, essentially making the incomplete data look complete. This is known as listwise deletion or complete case analysis. Strictly speaking, this method of dealing with missing data yields unbiased results only under the MCAR assumption and is therefore not unproblematic in the present setting. Furthermore, one obvious caveat is statistical inefficiency: not all available data is used, and we effectively have a smaller sample size than initially. However, in general this crude method is quite insensitive to departures from the MAR assumption and might even perform better than more sophisticated methods in the context of regression (Allison, 2002).

We consider five potential predictor variables in our model: moment magnitude, Joyner-Boore distance, local shear wave velocity (V_S30), fault mechanism (based on the rake angle) and depth-to-the-top-of-the-rupture, hereafter referred to as rupture depth. Hence, all records that do not have complete information on these parameters are excluded from further analysis. The majority of records that are not considered are from earthquakes that do not have a finite fault model, so no information on Joyner-Boore distance and depth-to-the-top-of-rupture is available. These earthquakes also include those that are missing estimates of moment magnitude and rake angle. Additionally, 6 records are excluded due to missing values for V_S30. The target variables of our model the are horizontal spectral ordinates (including PGA). The two horizontal components are combined via the geometric mean. Hence, we exclude all records for which only one component is available.

The model we develop is supposed to represent free-field ground motions from active tectonic regions. Hence, we include only those records in the analysis that are representative of free-field conditions. The classification of a free-field recording is based on the Geomatrix classification C1 (see e.g. Abrahamson and Silva, 1997) for the corresponding station. As non free-field sites we consider those stations which have a GMX C1 code C, D, E, F and G.

Additionally we excluded some records from the Chi-Chi-sequence in Taiwan, following Abrahamson and Silva (2008). Several stations in Taiwan house more than one strong motion instrument, and only those records from the newer instrument are included, as recommended by Lee et

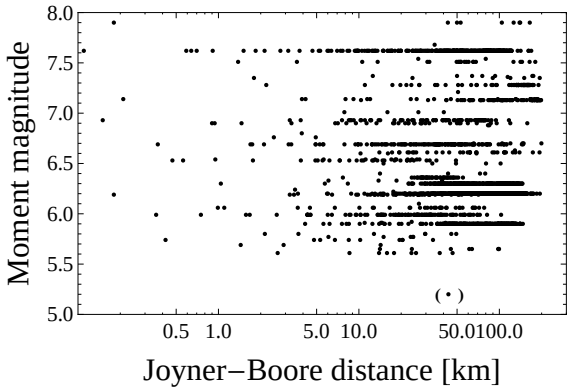

Figure 2.2: Magnitude-distance distribution of selected records. The record with $M_W = 5.2$ is excluded to avoid it playing a dominant role in the small magnitude range.

al. (2001). Furthermore, a couple of stations are classified as poor quality by Lee et al. (2001) and are therefore not included in the further analysis.

Finally, we do not include records with Joyner-Boore distances greater than 200 km. Due to their large source-to-size distance, they are of low engineering significance. Furthermore, excluding these records reduces possible correlations between magnitude and distance, making the distribution with respect to magnitude and distance more uniform. This way, also a possible bias due to different attenuation properties of the data-contributing regions is reduced.

In total, we select a subset of 2661 records from 61 earthquakes. The resulting magnitude-distance distribution is shown in Figure 2.2. As one can see, there is only one record from the smallest earthquake ($M_W = 5.2$), with the next larger earthquake having a magnitude of $M_W = 5.61$. This one record plays a dominant role for the model in the small magnitude range and is therefore excluded, leading to a dataset of 2660 records from 60 earthquakes.

2.4 Model Development

In this section we describe the model building process. This process consists of two steps: First, a simple, non-physical regression model is developed that is optimized for low generalization error. Second, a stochastic point-source model is calculated using the approach proposed by Scherbaum et al. (2006).

2.4.1 Regression Model

In this section we restrict attention to regression models. The model is derived for peak ground acceleration and spectral acceleration (5% of critical damping) for 39 periods between 0.01s and 3s. For the derivation of the response spectral models, we select only those records that have a highest usable period, as indicated by the NGA flatfile, above the current period. The target variable of the regression model is a response spectral ordinate (including PGA), which is taken as

Table 2.1: Classification of fault mechanisms based on their rake angle and corresponding number of earthquakes and records in the dataset under study

fault mechanism	rake angle	no. of earthquakes	no. of records
reverse	$30 < \lambda < 150$	19	1870
normal	$-150 < \lambda < -30$	11	49
strike slip	else	30	741

the natural logarithm of the geometric mean of the two horizontal components. Hence, by taking the logarithm we make (as almost all published ground motion models) the basic assumption that the ground motion parameter is log-normally distributed.

In contrast to most of the published ground motion models we take a radically different approach in the development of the functional form of the regression model. We make almost no a-priori assumptions regarding the relationship between the target and the predictors but let the regression equation evolve in a simple way based on low generalization error. In other words, we adhere to the general framework of model selection, where the constituents are: (a) a search space that defines the space of potential regression models that should be considered, (b) a scoring criterion that assigns each model in the search space a value depending on its quality, (c) a traversal strategy that more or less intelligently moves through the search space in a non-brute-force manner as to find the optimal model. This decomposition stipulates a (semi)-automatic approach to model selection. Only few restrictions are imposed on the functional form of the regression model, hence we "let the data speak". We consider this an essential feature of our approach.

The search space mentioned in point (a) consists of linear regression models with polynomial basis functions. The scoring criterion mentioned in point (b) is the GE approximated via cross-validation. Hence, we model the target variable as a linear combination of polynomials where the order of the polynomials is determined by low generalization error. The approach taken does without any Fourier spectra-based theory (as is usually employed in ground motion models) which is not one-to-one applicable to the response spectrum as was demonstrated by Cotton et al. (2008). Five earthquake- or site-related parameters are considered as predictor variables: moment-magnitude, Joyner-Boore distance, V_S30, fault mechanism and rupture depth. All of these parameters are known or have been suggested to influence ground motions. Nevertheless, except for magnitude and distance, which are known to have the greatest impact on ground motions, the inclusion of a particular predictor variable depends on its ability to reduce the generalization error. Joyner-Boore distance is used since some trials with a simple functional form have shown that it gives a smaller generalization error than, for example, rupture distance. The fault mechanism is based on rake angle - the classification scheme is depicted in Table 2.1.

The traversal strategy in point (c) starts off with a very simple model

$$f = a_0 + b_0 \ln(\sqrt{R_{jb}^2 + a_r^2}) + c_1 M_W^1 \tag{2.4}$$

In equation (2.4), we have made two assumptions: (a) that magnitude and distance appear in the

model, which is certainly reasonable, and (b) that the scaling of ground motions with distance is logarithmic.

Subsequently, new terms of higher order are added to the model based on generalization error. This is done as follows:

1. Individually add the terms $c_i\, M_W^i$, $e_i\, R_{jb}^i$, $f_i\, V_S30^i$, $g_i\, d_r^i$ and $b_i\, M_W^i\, \ln(\sqrt{R_{jb}^2 + a_r^2})$, where i denotes the next highest polynomial order for each term that is already in the model. M_W, R_{jb}, V_S30 and d_r denote moment magnitude, Joyner-Boore distance, average shear wave velocity in the upper 30 m and depth-to-the-top-of-the-rupture, respectively, while b_i, c_i, e_i, f_i and g_i are parameters to be estimated.

2. For each new model, calculate an estimate for the generalization error via 10-fold cross-validation.

3. Select the term which gives the lowest generalization error and add it permanently to the model.

The fault mechanism is not modeled as polynomials but as a simple function

$$fm = \begin{cases} a_{NM}, & \text{normal faulting} \\ a_{RV}, & \text{reverse faulting} \\ 0, & \text{else} \end{cases}, \tag{2.5}$$

and once this term is incorporated in the model, the fault mechanism is not considered further. In step one, we allow one interaction term, between magnitude and logarithmic distance. This is because several studies have found that the decay of ground motions with distance is magnitude dependent (e.g. Anderson, 2000). Obviously, one could think of many more combinations, but to keep things simple and computationally easy we decided against the incorporation of other cross-terms.

The steps outlined above are repeated until the generalization error of the current model is higher then the one of the base model [eq. (2.4)], and the model with lowest generalization error is selected as the final model. It is important to note that in step 2 described above, always a new term is selected to be added to the model, even if none of the terms decreases the generalization error. In this way, the effect of a predictor can be included even if it shows up only at a higher level.

The resulting model is rather complex, comprising in total 48 parameters. These break down into 14 for the magnitude, 5 for Joyner-Boore distance, 4 for V_S30, 6 for the rupture depth and 15 for the combination of magnitude and distance as well as the intercept parameter a_0, the pseudo-depth a_r and the two fault mechanism parameters. However, since the model has very low generalization error, it is not overfit with respect to the dataset and the quality criterion we apply. By a traditional analysis, some of the parameters do not appear statistically significant at a 5% level. Nevertheless, since 5% is an arbitrarily chosen value and since the exclusion of these parameters results in a higher generalization error, they are retained in the model.

Table 2.2: Comparison of the generalization error, computed with 10-fold cross-validation, for the polynomial function and a physical functional form commonly employed in ground motion models (e.g. Akkar and Bommer, 2007a,b) as well as a complex, overfit polynomial function with 58 parameters. The last column gives the standard deviation of the generalization error over the 10 subsets that were used by cross-validation.

Model	Generalization error	Standard deviation over 10 subsets
polynomial function	0.275	0.033
physical function	0.344	0.027
overfit polynomial function	0.6678	0.621

In Table 2.2, we compare the generalization error of the best model with polynomial basis functions with a simple, physical-based functional form that is commonly employed in ground motion models (e.g. Akkar and Bommer, 2007a,b):

$$\begin{aligned} \ln(PGA) &= a_1 + a_2 M_W + a_3 M_W^2 + (a_4 + a_5 M_W) \ln \sqrt{R_{JB}^2 + a_6^2} \\ &\quad + a_7 S_S + a_8 S_A + a_9 F_N + a_{10} F_R, \end{aligned} \tag{2.6}$$

where S_A and S_S are switches for soft soil and stiff soil and F_{NM} and F_{RV} are switches for normal and reverse faulting, respectively. To be able to compare the generalization errors obtained for the model with polynomial basis functions and the functional form of equation (2.6), we use natural logarithms in the latter. This ensures that the target variable is the same in both cases.

As one can see, the model with polynomial basis functions has considerably lower generalization error and is therefore expected to estimate ground motions from future scenarios more precisely. We also compare the two aforementioned models with a complex polynomial model with 58 parameters. The generalization error of this model is much higher than for the other two models which indicates that it is strongly overfit. Also shown in Table 2.2 is the standard deviation of the generalization error over the 10 subsets of cross-validation. It is slightly higher for the model with polynomial basis functions than for the physical-based model, which reflects the higher variability of the polynomials.

After having selected the functional form as described above, the model is learned for the whole dataset using the regression algorithm of Abrahamson and Youngs (1992) to account for intra- and inter-event correlations. The same functional form is used for all periods. Certainly the same algorithm could be applied to learn a best-generalizing model separately for each period. However, this does not result in a very large reduction of generalization error, so to keep the computational effort low we chose to retain the functional form found for PGA.

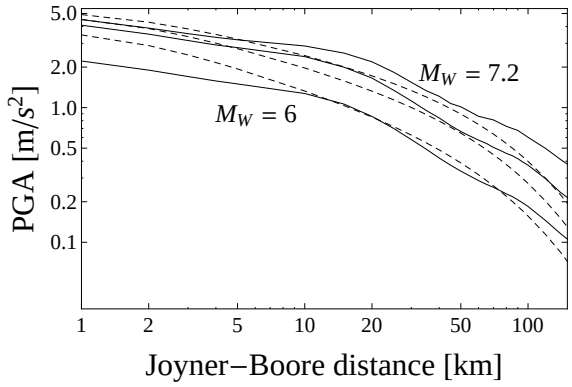

Figure 2.3: Comparison of the subsymbolic model (solid line) with the NGA model of Boore and Atkinson (2007) (dashed line) for PGA dependent on distance, for magnitudes $M_W = 6,\ 6.6,\ 7.2$. The comparison is made for rock sites ($V_S30 = 760m/s$) and reverse faulting earthquakes. For rupture depth in the subsymbolic model a value of 6.5 km is used, which corresponds to the median of the underlying dataset.

2.4.1.1 Results

Figure 2.3 shows a comparison of median PGA of the model with polynomial basis functions with the NGA model of Boore and Atkinson (2008). This model is chosen since it is regarded as being representative of the NGA models (Stafford et al., 2008), and uses - save rupture depth - the same predictor variables as the model learned above. Since the model of Boore and Atkinson (2008) uses a new measure for combining the two horizontal components, GMRotI50 (Boore et al., 2006), it is converted to geometric mean using the relationships of Beyer and Bommer (2006). One can see that both models show a similar scaling with magnitude but a different distance decay at small and large distances. At small distances ($R_{jb} < 5$ km), the subsymbolic model predicts smaller median PGA values than the model of Boore and Atkinson (2008). However, in this range the model is not well represented by data. This can pose a problem for a subsymbolic model, since the flexible basis functions require a good data coverage. Hence, care should be taken when interpreting the values of such a model in the limits of its data range.

At intermediate to large distances, the model with polynomial basis functions does not decay smoothly with distance. Instead, there is a range between about $R_{jb} = 50$ km to $R_{jb} = 90$ km where PGA decays less rapidly with distance. This effect can be associated with post-critical reflections of seismic waves at the Moho and has been found for eastern (Burger et al., 1987) and western North America (Campbell, 1991; Somerville and Yoshimura, 1990). Contrary to other ground-motion models that incorporate this so-called Moho bounce (e.g. Atkinson and Boore, 2006), the subsymbolic model does not include this effect a priori - it is solely a result of the model building, which relies completely on the dataset and generalization.

In Figure 2.4 we show the scaling of PGA in the subsymbolic model with magnitude, distance, V_S30 and rupture depth as partial dependence plots, introduced by Friedman (2001). Partial de-

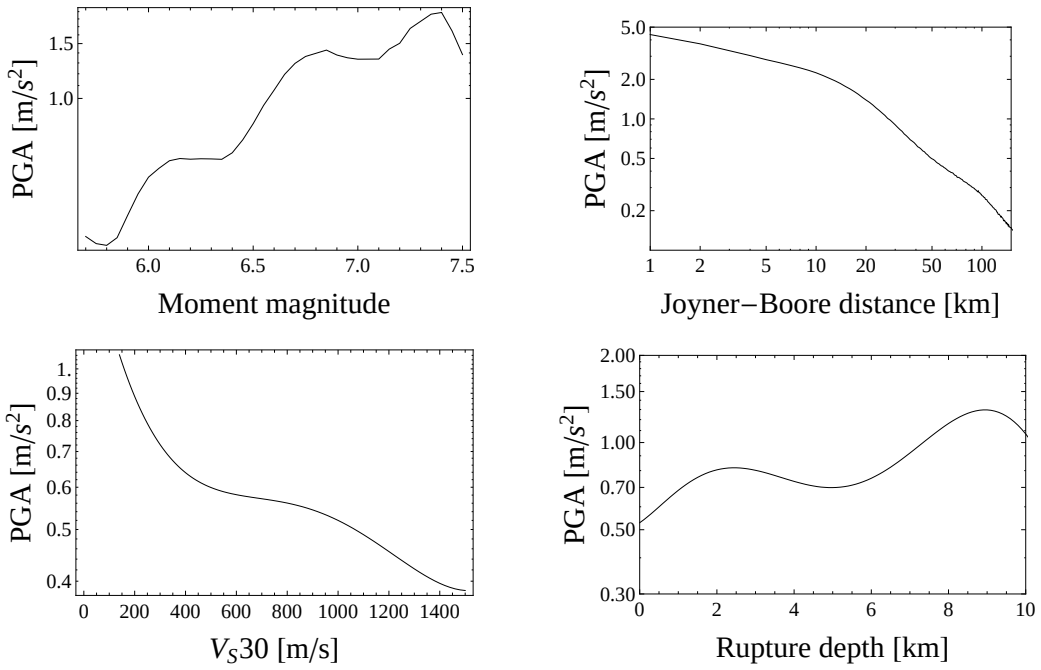

Figure 2.4: Partial dependence plots of the subsymbolic model for the effect of magnitude (top left), distance (top right), V_S30 (bottom left) and rupture depth (bottom right) on PGA.

pendence plots are a useful tool for examining the effect of a particular predictor x_S. The partial dependence function is defined as

$$f_S(x_S) = \frac{1}{N}\sum_{i=1}^{N} f(x_S, x_{iC}), \tag{2.7}$$

where x_S is the predictor of interest and x_{iC} is the ith observation of the other predictors in the data. For example, to compute the partial dependence for the effect of magnitude on PGA, we replace the magnitude in all records of our dataset by M_W, compute PGA for the whole dataset, and average these predictions. Thus, we get a function that is dependent on M_W, which is then plotted in Figure 2.4 (top left).

As one can see in Figure 2.4, the magnitude scaling of the subsymbolic model looks very complicated as a result of the high order of the magnitude polynomial. However, we stress that the model is not overfit. The behavior rather reflects characteristics of the magnitude dataset, which is not a representative sample of the underlying distribution. Nevertheless, the trend of the magnitude scaling is correct, and the partial dependence plot can aid in the interpretation of the magnitude-PGA relationship. In particular, Figure 2.4 might suggest a tri-linear scaling of PGA with magnitude, as used in the NGA model of Campbell and Bozorgnia (2008). Figure 2.4 also supports the hypothesis that events with a buried rupture lead to higher ground motions. Here, however, we face again the problem of a rather discrete underlying dataset. By contrast, this is not so much of

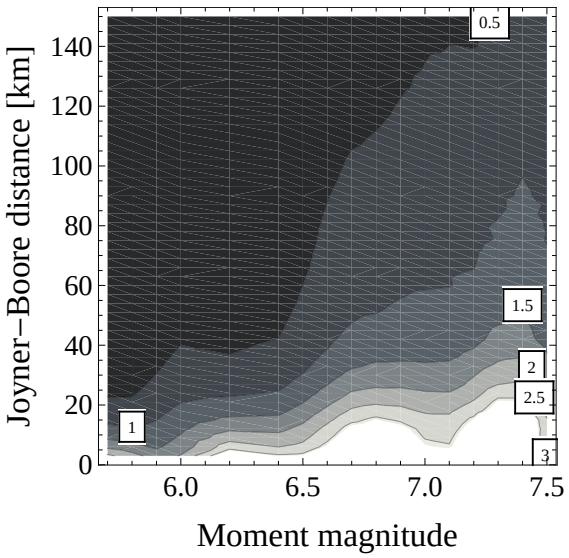

Figure 2.5: Partial dependence plot of the subsymbolic model for the effect of magnitude and distance on PGA. The values of PGA in m/s^2 are given in boxes.

a problem for Joyner-Boore distance and V_S30.

In the development of the polynomial model, we have allowed interaction terms between magnitude and distance to enter the model, since there is strong evidence that the ground motions of large earthquakes decay less rapidly with distance than those of small earthquakes (Anderson, 2000). Of the 47 terms comprising the model with polynomial basis functions, 15 are interaction terms, which already illustrates the importance of the interdependency between magnitude and distance. The effect of these interaction terms can be seen in Figure 2.5, which shows the partial dependence of PGA simultaneously on magnitude and distance. As one can see, with increasing magnitudes and distance, the contour lines are spaced further apart, reflecting a weaker attenuation with distance for large earthquakes, as is expected.

The strength of the model with polynomial basis functions (the subsymbolic regression approach) is its low generalization error, and that it can reveal the characteristics that are actually supported by the data, such as the Moho bounce in this case. On the other hand, it has no physically interpretable parameters, and by no means can be extrapolated. It can also have problems if the underlying dataset is not representative of the underlying data generating process. These issues are addressed in the next section.

2.4.2 Stochastic Model

In this section, we give a physical interpretation of the regression model learned in the previous section. Therefore, we invert for the (physical) parameters of the stochastic point source model - this corresponds to a transition from a subsymbolic to a symbolic model. The stochastic method, proposed by Boore (1983) following work by McGuire and Hanks (1980) and Hanks

and McGuire(1981), is a simple, yet powerful technique to simulate ground motions and response spectra based on simple seismological methods. It is widely used to predict strong ground motion in regions with only a scarce number of records, such as eastern North America. A detailed summary of the stochastic method as well as a comprehensive list of applications can be found in Boore (2003). The stochastic method also serves as the basis for the hybrid empirical method of Campbell (2003, 2004), which provides a methodological framework to transfer empirical ground-motion models from their host region to an arbitrary target region.

Scherbaum et al. (2006) have proposed a method to invert equivalent stochastic models for an empirical ground-motion relation. The inversion of the stochastic model parameters is done using a genetic algorithm (GA) (Goldberg, 1989). This approach also allows one to quantify parameter uncertainty, which is important to assess error propagation. Scherbaum et al. (2006) find a good agreement between the inverted stochastic model spectra and the empirical model spectra that were used as "data". For a complete description of the method, see Scherbaum et al. (2006).

We use the same method as Scherbaum et al. (2006) to derive equivalent stochastic models for the regression model with polynomial basis functions learned above. Therefore, first a set of polynomial model response spectra are generated for specific magnitude/distance combinations. These are then used as input for a genetic algorithm (probabilities of crossover and mutation set to 0.6 and 0.04, respectively) that finds the stochastic model parameters.

The spectra to be inverted are generated for magnitudes of 6, 6.5, 7 and 7.4. The Joyner-Boore distance ranges from 1 km to 150 km. Since there are only a few records with large magnitudes and small distances, these combinations are not used in the inversion. In the stochastic models, we tested hypocentral distance, rupture distance and the distance measure of Atkinson and Silva (2000), since Scherbaum et al. (2006) find that these distance measures give good results in the inversion process. To convert the Joyner-Boore distances of the polynomial model to the aforementioned distance measures, the method of Scherbaum et al. (2004b) is used. Since an expressively large set of free parameters would lead to a poorly constrained inverse problem, not all stochastic model parameters are inverted for. Some, such as density, shear wave velocity or radiation pattern, are kept fixed (see Table 2.3). For the source model, we use the two-corner point source model of Atkinson and Silva (2000), since trials have shown that it gives the lowest misfit.

The best fit is achieved with rupture distance as the distance measure in the stochastic model. The inverted model parameters of the best-fitting stochastic model are depicted in Table 2.3, together with the values of the western North America model of Campbell (2003, 2004). As one can see, the parameters are generally quite similar, indicating that there is a reasonable physical interpretation of the subsymbolic regression model by means of the regression model. Differences between the two depicted models are exhibited by the local site profile and the exponent of the frequency-dependent quality factor Q. However, when we rerun the inversion with these parameters fixed, we see that there is a trade-off between these two parameters and the quality factor Q_0, while the other parameters do not change very much. Hence, we conclude that it is not possible to resolve the local site profile and the path attenuation Q very well. This might be due to the fact that the underlying dataset aggregates records from different areas of shallow active tectonics.

The regression model spectra are shown, together with the five best fitted stochastic models, in Figure 2.6. As one can see, there is good agreement between the polynomial and the fitted spectra for small magnitudes. By contrast, for large magnitudes differences persist. However, this is not

Table 2.3: Parameters of the western North America model of Campbell (2003, 2004) and of the best-fitting stochastic model

Parameter	Campbell WNA	best-fitting model
Stress drop, $\Delta\sigma$ (bar)	100	71
Site attenuation, κ_0 (s)	0.04	0.037
Path attenuation, Q	$180f^{0.45}$	$171f^{0.92}$
Geometric attenuation	R^{-1}	$R^{-0.95+0.04(M_W-4)}$
	$R^{-0.5}$	$R^{-0.34+0.09(M_W-4)}$
		$R^{-0.68+0.13(M_W-4)}$
Length of Segments (km)	$40/\infty$	$50/84/\infty$
Local site profile (m/s)	620	1650

The rupture distance was used as distance metric for the stochastic model. Fixed parameters are $R_{\Theta\Phi}$ = 0.55, $V = 1/\sqrt{2}$, F = 2. Density and shear wave velocity are set to $\rho_S = 2700kg/m^3$ and $\beta_S = 3500m/s$. For a description of the parameters, see Scherbaum et al. (2006). The two-corner point source model of Atkinson and Silva (2000) is used.

unexpected, since the fault extension cannot be neglected for an earthquake with a magnitude of m=7.4. Hence, the point source assumption has its limits in this case.

2.5 Discussion and Conclusions

There are three points to discuss: generalization, taking a data-driven approach, and the resulting stochastic model parameters. We think that generalization is a very important aspect, if not the most important, when it comes to evaluating the quality of a ground motion model, since good predictive power is essential for ground-motion models in seismic hazard analysis. However, in almost all published ground-motion models the quality of a model is judged primarily by residual plots and the goodness-of-fit, even though as shown in Figure 2.1 training error is not a good measure of predictive power. In this context, it is important to note that the estimation of the generalization error is independent from the approach taken here. Using cross-validation or setting aside a test dataset works for any model. Hence, we believe that generalization error should play an important role when building a ground motion model. For example, it can be used to decide whether a new term should enter the model or not. However, even if the functional form for the ground motion model is determined completely independent, e.g. based completely on physical assumptions, generalization error should be estimated, since it provides a good measure of the quality of the model and how well it is able to predict new data that was not used for learning.

Generalization error is essential to the data driven approach we have taken here to find the functional form for our regression model, since it prevents the model from being overfit. The results of this approach have shown that it can be a useful alternative to traditional modeling

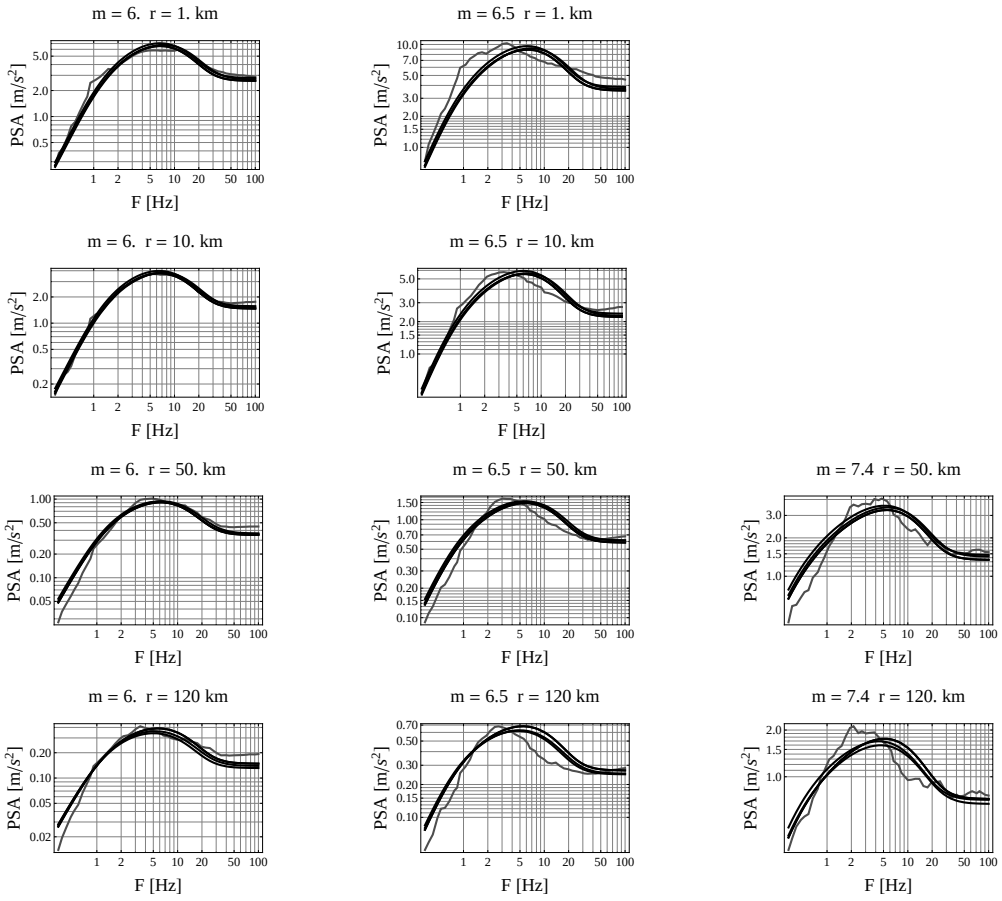

Figure 2.6: Model spectra for the five best-fitting stochastic models (black lines) and the regression model (gray line) for selected magnitudes and distances. For larger earthquakes and smaller distances the inversion is not carried out since here data is sparse.

strategies. By making few a priori assumptions, we also reduce the number of constraints on the model. This allows us to find new, unexpected behavior that is nevertheless supported by the data (such as the Moho bounce in the distance scaling, cf. Figure 2.3). As seen in our results, this can lead to lower generalization error, which is very important for seismic hazard analysis.

The data-driven approach works well in the case of the distance-scaling, which shows its potential. On the other hand, its limitations are shown by the magnitude scaling. Here, the dataset is not representative of the underlying distribution, which leads to a dependence of PGA on magnitude which is not physical (cf. Figure 2.4). However, except for magnitudes above m=7.4 we get an overall increasing trend that makes physical sense. Hence, in this case the polynomial model can aid in the selection of a good functional form for the magnitude scaling. The decrease of PGA with large magnitudes is a feature that is inherent in the NGA dataset, and is also observed in the NGA models. These deal with this issue by forcing the magnitude scaling to be positive (Boore, 2008, personal communication).

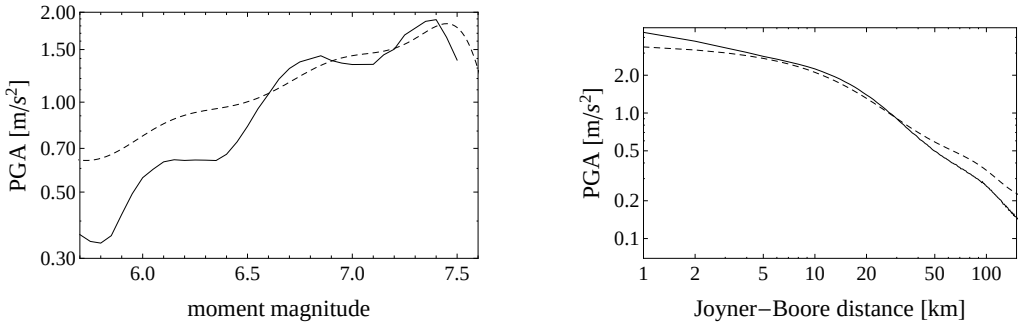

Figure 2.7: Partial dependence plot of the subsymbolic model learned in section 2.4.1 (black line) and a subsymbolic model learned on the dataset of the NGA model of Campbell and Bozorgnia (2008) (dashed line) for magnitude (left) and distance (right).

To illustrate the effect of different distributions of the predictors on the subsymbolic model, we show in Figure 2.7 the partial dependence plots for magnitude and distance for two models with polynomial basis functions - the one learned in section 2.4.1 and a preliminary one learned on the dataset that was used in the NGA model of Campbell and Bozorgnia (2008). Here we can see a rather similar distance scaling of the two subsymbolic models, while the scaling of PGA with magnitude is more different. This is due to the difference of the distributions of magnitude and distance underlying the analyses. As one can see in Figure 2.8, the magnitude distributions show greater discrepancies than the distance distributions, thus causing the differences seen in Figure 2.7.

A related issue is the behavior of the subsymbolic model at the boundaries of the data, where data is generally sparse. This already poses a problem when ground-motion models that are based on physical assumptions are derived, as was recently demonstrated by Bommer et al. (2007). Even though such physical functions can be believed to hold over a range that is wider than the data range, ground-motion models based on them should not be extrapolated. The subsymbolic model, however, can by no means extrapolated, since it can give physically completely unreasonable values outside the data range. This is due to the fact that, in order to capture the characteristics of the dataset, the basis functions are chosen to be rather flexible. This flexibility is not constrained beyond the borders of the dataset - if there is no data, the data cannot “speak”. We already encountered this behavior in Figure 2.3 at very short distances. However, this does not affect the model in ranges with good data coverage.

From the above considerations we see that it is important to carefully balance the predictive performance. Predictive power is not the only quality criterion one can apply for model selection. Physical interpretability or parsimony might be others. Here, we have built our regression model primarily on its predictive performance given the underlying dataset, since this is a crucial part for any ground motion model that is to be applied in seismic hazard analysis. This results in a model with considerable lower generalization error than a physical based model (cf. Table 2.2). However, physical considerations can supplement or constrain the approach in regions where data is scarce.

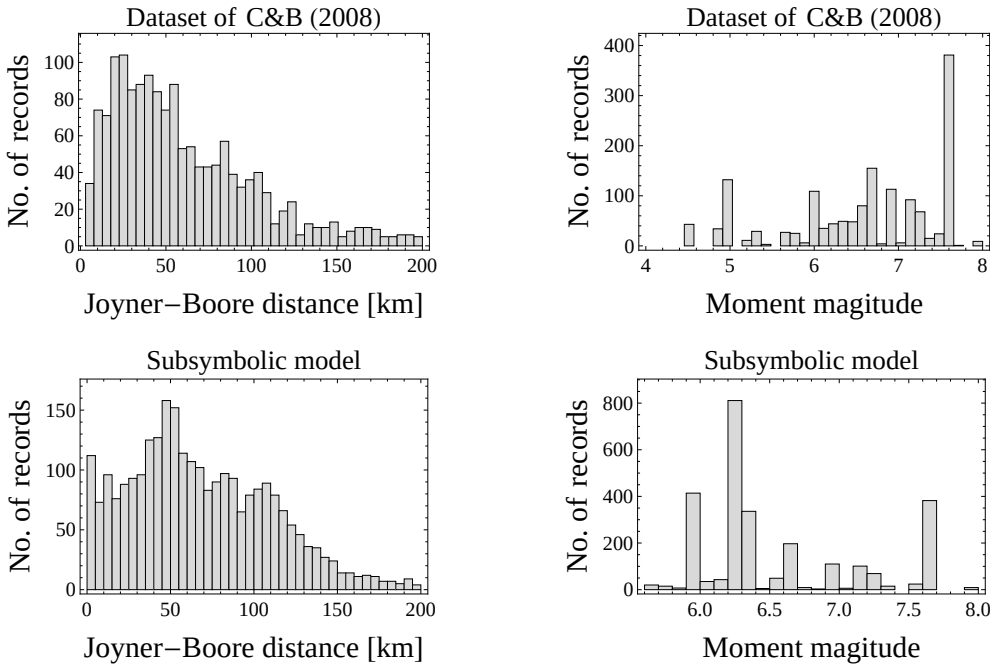

Figure 2.8: Distributions of distance (left) and magnitude (right) for the datasets underlying the subsymbolic model (bottom) and the NGA model of Campbell and Bozorgnia (2008) (top).

In this work, we have used polynomials in the development of the subsymbolic model. We have done so because polynomials are quite flexible and easy to understand, so they serve well to illustrate the method. However, the approach outlined in this paper is not confined to polynomials. One could think of piecewise linear models, splines or neural networks in the development of a subsymbolic model with optimized generalization error to reveal the characteristics of the data. Subsequently, such a model can be interpreted physically.

We have seen already in the analysis of the partial dependence plots of the regression model that a physical interpretation is possible. The general trend of the scaling of all predictor variables makes physical sense. The regression model also shows that earthquakes with a buried rupture generate higher ground motions than those that rupture the surface (cf. Figure 2.4, lower right), an effect that is also incorporated in the NGA models of Abrahamson and Silva (2008), Campbell and Bozorgnia (2008) and Chiou and Youngs (2008). When we go a step further and interpret the regression model in terms of the stochastic model, we also get a reasonable physical representation. The stochastic model parameters are generally similar to the parameters of the western North America model of Campbell (2003, 2004). In addition, at intermediate distances the fitted stochastic model shows a decrease in distance scaling, which was already inherent in the regression model and that we attribute to the Moho bounce. Furthermore, the inverted stochastic model shows a magnitude dependent decay with distance for larger distances (cf. Anderson, 2000).

An obvious next step would be to develop full fledged ground motion models using the model building strategy presented in this article. This would also involve the analysis of aleatory vari-

ability, the discussion of which is beyond the scope of the present work. For example, it would be interesting to compare the error estimates of the subsymbolic regression model with the limits of accuracy for ground motion models given by Douglas and Smit (2001). In this process, also additional predictor variables could be evaluated, such as rupture dip or directivity effects. Furthermore, it might be interesting to use other functions than polynomials in the development of the subsymbolic model, such as splines, neural networks or regression trees. These are, like polynomials, very flexible, but have the advantage that they are better suited to model local dependencies. In conclusion, we believe that we have presented a viable new perspective on the development of ground motion models for seismic hazard analysis.

Data and Resources

Ground motion data used in this study were compiled for the NGA project. Data and accompanying information can be downloaded from http://peer.berkeley.edu/nga (last accessed September 2007). The coefficients of the regression model can be obtained by email from the first author.

Acknowledgments

We thank Matthias Ohrnberger for valuable discussions. We also thank Dave Boore and Ken Campbell for kindly providing us with the exact datasets they used for development of their NGA models. We thank the reviewers Peter Stafford and John Douglas for their detailed comments, which greatly helped to improve the article.

3

MODELING THE JOINT PROBABILITY OF EARTHQUAKE, SITE, AND GROUND-MOTION PARAMETERS USING BAYESIAN NETWORKS

Kuehn, N. M., C. Riggelsen, and F. Scherbaum
Bulletin of the Seismological Society of America, in press

Bayesian networks are a powerful and increasingly popular tool for reasoning under uncertainty, offering intuitive insight into (probabilistic) data generating processes. They have been successfully applied to many different fields, e.g. bioinformatics. In this paper, Bayesian networks are used to model the joint probability distribution of selected earthquake, site and ground-motion parameters. This provides a probabilistic representation of the (in)dependencies between these variables. In particular, contrary to classical regression, Bayesian networks do not distinguish between target and predictors, treating each variable as random variable. The capability of Bayesian networks to model the ground-motion domain in probabilistic seismic hazard analysis is shown for a generic situation. A Bayesian network is learned based on a subset of the NGA dataset, using 3342 records from 154 earthquakes. Since no prior assumptions about dependencies between particular parameters are made, the learned network displays the "most probable model given the data". The learned network shows that the ground-motion parameter (horizontal peak ground acceleration, PGA) is directly connected only to the moment magnitude, Joyner-Boore distance, fault mechanism, source-to-site azimuth, and depth to a shear wave horizon of 2.5 km/s (Z2.5). In particular, the effect of V_S30 is mediated by Z2.5. Comparisons of the PGA distributions based on the Bayesian networks with the NGA model of Boore and Atkinson (2008) show a reasonable agreement in ranges of good data coverage.

3.1 Introduction

There have been important developments in the evolution of probabilistic seismic hazard analysis (PSHA) since the pioneering works of Cornell and Esteva in the 1960's. For example, it is now considered bad practice if the ground-motion parameter is not treated as a random variable (Abrahamson, 2000). It has been recognized that it is important to incorporate its full distribution into the calculation of PSHA for a proper treatment of the uncertainties (Bommer and Abrahamson, 2006). In fact, ground-motion uncertainty is one of the key factors that controls the exceedance frequency for a given ground-motion value (e.g. Bommer and Abrahamson (2006)). Hence, it is very important to have a good understanding of the variables that can affect the ground-motion variability (Strasser at al., 2009).

In the derivation of a ground-motion model, the goal is to estimate the conditional distribution $\Pr(Y|\boldsymbol{x})$, where Y is a ground-motion parameter, e.g. peak ground acceleration (PGA) or spectral acceleration (PSA), while $\boldsymbol{x}$ is a vector containing some earthquake and site-related parameters, e.g. magnitude, distance or local shear wave velocity V_S30, that act as predictor variables. Traditionally, it is assumed that the ground-motion parameter Y is sampled from a log-normal distribution (or its logarithm $\log Y$ from a normal distribution) whose median μ (and possibly the standard deviation σ) is input dependent, i.e., a function of the predictor variables:

$$\log Y \sim \mathcal{N}(\mu = f(\boldsymbol{x}), \sigma). \tag{3.1}$$

This is equivalent to the following, more traditional notation:

$$\log Y = f(\boldsymbol{x}) + \epsilon, \tag{3.2}$$

where the residual ϵ is normally distributed with mean zero and standard deviation σ. Besides the assumption of log-normality, usually assumptions on the functional form of $f(\boldsymbol{x})$ and the kind and type of predictor variables $\boldsymbol{x}$ to include are made. These constraints are commonly based on exploratory data analysis, visual inspection of residual plots and expert knowledge concerning seismological processes. Once a functional form for $f(\boldsymbol{x})$ is chosen, its parameters are derived by regression, usually taking into account multiple recorded earthquakes (e.g Joyner and Boore, 1993, 1994; Abrahamson and Youngs, 1992). Regression based ground-motion models of the type of eq. (3.2) have been employed successfully in PSHA for many years. However, they are not without problems. For example, different ground-motion models can in fact employ quite different functional forms for $f(\boldsymbol{x})$ and different sets of predictor variables $\boldsymbol{x}$. One problem in this context is that the exact physical relationships between some of the predictor variables and the ground-motion parameters are not known. One way of balancing physical constraints and predictive power for $f(\boldsymbol{x})$ is described in Kuehn et al. (2009a).

Furthermore, for regression as such, the implicit assumption is that only the target is considered a random variable while the predictor variables are not. That is, the predictor variables are unassociated with any distributional variation, and are assumed to be error-free and completely observed. However, this is not the case for earthquake catalogs, where there is uncertainty about the true values of the earthquake or site-parameters. There are methods to take into account these uncertainties (e.g. Rhoades, 1997), but in most applications, the predictor variables are implicitly

assumed to be error-free.

In this paper, we take a different view on the derivation of ground-motion models for PSHA, a view that is adopted from the fields of machine learning and artificial intelligence: We aim at directly modeling the joint probability of the ground-motion parameter Y and the predictor variables $\boldsymbol{X}$, i.e. we take a multivariate stance, and aim at modeling the joint probability $\Pr(Y, \boldsymbol{X})$. In particular, this allows us to compute the conditional distribution $\Pr(Y|\boldsymbol{X})$, but also any other conditional or marginal probability. In general, modeling directly the joint probability distribution $\Pr(Y, \boldsymbol{X})$ offers a lot more flexibility than just modeling $\Pr(Y|\boldsymbol{X})$. Furthermore, modeling $\Pr(Y, \boldsymbol{X})$ means that there is no distinction between target and predictor variables - all variables are treated equally as random variables. It has to be stressed that there is a fundamental difference between regression and the approach taken here. In regression, one tries to estimate a conditional probability distribution $\Pr(Y|\boldsymbol{x})$ as compared to the joint probability distribution $\Pr(Y, \boldsymbol{X})$.

When we model the joint probability of earthquake, site, and ground-motion parameters, it is possible to incorporate all sorts of constraints or prior information/assumptions about the dependencies between the variables. However, in this work we refrain from making such prior assumptions but take a completely data-driven approach instead. This way, we aim at modeling only those relations that are actually required by the data.

We model the joint probability of earthquake, site, and ground-motion parameters by means of Bayesian networks (hereafter called BNs; see e.g. Pearl (1988)), which provide a compact, graphical interpretation of $\Pr(Y, \boldsymbol{X})$. This graphical representation of the joint probability allows for an intuitive insight into the probabilistic dependencies between the individual variables. Furthermore, the modular structure of BNs offers easy extensibility. In particular, BNs can be enhanced by utility and decision nodes, which together with the basis of BNs in probability theory enables proper reasoning under uncertainty. Thus, BNs provide a methodological framework for risk management systems, which has been recognized e.g. in Bayraktarli et al. (2006), Blaser et al. (2009) and Straub (2005). It should be noted that the name Bayesian network comes from the fact that Bayes rule is used to draw inferences from the network. It does not pertain to any statistical philosophy.

In Bayraktarli et al. (2006), the capability of BNs for earthquake risk management is shown. There, the ground motion domain is modeled very simply by an earthquake, distance and PSA node, based on the model of Boore et al. (1997) (for details, see Bayraktarli et al. (2006)). Kuehn et al. (2009b) used a BN constructed from the model of Campbell and Bozorgnia (2008) to investigate the sensitivity of PSHA results to certain predictor variables. These studies show the potential of BNs for PSHA. However, these BNs are based on theoretical assumptions and/or existing ground motion models. By contrast, in this work we directly learn BNs from data. That way, we can directly assess all probabilistic (in)dependencies in the available data.

The paper is organized as follows: First, a brief intuitive introduction to BNs is given in Section 3.2. In Section 3.3 we describe the dataset underlying this work. We perform some synthetic tests in Section 3.4 to demonstrate that the approach is valid. In Section 3.5 the results of the application to the real world dataset is shown, and we end with a discussion and conclusions.

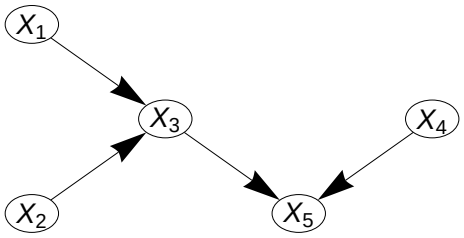

Figure 3.1: Example of a directed acyclic graph over a domain $\boldsymbol{X}$ with five variables.

3.2 Bayesian Networks

Here, we can give only a very short introduction into BNs. A more detailed overview can be found in Pearl and Russel (2000). There exist also several textbooks on the topic, e.g. Pearl (1988), Jensen and Nielsen (2001) or Koller and Friedman (2009).

A BN is a concise representation of a joint probability distribution $\Pr(\boldsymbol{X})$ over a set of variables $\boldsymbol{X} = \{X_1, \ldots, X_N\}$. The compactness of a BN is achieved by factoring the joint distribution into local conditional probability distributions by exploiting conditional independences between the variables. These conditional independence statements can be encoded graphically in a so-called *directed acyclic graph* (DAG), which comprises a set of nodes and directed edges (arrows, arcs) between the nodes. If there is an arrow pointing from node X_i to node X_j (i.e. an arc $X_i \rightarrow X_j$), we say that X_i is a parent of X_j, and X_j is a child of X_i.

Formally, a BN is a pair: The first part is a DAG $\mathcal{G}$ whose nodes correspond to the variables $X_1, \ldots, X_N$. $\mathcal{G}$ encodes conditional independence statements between the variables via the so-called Markov properties. For instance each node X_i is conditionally independent of its non-descendants given its parents. The second part of the BN is its set of parameters. These are the local conditional probability distributions for each node of $\mathcal{G}$, $\Pr(X_i|Pa(X_i))$, where $Pa(X_i)$ is the parent set of X_i in $\mathcal{G}$.

The joint probability distribution then factorizes as a product,

$$\Pr(\boldsymbol{X}) = \Pr(X_1, \ldots, X_N) = \prod_{i=1}^{N} \Pr(X_i|Pa(X_i)). \tag{3.3}$$

Thus, it is possible to recover the joint probability distribution $\Pr(\boldsymbol{X})$ from the DAG structure and the parameters - it does not need to be specified explicitly.

As an example, a simple BN is shown in Figure 3.1 for a domain with five variables. Formally, this DAG entails many conditional independence statements. For example, one can read off that X_5 is independent of X_1 given X_3. Intuitively, by looking at the DAG as a whole, one can read off the qualitative influences various variables have on each other. The previous independence statement intuitively can be understood as "X_1 influences X_5 indirectly via X_3". According to eq.

(3.3), the joint probability distribution for the example DAG of Figure 3.1 can be written as

$$\Pr(\boldsymbol{X}) = \Pr(X_1)\Pr(X_2)\Pr(X_3|X_1,X_2)\Pr(X_4)\Pr(X_5|X_3,X_4) \tag{3.4}$$

The example DAG of Figure 3.1 also includes the substructure $X_1 \rightarrow X_3 \leftarrow X_2$. This is an example of a so-called v-connection. In a v-connection, the parents of a variable become dependent once the value of the child is known. This means, if we know the state of X_3 and receive information about X_1, this will change our belief about the distribution of X_2.

From the joint distribution $\Pr(\boldsymbol{X})$ and its factorization property, it is possible to compute any conditional or marginal probability distribution of interest, using normal probability calculus:

- Sum-rule: $\Pr(X_i) = \sum_{\boldsymbol{x} \setminus X_i} \Pr(\boldsymbol{x})$
- Chain-rule: $\Pr(\boldsymbol{X}) = \Pr(X_i|\boldsymbol{X} \setminus X_i)\Pr(\boldsymbol{X} \setminus X_i)$

Fast inference algorithms exist that facilitate reasoning in BNs without having to apply the above rules systematically, but rather exploit the Markov-properties of the DAG to speed up inference. For details, see e.g. Jensen and Nielsen (2001).

The direction of the arcs is not necessarily tied to a causal interpretation. In particular, $X_1 \rightarrow X_2$ can, but does not need to mean "X_1 causes X_2". For purely probabilistic queries, it does not matter whether a BN is causal or not - the BN has to be a representation of the joint probability distribution to answer such queries. It is at best very difficult to learn causal relationships from data alone. In this work, we are only interested in learning the probabilistic dependencies between different variables from data, and therefore refrain from interpreting the learned network causally.

There are several characteristics that make BNs appealing to use in PSHA. For example, it is possible to reason with partial observability. This means, even if we do not completely know the states of all relevant variables but only of a subset, we can still compute conditional probabilities given this subset. The additional uncertainty of not knowing the states of the other variables is incorporated in the model framework. Applied to ground-motions, we could e.g. calculate the conditional probability of PGA given magnitude, even if we do not know the distance, which would lead to a rather wide distribution. Once we obtain information about the distance, the probability can be easily updated, resulting in a reduced spread of the ground-motion distribution.

On the other hand, it is possible to reason in any direction. One cannot only compute the probability of PGA given magnitude, but also the probability of magnitude given PGA, which might offer new perspectives in disaggregation of a PSHA.

BNs can also be used to perform sensitivity studies in PSHA. A BN directly provides the conditional probability distribution of a ground-motion parameter given some predictor variables (either fully or partial observed), which is directly related to the hazard curve. Together with the fast inference algorithms it is hence possible to immediately see the influence of different parameters on the hazard curve.

Given sufficient data, it is possible to learn a BN, both the structure as well as the parameters. Learning BNs is a large research topic and we refer to Koller and Friedman (2009), chapters 16-19 and references therein, for more on this. It is important to note that both structure and parameters

of a BN can be learned using either maximum likelihood estimation or Bayesian inference. Below, we give some details about the learning technique applied in this work.

3.2.1 Learning

In short, learning a BN is an instance of so-called "model selection", where we learn the most probable BN that could have been responsible for generating the data at hand. We note that a principled model selection technique takes care of overfitting. Using a scoring metric it is possible to "find" the best BN consisting of N nodes using a suitable BN traversal strategy: various potential BNs are scored, and the best one is selected as the generating joint distribution. Several scoring metrics have been proposed in the literature. Here we apply a scoring metric (the MAP BN metric) introduced in Riggelsen (2008) which has been shown, both theoretically and empirically, to outperform other scoring metrics. The scoring metric assigns a score to a BN depending on how well the BN in question is able to predict the data cases/records. Hence, a BN that is able to explain the data the best way possible is assigned a relatively high score. For the traversal strategy we use the method of inclusion-driven learning (Castelo and Kocka, 2003) which provably, in conjunction with the MAP BN metric, will come across the best BN.

In a nutshell, the "best" BN is selected in the following way: We start with an empty network, and perform subsequently one of the following basic operations:

- arc addition: Change $X_i \nsim X_j$ (no arc) to $X_i \rightarrow X_j$
- arc removal: Change $X_i \rightarrow X_j$ to $X_i \nsim X_j$
- arc reversal: Change $X_i \rightarrow X_j$ to $X_i \leftarrow X_j$

The resulting networks are scored with the MAP BN metric, and the highest scoring network is selected as the best one. The score of an arc depends on the sufficient statistics $n(X_i, Pa(X_i))$, i.e. the joint number of instances in different states:

$$n(x_i, \boldsymbol{x}_{Pa,i}) = \#(X_i = x_i \wedge \#Pa(X_i) = \boldsymbol{x}_{Pa,i}). \tag{3.5}$$

The sufficient statistics, and thus the score of an arc, depend on the amount of data available as well as the cardinality of the state space (i.e. the discretization), which has implications for learning, as we will see below.

In Riggelsen (2006) a way to handle partly missing or incomplete data (under the so-called *ignorability assumption*) when learning BNs is described using the so-called Markov blanket predictor approach. Combining this technique, the MAP BN scoring metric and the inclusion-driven search methodology, we have all the ingredients required for doing model selection in terms of BNs.

The approach we apply for learning will yield the best network that is able to predict the data meaning that it will not pick up noise and (spurious) details. A data set is detailed and potentially noisy when it consists of many variables and/or a large cardinality state space ("many bins") in comparison to the number of records/cases/instances available. When having such a data set, it is difficult, based on data alone, to generalize the data, as it requires a certain minimal number of

observations in order to be able to establish connection (e.g., metaphorically speaking, how many co-occurrences are required for values of 2 variables in order to claim that they are related?). This "minimal number" is not explicitly set in the learning approach we are taking, rather, it is determined automatically as the degree to which we are able to predict the data. The better we are able to predict the data, the better the network. This approach automatically means that noise should not be learned as this does not contribute to a good prediction. It also means that the network will only pick up the detail if it pays off in terms of predictiveness, that is, if it is "worth modeling" compared to the pay-off in how well it predicts the data.

3.3 Dataset

The ground-motion dataset we use for learning BNs is the one compiled for the Next Generation of Attenuation (NGA) relationships project (Power et al., 2008; Chiou et al., 2008). This dataset comprises 3551 strong-motion recordings from 173 earthquakes, mainly from California, but also from other regions representing shallow active tectonics, e.g. Taiwan or Alaska. The NGA dataset not only contains information about magnitude and source-to-site distance, but also includes numerous meta-data for each recording such as finite-rupture models, directivity parameters or local site effects such as V_S30 or sediment depth. This makes the dataset an ideal basis to investigate the (probabilistic) relationships between various earthquake-source, path, site and ground-motion parameters. In the following, we describe the criteria by which we select the records from the NGA dataset that are used for subsequent analysis.

The model we develop is supposed to represent free-field ground-motions from active tectonic regions. Hence, we include only those records in the analysis that are representative of free-field conditions. The classification of a free-field recording is based on the Geomatrix classification C1 (see e.g. Abrahamson and Silva, 1997, for details) for the corresponding station. As non free-field sites we consider those stations which have a GMX C1 code C, D, E, F and G.

Additionally we exclude some records from the Chi-Chi-sequence in Taiwan, following Abrahamson and Silva (2008). Several stations in Taiwan house more than one strong motion instrument, and only those records from the newer instrument are included, as recommended in Lee et al. (2001). Furthermore, a couple of stations are classified as poor quality in Lee et al. (2001) and are therefore not included in the further analysis.

We include only records that have a rupture distance smaller than 200 km. Records with a larger source-to-site distance are of low engineering significance. Furthermore, also a possible bias due to different attenuation properties of the earthquake source regions is reduced. Since some records do not have an estimate of the rupture distance, we calculate corresponding values from the epicentral distance using the approach of Scherbaum et al. (2004b) and exclude those with rupture distance values greater than 200 km. However, the converted values are only used for data selection, not for the learning process.

The ground-motion parameter we investigate is the horizontal peak ground acceleration (PGA). The geometric mean is used to combine the two horizontal components. No restrictions for the dataset come from this parameter. In particular, records with information missing on PGA are retained in the analysis since they can provide important information about the relationships of the other variables.

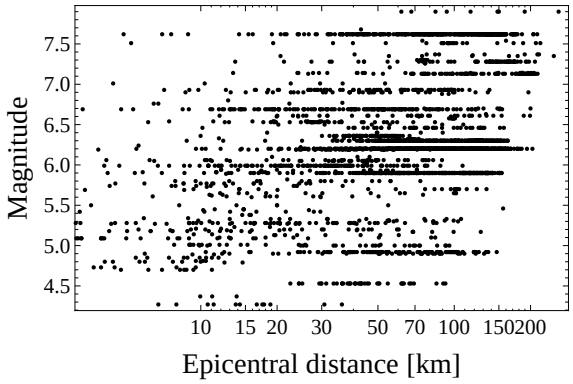

Figure 3.2: Magnitude vs. epicentral distance distribution of the dataset used in this study (3342 data points). Records with an epicentral distance larger than 200 km are included since their rupture distance, calculated using the method of Scherbaum *te al.* (2006) is smaller than 200 km.

In total, 3342 records from 154 earthquakes and 1314 stations are selected. The magnitude-distance distribution of this dataset is shown in Figure 3.2. In the electronic supplement, we provide a detailed list which records of the NGA flatfile are used for the present work.

3.4 Synthetic Tests

Despite it being the best strong-motion dataset currently at hand, from a statistical point of view the NGA dataset is not without challenges. One difficulty that arises is that multiple recordings from the same earthquake are not independent. Furthermore, even though it is possible to deal with missing data, this issue nevertheless complicates matters. On the applied side, BNs require that continuous data is discretized. However, BNs are a sub-class of a general framework referred to as graphical models, where the treatment of continuous and mixed variables is possible. We are currently exploring this, but for the present paper we restrict ourselves to BNs, considering it as a first step towards a full multivariate approach. To investigate the influences of these issues on the learning process of BNs, we also perform inference on synthetic datasets, where we have control over the number of records and the amount of missing data. Moreover, this allows us to look into the significance of different discretization schemes on the learned networks.

The synthetic datasets are generated from the empirical NGA model of Boore and Atkinson (2008) in the following way: First, synthetic values for the predictor variables magnitude, Joyner-Boore distance, V_S30 and faulting style are created randomly. Then median and standard deviation of PGA for the given predictor variables are calculated from the model of Boore and Atkinson (2008), and a PGA value is sampled from the corresponding log-normal distribution. Then the continuous variables are discretized, and missing data are created randomly. This is repeated for different numbers of records, different discretization schemes and different fractions of missing data.

To sample the predictor variables for the synthetic datasets, we need to specify their respec-

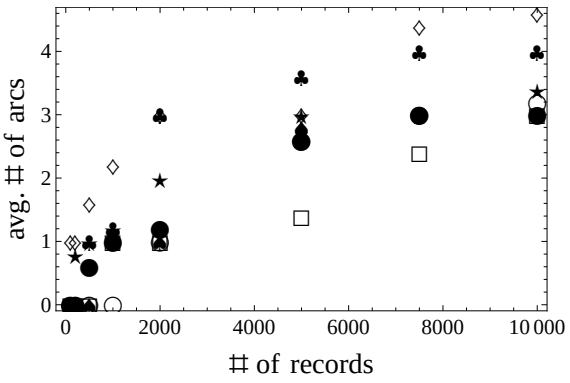

Figure 3.3: Average number of learned arcs for networks learned on synthetic datasets with different discretization schemes and dataset sizes. The cardinality of the state space for the different networks is given in by the corresponding symbols in Table 3.1.

tive distributions. For the magnitudes, we use a doubly truncated Gutenberg-Richter distribution between $M_W = 5$ and $M_W = 8$, with a b-value of one. The distances are sampled from a distribution that resembles the distribution of Joyner-Boore distances between 0 and 200 km in the NGA dataset. V_S30 values are drawn from a uniform distribution between 200 and 1500 m/s, and the probabilities for the focal mechanism to be normal, strike slip or reverse are 1/3, 1/3 and 1/3, respectively. The predictor variables are assumed to be independent.

The model of Boore and Atkinson (2008) is of the form

$$\ln(PGA) \sim \mathcal{N}(\mu = f(M_W, R_{JB}, V_S30, MECH), \sigma), \tag{3.6}$$

where M_W is the moment magnitude, R_{JB} is Joyner-Boore distance, V_S30 is the average shear wave velocity in the upper 30m and $MECH$ is the focal mechanism. Eq. (3.6) is an expression for $\Pr(PGA|M_W, R_{JB}, V_S30, MECH)$. Hence, we can define a network based on the theory of our simulations. In this network, PGA should be connected to all predictor variables, while these are marginally independent of each other and thus unconnected. Thus, it has four arcs. We will call this theory-based network the 'gold standard', since it is the theoretically best network possible.

As we have described above, there are four predictor variables in the model of Boore and Atkinson (2008). However, the effect of these predictor variables on the distribution of PGA is rather different. While changes in magnitude or distance severely impact the median of PGA, the effects of V_S30 and the faulting style are comparatively small. Therefore, we call the first associations strong dependencies and the latter ones weak dependencies. This has important consequences for learning, as is shown subsequently.

In Figure 3.3 we show the average number of arcs for networks learned on synthetic datasets with different numbers of records. The corresponding discretization schemes are summarized in Table 3.1. A number of important points can be seen in Figure 3.3. As one can see, for a small number of records, the learned networks have less than four arcs, i.e. it is not possible to

Table 3.1: Cardinality of state space for the networks in Figure 3.3

Symbol used in Figure 3.3	No. of PGA bins	No. of magnitude bins	No. of distance bins	No. of V_S30 bins	No. of mechanism bins
⋆	7	6	10	3	3
♣	4	6	10	3	3
♠	10	6	10	3	3
□	7	10	10	3	3
◯	7	6	16	3	3
•	7	6	10	3	3
♢	4	6	7	2	3

learn all dependencies. In these cases, only (if at all) the strong dependencies, i.e. $M_W \leftrightarrow PGA$ and $R_{JB} \leftrightarrow PGA$ can be learned. However, if the discretization is too fine, i.e. there are too many bins per variable, it is not even possible to learn these arcs. On the other hand, for large numbers of records, it is possible to reproduce the four arcs of the gold standard network. However, care must be taken: If the discretization is too coarse, more than four arcs are learned when the amount of data is sufficient. This is due to the way how the different networks (with and without additional arcs) are scored (cf. section 3.2.1). The score of an additional arc depends on the joint number of instances in different states, i.e. the sufficient statistics (cf. eq. (3.5)). That means, if the discretization is very fine, there are not many joint instances, so additional arcs, even for strong dependencies, are not scored high enough to be added to the model. On the other hand, if the discretization is coarse, there are a lot of joint instances, which means that even arcs for independent variables are added.

In general, we can draw the following conclusions regarding learning the structure of BNs on the synthetic datasets:

- **Small** amount of data: only strong dependencies can be learned; coarse discretization required
- **Large** amount of data: weak dependencies can be learned; finer discretization possible
- Missing data does not have a very large influence on the learned networks, especially for the strong dependencies.

We note that the first two points can be associated with the notion of generalization.

In the following, we compare one learned network with the generating model of Boore and Atkinson (2008). The dataset on which the network is learned comprises 10,000 records, the variables are discretized as follows:

- PGA: six equal-sized bins between $\min(\ln(PGA))$ and $\max(\ln(PGA))$
- Magnitude: six bins of width 0.5 between 5 and 8

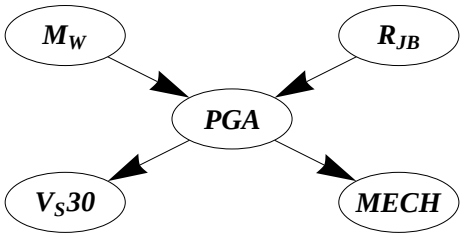

Figure 3.4: Structure of a BN learned on a synthetic dataset, generated from the model of Boore and Atkinson (2008), with 10,000 records.

- Distance: 10 bins of width 20 km between 0 and 200 km
- V_S30: binned into "SOFT SOIL" ($V_S30 < 360m/s$), "STIFF SOIL" ($360m/s \leq V_S30 < 760m/s$), "ROCK" ($760m/s \leq V_S30$)
- Mechanism: "normal", "strike slip", "reverse"

The structure of the network is shown in Figure 3.4.

As one can see, the learned network corresponds to the gold standard in that all "predictor" variables are connected to PGA, while there are no direct arcs between them, meaning that they are marginally independent. Figure 3.4 also shows that magnitude and distance are parents of PGA, while V_S30 and the focal mechanism are its children. This result is tied to some inner semantics of BNs, as M_W and R_{JB} become dependent once the value of PGA is known - $M_W \rightarrow PGA \leftarrow R_{JB}$ is a so-called v-connection. This means that, given that we know the value of PGA, learning the value of M_W will change our belief in the distribution of R_{JB} (or vice versa). In principle, this should also apply for V_S30 or the mechanism - however, since their effect on the distribution of PGA is comparatively weak, the v-connection is not learned in that case. Learning a v-connection also requires a sufficient amount of data. For a small amount of data, it is not even possible to learn the v-connection $M_W \rightarrow PGA \leftarrow R_{JB}$.

We also stress again that the network depicted in Figure 3.4 is learned from data alone, with no prior assumptions. This makes it difficult to associate a causal interpretation to the direction of the learned arcs.

Now, we show comparisons of the predictions using the network of Figure 3.4 and the model of Boore and Atkinson (2008) in Figure 3.5. Therefore, we have fixed the states of the"predictor" variables in the BN and computed the conditional distribution of PGA. In Figure 3.5, four comparisons are shown, for different distance and magnitude ranges. V_S30 is set to the ROCK state, the mechanism is normal. For the generating model distributions (black lines in Figure 3.5), we set the values for magnitude and distance to the mean of the corresponding BN range. For V_S30, we use a value of 1100 m/s. The faulting style is set to normal. As one can see, the agreement between the BN and the generating model of Boore and Atkinson (2008) is good in ranges with good data coverage.

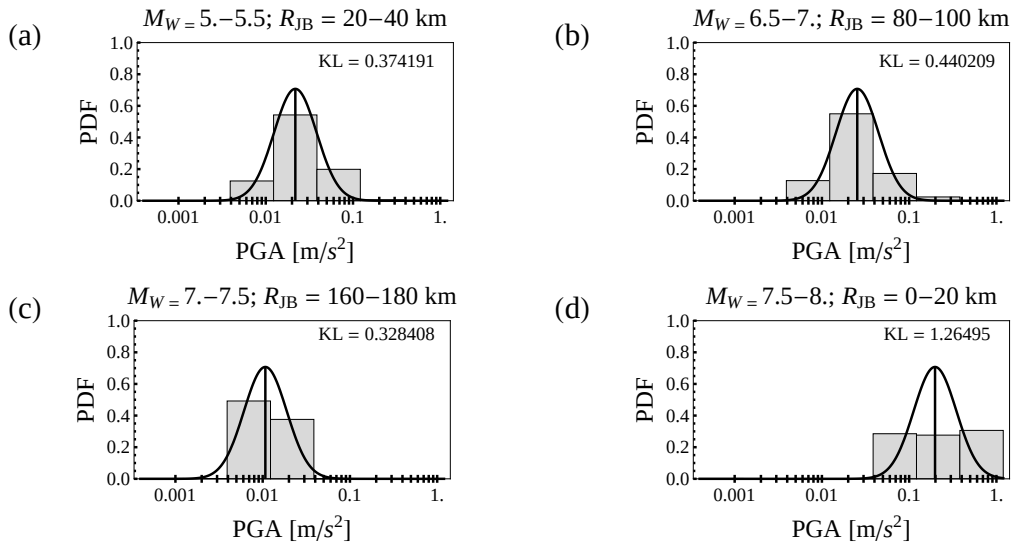

Figure 3.5: Comparison of conditional distributions of PGA, computed with a BN (gray rectangles) based on a synthetic dataset and the model of Boore and Atkinson (2008) (black line), for different magnitude and distance ranges. For the model of Boore and Atkinson (2008), magnitude and distance are taken to be the means of corresponding ranges. V_S30 is 1100 m/s, the focal mechanism is normal. The KL-divergences between the model of Boore and Atkinson (2008) and the Bayesian network for the displayed cases are given in the plots.

Since we cannot show comparisons for all possible combinations of magnitudes and distances, we use the relative information loss between the generating model, i.e. the NGA model of Boore and Atkinson (2008), and the BN to assess their difference, similar to what has been suggested for ground-motion model selection (Delavaud et al., 2009; Scherbaum et al., 2009). The relative information loss is expressed by the Kullback-Leibler (KL) divergence, which is shown in Figure 3.6 (a) for the whole magnitude/distance range.

The KL-divergence D_{KL} is a measure of distance between two probability distributions P and Q. D_{KL} is a positive quantity and is zero if both P and Q are identical (e.g. Scherbaum et al., 2009). In our case, we calculate the KL-divergence between the conditional probabilities of PGA given the other variables, calculated both with the BN as well as the model of Boore and Atkinson (2008), respectively. V_S30 and the focal mechanism are set to fixed values (1100 m/s and normal, respectively), while magnitude and distance vary over their whole range. For our case, the KL-divergence can be calculated by

$$D_{KL}(p(x)||q(x)) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx, \tag{3.7}$$

where $p(x) = \mathrm{Pr}_{BN}(PGA|M_W, R_{JB}, V_S30, MECH)$ is the conditional distribution of PGA given the predictor variables, calculated with the BN, and
$q(x) = \mathrm{Pr}_{BA}(PGA|M_W, R_{JB}, V_S30, MECH)$ is the conditional distribution of PGA given the

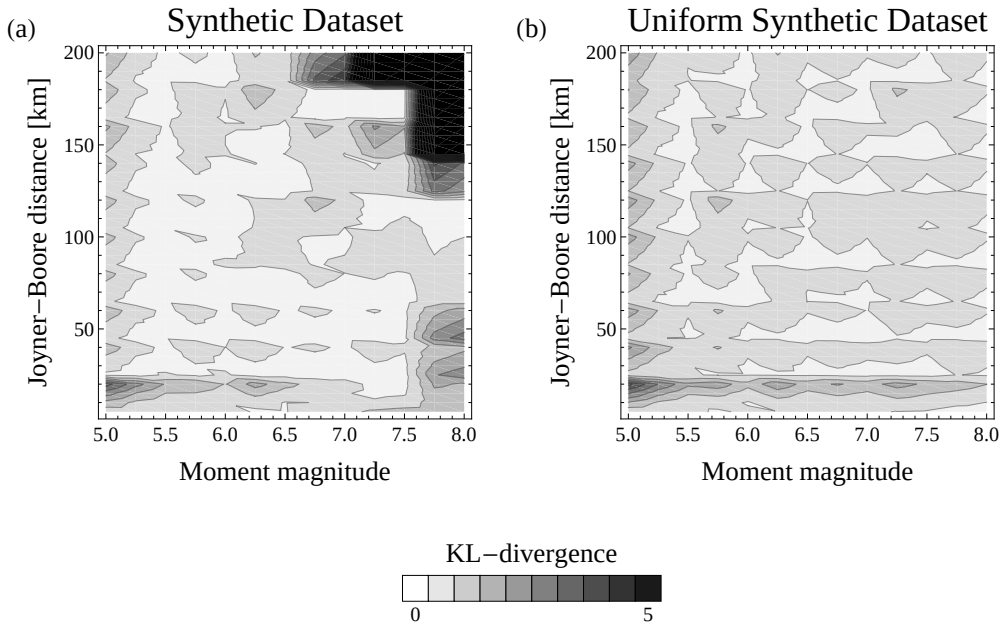

Figure 3.6: KL-divergences between the conditional distributions of PGA given M_W, R_{JB}, V_S30 and MECH for different values of M_W and R_{JB}, calculated with the model of Boore and Atkinson (2008) and a BN that was learned on different datasets: (a) a synthetic dataset, sampled from a doubly truncated GR-distribution and a distance distribution that resembles the NGA dataset; (b) a synthetic dataset, sampled from a uniform magnitude and distance distribution.

predictor variables, calculated with the model of Boore and Atkinson (2008).

In Figure 3.6 (a) the KL-divergences between the generating model of Boore and Atkinson (2008) and the BN are shown over the the whole magnitude/distance range of the synthetic datasets. As one can see, the values of D_{KL} are small for small magnitudes and distances, which corresponds to a good agreement between the two distributions. On the other hand, for large magnitudes and distances the KL-divergences increase, thus indicating a mismatch. This is due to the distributions from which magnitude and Joyner-Boore distance are sampled - in both cases, the probabilities for large values are small, so the amount of data in these ranges is small. Hence, it is possible to recreate the generating model as a BN (structure and parameters), if the amount of data is sufficient. Otherwise, prior information/constraints are needed.

3.5 A Bayesian Network for the NGA Dataset

In this section, we apply the BN formalism to the NGA dataset. This dataset is in terms of metadata that is available for each earthquake, record and site the most comprehensible at the moment. Therefore, the NGA dataset is well suited to investigate the probabilistic dependencies between the individual ground-motion, earthquake, and site parameters. On the other hand, the NGA dataset

still presents some problems, as we have pointed out before. In the previous section we have seen that it is possible to deal with missing data, and that meaningful results can be obtained if the discretization strategy is chosen carefully. However, one problem that arises with strong motion data is that the data in general cannot be considered independent and identically distributed (iid), since we can have multiple records from one earthquake and one station recording several earthquakes. This poses a challenge to any statistical inference method, and learning BNs becomes more complicated.

The variables we consider in the learning process are shown in Table 3.2. We have included variables that account for different aspects of the ground-motion domain: size, orientation, site effects. Table 3.2 also shows the discretization scheme that is used for each variable. The ground-motion parameter PGA is log-transformed before learning. However, even if the ground-motion variable is thus strictly speaking Log[PGA], we refer to it as PGA in the following.

We use two different approaches to find a discretization scheme for each variable. The first is expert knowledge. For example, it makes sense to discretize the fault mechanism into normal, strike slip and reverse bins, or V_S30 into bins of soft soil, stiff soil and rock, as has been done in many published ground-motion models (e.g. Ambraseys et al., 2005; Bommer et al., 2007; Danciu and Tselentis, 2007; Zhao et al., 2006). The other approach is based on the entropy method of Fayyad and Irani (1993), which tries to maximize the information content of the discretized variable with respect to another variable. Since the variable we are most interested in is PGA, we first discretize PGA, and then use the method of Fayyad and Irani (1993) to calculate the boundaries of the other variables. As the method of Fayyad and Irani (1993) can have a problem with very skewed or noncontinuous distributions, we combine the two approaches. Table 3.2 also shows the discretization scheme that we have settled on for each variable. We use seven bins for PGA since we have seen in Section 3.4 that this gives a good tradeoff between the amount of data and the learning of dependencies. The bins for PGA are chosen so that each bin has the same width.

As we have pointed out above, learning BN becomes more complicated when the data is non-iid. Therefore, we use two different approaches to deal with this problem, which both exploit the underlying topology in the ground-motion domain and are later combined using prior (expert) knowledge. This topology consists of the three distinct entities of the domain – earthquake, record, and site – and is illustrated in Figure 3.7, which is based on ideas presented in Heckerman et al. (2007). Each variable, such as magnitude, distance, PGA or V_S30, belongs to one entity. The interpretation of this topology is simple: An earthquake, represented by its associated variables, is recorded at a site, also represented by its associated variables. The variables that represent the record entity are for example distance, azimuth or PGA and are unique for each earthquake/station pair. We can connect each entity with a table that contains the values of its associated variables. In our case, the length of the tables is 154 for the earthquake entity, 1314 for the station entity and 3342 for the record entity.

For the first approach to structure learning, we divide the ground-motion domain into its three entities and learn a BN structure for the variables of each entity separately. The ground-motion parameter is included in all three entities and serves as a connection between them. Since there are more records than earthquakes or sites, we average the values of PGA for the same earthquake or site. In principle this is not correct, since multiple records that have different distances are

Table 3.2: Variables that are used for learning the Bayesian network on the NGA dataset

Parameter	Notation	Unit	No. bins	Bin boundaries inside the data-range
Log[PGA]	PGA	g	7	equal-sized bins between $min(\ln[PGA])$ and $max(\ln[PGA])$
Moment magnitude	M_W		6	5.5,6.,6.5,7.,7.5
Depth to the top of the rupture	Z_{TOR}		2	TRUE,FALSE
Fault mechanism	MECH		3	normal, strike slip, reverse
Fault dip	DIP	°	3	50,80
Source-to-site azimuth	AZ	°	4	-90,0,90
Epicentral distance	R_{EPI}	km	10	20,40,60,80,100,120,140,160,180
Hypocentral distance	R_{HYP}	km	10	20,40,60,80,100,120,140,160,180
Rupture distance	R_{RUP}	km	10	20,40,60,80,100,120,140,160,180
Joyner-Boore distance	R_{JB}	km	10	20,40,60,80,100,120,140,160,180
V_S30	V_S30	m/s	3	360,760
Depth to V_S=1.0 km/s	Z1.0	m	3	290,532
Depth to V_S=1.5 km/s	Z1.5	m	3	543.5,905
Depth to V_S=2.5 km/s	Z2.5	m	3	1615,2454

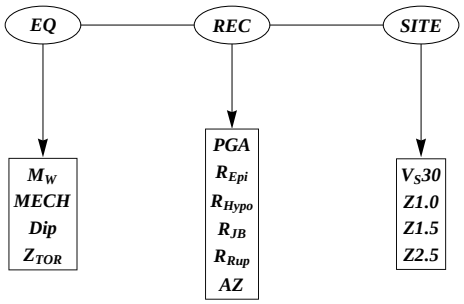

Figure 3.7: Topology of the ground-motion domain with three different subdomains/entities: earthquake (EQ) and site (SITE) entity, connected via the record (REC) entity. Each entity has its own associated variables.

averaged. However, the averaged values are only used to learn the substructure for each entity, i.e. for learning direct probabilistic dependencies between the variables, not for parameter estimation. On average, higher magnitudes produce higher ground-motions, so these dependencies will be detected, even though it is a rather crude, ad hoc approach for dealing with the problem of non-iid data. This approach to learning the structure makes one fundamental assumptions - that variables from different entities are only connected via PGA. We think that this is a reasonable assumption.

The second, alternative approach is based on Getoor et al. (2007) which we adapt for our needs and combine with the learning algorithm of Riggelsen (2008) and the method for dealing with missing data presented in Riggelsen (2006). In this approach, the joint probability of all variables is estimated at the same time. To account for the differences in the number of data for the different entities, the sufficient statistics, which are the basis for BN learning, are adjusted for arcs between variables of different entities. This is done by weighting each value by the number of times it is recorded. Hence, data entries only count as often as they appear in their respective entity table. However, the exact relation of the approach taken here to the semantics of BNs is unclear and requires further research. Therefore, we use its result, as the result of the first approach, as a guidance to constructing the structure of the BN for the NGA dataset.

Both approaches described above give insight into the probabilistic dependencies between the earthquake, site, and ground-motion parameters. Based on these approaches, we have selected the network that is depicted in Figure 3.8 as the one that best describes the data. We have added one arc as expert knowledge to the network that was not learned by the structure learning algorithms: $PGA \rightarrow mechanism$, since the focal mechanism is known to have an influence on the distribution of PGA (Bommer et al. (2008)). However, the effect of different focal mechanisms on PGA is small compared to the ones of magnitude and distance (cf. section 3.4). Therefore, the direction of the added arc is $PGA \rightarrow mechanism$ and not $mechanism \rightarrow PGA$ to avoid a v-connection between magnitude, distance and the focal mechanism (cf. section 3.4).

As one can see, in the final network PGA is only connected to magnitude, Joyner-Boore distance, the focal mechanism, the depth to the shear-wave horizon of 2.5 km/s and the azimuth. Hence, once we know the values of these variables, PGA is shielded from the influence from

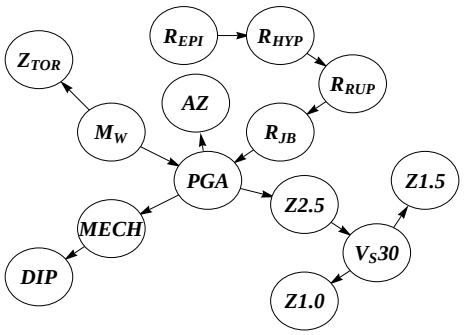

Figure 3.8: Structure of the Bayesian network for the NGA dataset.

other parameters, such as the fault dip (the variables connected to PGA form the so-called Markov blanket of PGA).

One important finding is that there is no direct arc between V_S30 and PGA. Instead, the influence of V_S30 on PGA is mediated via the depth to the 2.5 km/s shear wave horizon (Z2.5). This does not mean that there is no correlation between PGA and V_S30, but that once we know the value of Z2.5, V_S30 does not provide any *further* information for the prediction of the PGA distribution. This gives rise to the question whether V_S30 is the best site effects predictor or whether other variables should be explored (e.g. Castellaro et al., 2008).

It is also interesting that the final network depicted in Figure 3.8 contains an arc from PGA to the azimuth, which is a proxy for directivity effects. This arc is learned by the structure learning algorithms we applied. Hence, our findings indicate that there is a probabilistic dependency between directivity and ground-motions. This has also been found in other studies, e.g. Spudich and Chiou (2009). It should be emphasized that the arcs PGA → Azimuth and PGA → Z2.5 are the product of a solid statistical approach, indicating a real dependency which is inherent in the data.

A comparison between the predictions of the NGA BN and the NGA model of Boore and Atkinson (2008) is displayed in Figure 3.9. Here, we show the conditional distribution of PGA, computed with both models. The magnitudes and distances are the same as in Figure 3.5. The comparisons are made for a V_S30 value of 1100 m/s and a normal focal mechanism. Since the model of Boore and Atkinson (2008) does not contain the azimuth as a predictor variable, this node is left undefined in the BN. KL-divergences between the two models are shown in the plots. As one can see, the differences are larger than for the synthetic dataset. In particular, in Figure 3.9 (a) the PGA distribution is a bimodal one, which unrealistic from a physical perspective. This is due to the fact that the parameters of the BN are in good part controlled by the relative frequencies in the dataset. If the dataset is uneven in certain ranges, this reflects in the BN. This is what happens for the range $5. < M_W < 5.5$ and 20 km $\leq R_{JB} < 40$ km. It would be possible to use a strong prior or augment the data to smoothen the results (in a way, the functional form of a regression model can be thought of as a "prior" that smoothes the ground motion distribution) – however, that does not get along with our goal of minimum assumptions (see also section 3.6).

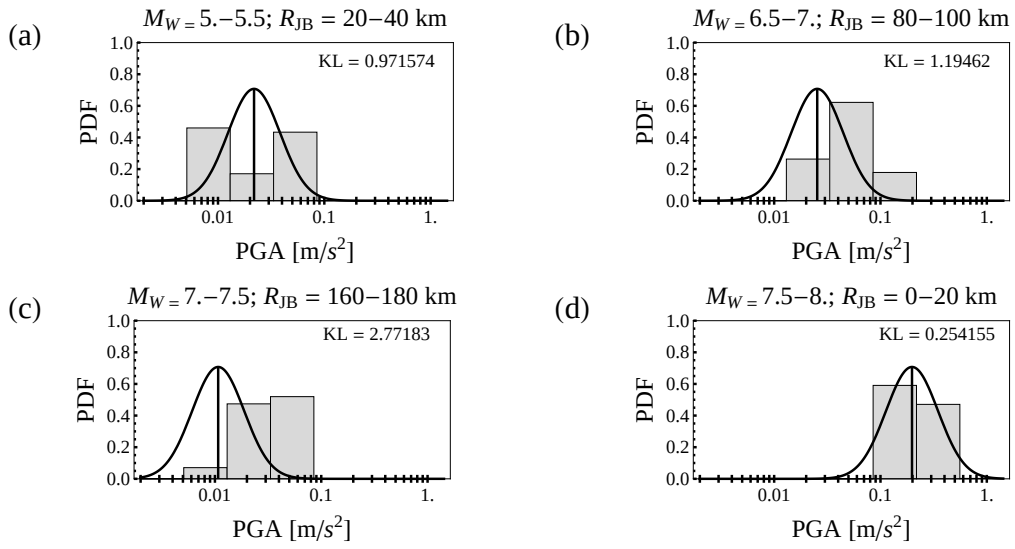

Figure 3.9: Comparison of the model of Boore and Atkinson (2008) with the Bayesian network: (a) – (d): Conditional distributions of PGA, computed with a BN learned on the NGA dataset (gray rectangles) and the model of Boore and Atkinson (2008) (black line), for different magnitudes and distances. For the model of Boore and Atkinson (2008), magnitude and distance are taken to be the means of corresponding ranges.. V_S30 is 1100 m/s, the focal mechanism is normal. The KL-divergences between the model of Boore and Atkinson (2008) and the Bayesian network for the displayed cases are given in the plots.

Figure 3.9 shows comparisons of the conditional distributions of PGA for individual magnitude/distance scenarios. By contrast, in Figure 3.10 (a) we compare the median predictions of the Bayesian network over the distance range 0-200 km with the median predictions of the model of Boore and Atkinson (2008) for two magnitudes, again for a V_S30 value of 1100 m/s and a normal focal mechanism. As one can see, for a magnitude of 6.25, where there is the bulk of the underlying dataset, there is reasonable agreement between the two models, though the BN predicts slightly higher PGA values at larger distances. However, for $M_W = 5.25$, there can be seen large differences in those parts where the Bayesian network is not fully supported by data. Here, the BN even predicts an increase of the median with distance. This is due to the fact that in some magnitude/distance ranges there are only very few data points, and thus the PGA distribution can be irregularly sampled in these cases. This leads to a distortion in the estimation of $\Pr(PGA|M_W, R_{JB})$, which is greatly affected by "outliers" in the case of a small dataset (the same effect can be seen in Figure 3.9 (a)).

Analogous to Figure 3.10 (a), (b) shows comparisons of median predictions over the magnitude range 5-8 for distances $R_{JB} = 50$ km and $R_{JB} = 90$ km. In some ranges, the two models are similar, in other ranges they are more dissimilar. An interesting feature in Figure 3.10 (b) concerns the largest magnitudes, i.e. the range between magnitudes 7.5 and 8. Here, we see a reduction of PGA with magnitude, i.e. oversaturation. This is a feature that also seen by the developers of the

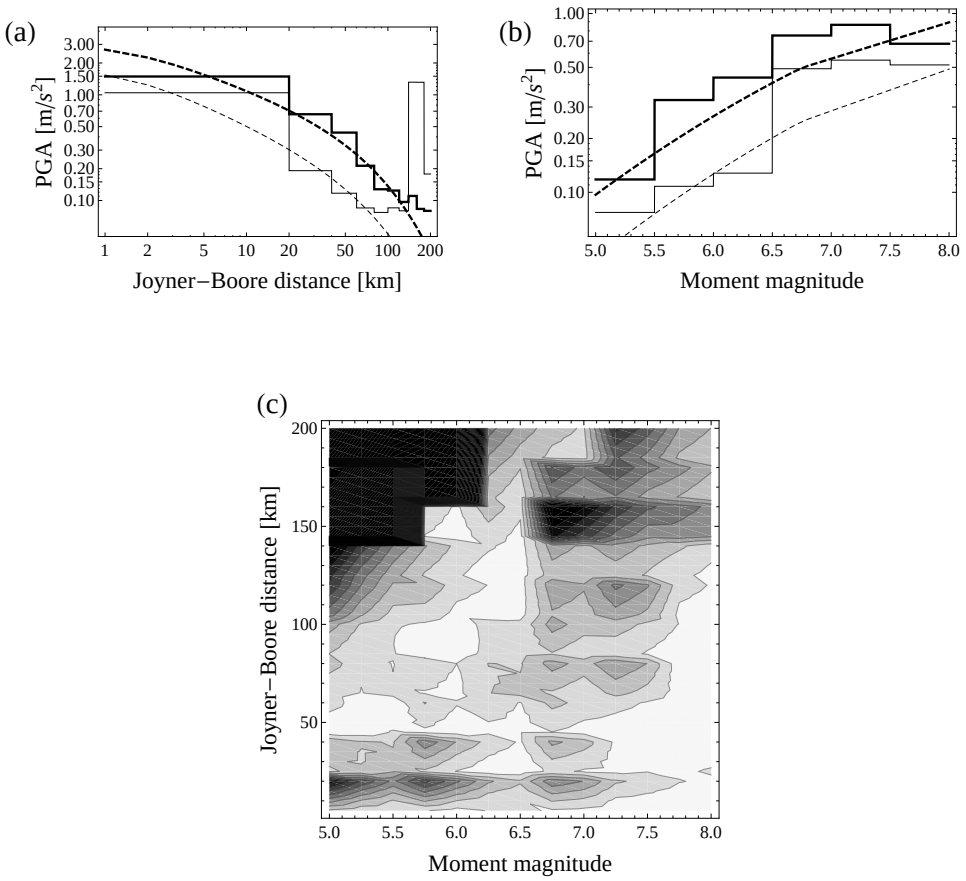

Figure 3.10: (a) Comparison of median PGA predictions of the model of Boore and Atkinson (2008) (dashed lines) with median PGA predictions of the Bayesian network (dashed line), for M_W = 5.25 (thin) and M_W = 6.25 (thick); (b); Comparison of median PGA predictions of the model of Boore and Atkinson (2008) (dashed lines) with median (black lines) PGA predictions of the Bayesian network, for R_{JB} = 50 km (thick) and R_{JB} = 90 km (thin); (c) KL-divergences between the conditional distributions of PGA, computed using the model of Boore and Atkinson (2008) and the Bayesian network, given M_W, R_{JB}, V_S30 and the focal mechanism, for different values of M_W and R_{JB}. The grayscale of the KL-divergence in (c) is the same as in Figure 3.6. V_S30 and the focal mechanism are 1100 m/s and normal, respectively, for (a), (b) and (c).

NGA models, but there the models were forced to be a monotone, non-decreasing function with magnitude (Boore and Atkinson, 2008; Campbell and Bozorgnia, 2008).

In Figure 3.10, but also in Figure 3.9, one can see that the BN predicts higher PGA values than

the model of Boore and Atkinson (2008) in some ranges. This is in part due to differences in the underlying datasets (3342 vs. 1604 records). We have compared th BN predictions with others using a continuous, non-parametric method based on the same 3342 records as thee BN, and they are similar in all ranges.

Both in Figure 3.10 (a) and (b) one can see the effect of discretization. The predictions of the BN are stepwise functions, with step sizes of 20 km and 0.5 units of magnitude, respectively. From a physical perspective, this is of course undesirable, since PGA, magnitude and distance are continuous variables. However, discretizing is required in order to learn the structure of BNs and make exact inferences without the assumption of a Gaussian distribution for the continuous variables.

Figure 3.10 (a) and (b) showed comparisons of median predictions. However, in PSHA the full ground motion distribution is important. These are compared for the two models over the full magnitude and distance range in Figure 3.10 (c), where we show the KL-divergences between the respective conditional distributions of PGA. As one can see, the two models are similar (i.e. have relatively low KL-divergences) in many ranges, and dissimilar in others. The average KL-divergence is larger than in the synthetic case (cf. Figure 3.6 (a)), which reflects that the datasets underlying the two models are different (1604 records in the case of Boore and Atkinson (2008) compared to 3342 records in the present work). For small earthquakes and large distances, the values of the KL-divergence are large since the amount of data in this range is insufficient (cf. Figure 3.2).

The BN learned for the NGA dataset can be found in the electronic supplement. It is stored as an xml-file and can used together with the freely available software GeNIe (`http://genie.sis.pitt.edu`).

3.6 Discussion

We have seen in section 3.4 that it is possible to model the ground-motion distribution using a BN. Given a sufficient amount of data, we obtain the expected number of arcs, and the learned conditional distribution of PGA is in good agreement with the underlying theoretical distribution (cf. Figures 3.3 and 3.6 (a)). However, Figure 3.6 (a) also shows that the PGA distributions are quite different for large magnitudes and distances. This is due to the distributions chosen for magnitudes and distances, which are sparse in these ranges. However, if we learn a network with a uniform magnitude and distance distribution, the KL-divergences are the same in all magnitude and distance ranges, as can be seen in Figure 3.6 (b).

Figures 3.6 and 3.10 also show the effects of the applied discretization. Since Joyner-Boore distance and magnitude are discretized into bins of 20 km and 0.5 magnitude units, respectively, $\Pr(PGA|M_W, R_{JB})$ does not change within one bin when calculated with the BN, whereas it does when calculated with the continuous model of Boore and Atkinson (2008), thus producing the small steps in Figures 3.6 and 3.10 (c). At first glance, the need to discretize variables for learning BNs might look like a great problem. However, it is often easy to find a reasonable discretization scheme for variables that are relevant in the ground-motion domain. For example, the fault mechanism is naturally binned into the three states “normal”, “strike slip” and “reverse”. Site effects have also traditionally been included as discrete variables in ground-motion models,

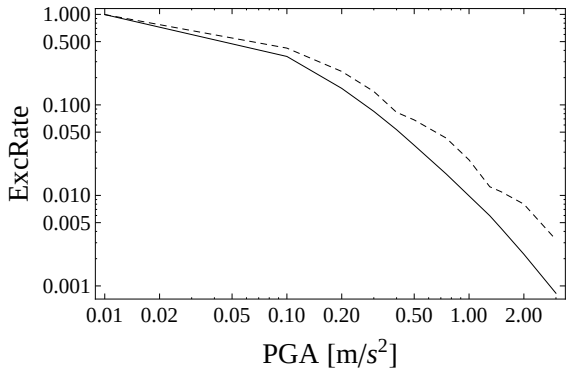

Figure 3.11: Hazard curves, calculated with the model of Boore and Atkinson (2008) (solid line) and the BN (dashed line), for a circular seismically active area of radius 140 km and a doubly truncated Gutenberg-Richter distribution (between M_W = 5 and M_W = 8. V_S30 is 1100 m/s, the fault mechanism is normal.

using various binning schemes such as Geomatrix classification or a discretization into "rock" and (stiff/soft) "soil". Only recent models incorporate V_S30 directly as a continuous variable (e.g. Abrahamson and Silva, 2008; Boore and Atkinson, 2008; Campbell and Bozorgnia, 2008; Chiou and Youngs, 2008). For the depth-to-the-top-of-the-rupture, a natural binning scheme is one where the variable has two states (TRUE/FALSE), like we used here. Nevertheless, there are variables where it is difficult to find a natural discretization, such as magnitude, distance or PGA. In these cases, one must choose the bin widths with care.

One direct consequence of the need to discretize continuous variables to be used in BNs is that it is not really possible to directly compare the results with a regression model. One can calculate a median from a discretized PGA distribution and plot them together with a regression model (cf. Figure 3.10 (a) and (b)), but this term and especially the standard deviation lose their meaning when also the predictor variables such as magnitude and distance are discretized. In Figures 3.6 and 3.10 (c) we have seen how it is still possible to compare two models using their KL-divergence. However, the primary use of a ground-motion model is to be employed in PSHA, and the effect of a different model representation on seismic hazard can sometimes be unforeseen (Musson, 2009). Therefore, two hazard curves are compared in Figure 3.11. Here, we have calculated hazard curves by sampling from the BN and the model of Boore and Atkinson (2008) for a doubly truncated Gutenberg-Richter distribution ($5 \leq M_W \leq 8$) with a b-value of one and a circular area around the site with radius 140 km. We set V_S30 to 1100 m/s and the fault mechanism to normal. As one can see, the hazard curve calculated from the BN is slightly increased compared to the one computed with the model of Boore and Atkinson (2008). This has two reasons: slightly higher median predictions by the BN, and the discretization that is applied. The discretization is also displayed by the little humps seen in Figure 3.11.

It is possible to directly use continuous data in more general graphical models than BNs, which can also offer intuitive insight into the domain. However, in these cases assumptions regarding the distributions of the individual variables need to be made, and the exact relations between them

need to be specified. In particular, structure learning becomes much more complicated. With BNs, we rely on fewer assumptions compared to these more general graphical models.

The discretization highlights an important aspect: Even though the NGA dataset is very extensive (the best strong motion dataset currently available), it is by no means exhaustive. Not all magnitude/distance ranges are equally well covered with data. The same is true for other (potential) predictor variables. This presents a problem for learning purely from data. However, it is important to see which parts of a model are actually supported by the data, and for which parts one needs to rely on assumptions. In this context, it is important to remember that our BN does not overfit the data, since this is taken care of during learning. Therefore, even if we rely on as little assumptions as possible and almost exclusively on data, the model still generalizes as far as the data allows.

In this work, our focus was learning on BNs purely from data, making as little assumptions as possible. As discussed before, learning BNs requires discretization of continuous variables such as magnitude, PGA, distance and others. There is a tradeoff between the width of the discretizations and the possibilities of learning. Finer discretizations (i.e. more bins) should lead to a better approximation of the continuous distributions. However, for smaller bin widths there will be many bins that contain no or only very few data points, which make them not useful, neither for parameter learning, structure learning nor for reasoning. One can think of ways to circumvent this problem. E.g. one could use simulations to increase the number of data, as is done in Blaser et al. (2009). Another possibility would be to place a strong prior on structure and parameters. However, both of these approaches require many and strong assumptions, which defies the purpose of this work.

Another problem in the context of learning a BN is the fact that multiple recordings from the same earthquake or recorded at the same site are not independent. Learning algorithms for BNs, both for structure and parameters, are generally designed for iid data. Hence, their application to the ground-motion domain is difficult. We described in section 3.5 how we dealt with this problem by partitioning the variables into the three entities earthquake, measurement and site (cf. Figure 3.7). However, this approach is kind of ad-hoc and lacks theoretical justification, even though we believe that it is a reasonable way to deal with the problem of non-iid data.

When breaking down the data into the three entities, every earthquake or site is used only once in the learning process. This means that we use 154 data points for earthquake related variables such as magnitude or mechanism and 1314 for site related variables, compared to 3342 data points for PGA or distance, which belong to the measurement entity. However, 154 data points might be sufficient only to discover strong dependencies, as we have seen in section 3.4. This is the reason why from the earthquake related variables, only the magnitude is connected to PGA by the structure learning algorithm. However, since it is known that fault mechanism has an (albeit small) effect on the distribution of PGA (Bommer et al., 2003), we have added this arc as expert knowledge. The direction of the arc was chosen so as not to create a v-connection. The other arcs are all learned. Thus, the connections in Figure 3.8 represent the statistical dependencies of several ground-motion related variables that are currently supported by the data.

An important feature of the learned network is that there is no learned direct arc connecting PGA and V_S30. However, both variables are not uncorrelated - the effect of V_S40 on PGA is mediated by the depth to a shear wave horizon of 2.5km/s. This might be an indication that V_S30 is not the

best predictor variable characterizing site effects. Hence, the use of other proxies for site effect characterization should be considered, as has been advocated lately (Castellaro et al., 2008).

One should keep in mind that the structure of the learned BN, as displayed in Figure 3.8, represents the current state of dependencies supported by the data. With more incoming data, the structure might (probably will) change. Also the parameters of the BN will change with increasing data, as more earthquakes will provide additional information on the interactions between earthquake, site and ground-motion parameters.

Another interesting feature of BNs is their extensibility. By adding nodes for e.g. the b-value or other ground-motion parameters like PSA, one could arrive at a full BN "hazard calculator" that takes into account all relevant variables and their corresponding uncertainties. For example, it is straightforward to calculate the hazard curve from the conditional distribution of the ground-motion parameter under consideration. A step further down the line would be to extend the BN to a decision support tool, as for example is done in tsunami early warning (Blaser et al., 2009), medical diagnosis (e.g. Nikovski, 2000) and many other fields. However, we acknowledge that this goal requires a lot of work in many fields. In this work, we have concentrated on learning a BN purely from data, both structure and parameters, which might be considered an early step towards these goals. This allowed us to assess which probabilistic (in)dependencies are actually supported by the available data. Future steps will include the combination of theoretical and empirical considerations, as we have seen that the BN is underrepresented by data in some ranges, which can lead to unphysical behavior if we rely only on the data.

3.7 Conclusions

We have presented a BN approach for the derivation of ground-motion models that directly estimates the joint probability distribution of several parameters related to the ground-motion domain in seismic hazard analysis. Directly modeling the joint-probability of earthquake, site, and ground-motion parameters gives insight into the data generating process hardly available otherwise. Since we use a Bayesian approach, the model we get is the maximum a posteriori model, i.e. the "most probable model given the data". Our results show that PGA is directly influenced by the magnitude, the Joyner-Boore distance, the source-to-site azimuth, the depth to a shear wave horizon of 2.5 km/s, and the fault mechanism. All other effects are mediated by one of these parameters. In particular, V_S30 affects the distribution of PGA only indirectly.

Data and Resources

Ground-motion data used in this study were compiled for the NGA project. Data and accompanying information can be downloaded from `http://peer.berkeley.edu/nga` (last accessed September 2007).

Electronic Supplement

A table with information on the records used in this study is available online at `http://www.geo.uni-potsdam.de/mitarbeiter/Kuehn/kuehn-esupp.html`. The Bayesian network can be downloaded from `http://www.geo.uni-potsdam.de/mitarbeiter/Kuehn/kuehn-esupp.html`.

Acknowledgements

Our implementation of Bayesian networks is based on the *SMILE* reasoning engine for graphical probabilistic models by the Decision Systems Laboratory, University of Pittsburgh (http://dsl.sis.pitt.edu). We would like to thank Yahya Bayraktarli for comments on an early draft of the manuscript. We also thank the editor Andrew J. Michael and two anonymous reviewers for helpful comments that clarified the manuscript.

4

A BAYESIAN GROUND-MOTION MODEL WITH CORRELATION OF GROUND MOTION INTENSITY PARAMETERS

Kuehn, N. M., C. Riggelsen, F. Scherbaum, and T. I. Allen
submitted to Bulletin of the Seismological Society of America

---

We present a Bayesian ground motion model that directly estimates both coefficients and the correlation between different ground motion intensity parameters. Therefore, we set up a graphical model which mimics our assumptions about the data generating process, i.e. which includes a source, path and station term. For each term, coefficients to predict the median of the intensity parameter distribution can be estimated, together with the associated covariance structure (i.e. between-event and within-event variability plus correlation coefficients). The graphical structure provides an easy, qualitative and intuitive insight into the model. The coefficients of the model are estimated in a Bayesian framework using Markov Chain Monte Carlo simulation. Thus, prior information can be included into the model in a principled way, and an estimate of the epistemic uncertainty of the parameters is provided. It also allows to easily update the model once new data becomes available. The parameters of the model are estimated on a global dataset using peak ground acceleration, peak ground velocity and the response spectrum at three periods as the target variables. There is correlation between all target variables, to a varying degree.

4.1 Introduction

Ground motion models (GMMs), also often called ground motion prediction equations (GMPEs), play a crucial role in probabilistic seismic hazard analysis (PSHA). Uncertainty in the estimation of the ground motion parameter of interest, e.g. peak ground acceleration (PGA) or spectral ac-

celerations, is one of the key factors that controls the exceedance frequency for a given ground motion value (e.g. Bommer and Abrahamson, 2006). There exist a wide variety of ground-motion models for different seismic provinces (shallow active tectonics, subduction zones and intraplate regions) and different regions in the world (e.g. California, Japan, Europe). There also exist many different functional forms that are employed to model the dependence of ground motions on predictor variables such as magnitude, distance or site effects. For a review of published GMMs, see Douglas (2003, 2006, 2008).

In technical terms, a GMM quantifies the conditional probability of a ground motion parameter Y given some earthquake and site related parameters $\boldsymbol{X}$, $\Pr(Y|\boldsymbol{X})$. In this context, it is usually assumed that the ground motion parameter Y is log-normally distributed, which leads to the following model:

$$\log Y = f(\boldsymbol{X}) + \Delta, \tag{4.1}$$

where Δ describes the total variability of the ground motion, which is usually decomposed into between-event variability, ΔB and within-event variability ΔW, which are independent of each other. Both ΔB and ΔW are normally distributed with mean zero and standard deviations τ and ϕ, respectively. Here, we follow the notation proposed by Al Atik et al. (2010) for the description of the variability of GMMs. We can rewrite eq. (4.1) to emphasize the probabilistic nature of ground motion as

$$\log Y \sim \mathcal{N}(\mu = f(\boldsymbol{X}), \sigma = \sqrt{\tau^2 + \phi^2}), \tag{4.2}$$

which reads as "$\log Y$ is distributed according to a normal distribution with mean $\mu = f(\boldsymbol{X})$ and standard deviation $\sigma = \sqrt{\tau^2 + \phi^2}$".

When dealing with GMMs in PSHA, epistemic uncertainty is commonly taken into account by selecting more than one GMM, which are then combined within a logic tree framework (e.g. Bommer et al., 2005). Problems are which models to select and how to assign the weights for the logic tree (e.g. Bommer and Scherbaum, 2005). These issues, however, are not the concern of the present work but are treated elsewhere (Cotton et al., 2006; Bommer et al., 2010; Scherbaum et al., 2004a, 2009). Here, we are concerned with the epistemic uncertainty that is intrinsic to a specific GMM.

Usually, one gets a point estimate of the parameters when estimating a GMM, i.e. a single value for each coefficient. Even with the best strong motion datasets currently available (e.g. the NGA dataset (Power et al., 2008; Chiou et al., 2008)), it is obvious that there is uncertainty associated with these parameter estimates. These uncertainties can be quantified by the respective standard errors. However, these do not lend themselves easily to a probabilistic interpretation. Here, we want to consider the aforementioned uncertainties by using a Bayesian approach. This results in a posterior probability distribution for the parameters which reflects their uncertainty, given our present state of knowledge and the current available data. A beneficial feature of the Bayesian approach to the estimation of GMMs is also that it allows for an easy update of the model once new data is available. The Bayesian approach has been used in e.g. Ordaz et al. (1994) or Wang and Takada (2009) for the prediction of seismic ground motion. Recently, Arroyo and Ordaz (2010a,b) have presented a study where they compare the relative merits of maximum-likelihood (ML) and Bayesian regression. They come to the conclusion that the Bayesian approach leads to better results than ML, in particular when data is sparse.

Traditionally, GMMs are derived separately for one ground motion parameter as the target variable, which is often PGA or pseudo-spectral acceleration (PSA) at discrete periods. However, it has been recognized that ground motion parameters recorded at one station are not independent from each other (e.g. Baker and Cornell, 2006). If this is not taken into account during PSHA and subsequent reliability analysis, it can lead to misleading or wrong results (e.g. Baker, 2007). Normally, the correlation between ground motion intensity parameters is investigated using the residuals given a GMM that was estimated separately for each parameter. By contrast, here we directly develop a model for all target variables under consideration which takes into account the covariance between these parameters. Thus, our work is similar to Arroyo and Ordaz (2010a,b), who investigate a multivariate Bayesian regression model for ground motions. Our model differs from their approach in the design of the covariance structure, which we set up in as a multilevel model, while Arroyo and Ordaz (2010a,b) follow Joyner and Boore (1993, 1994). Both ways are very similar, but the multilevel model allows higher computational flexibility.

We develop our model in the framework of probabilistic graphical models (see e.g. Koller and Friedman, 2009). These provide a general framework for reasoning under uncertainty, which can be exploited for use in PSHA. For example, it is possible to model measurement uncertainties or even different functional forms in the graphical model framework. Due to their modular structure, they are also easy to extend.

4.2 Introduction to Bayesian Inference

Bayesian inference is a key concept in our analysis. Therefore, we deem it necessary to provide a brief, though non-exhaustive introduction to the underlying principles of Bayesian inference/regression. A good overview of Bayesian statistics is presented by Spiegelhalter and Rice (2009, online available at `http://www.scholarpedia.org/article/Bayesian_statistics`). For a more thorough introduction, see e.g. Gelman et al. (2003).

A key notion of Bayesian statistics is a proper treatment of (epistemic) uncertainty in terms of probabilities. As such, the goal of Bayesian inference is not to to estimate one particular model, but rather a distribution of (likely) models. Therefore, all information/belief that we have about the physics of the problem at hand is specified in terms of a probability distribution defined on the parameters involved. This distribution is the so-called prior distribution, which is then subsequently updated given data using Bayes' law. In the following, Bayesian inference is illustrated by means of a simple regression example.

Imagine that we have data, $\boldsymbol{\mathcal{D}}$ on two variables, X and Y, with $i = 1, \ldots, N$ samples. We assume that there is a linear dependency between X and Y,

$$Y_i = w_0 + w_1 * X_i + \epsilon_i, \tag{4.3}$$

where ϵ is the error term which is Normal distributed with mean 0 and standard deviation σ. This defines a classical regression problem, which can be solved using e.g. maximum likelihood, giving us a point estimate of the parameters w_0, w_1 and σ.

We can rewrite eq. (4.3) to emphasize the stochastic nature of the data Y and to explicitly express

that the parameters are treated as random variables:

$$Y_i \sim \Pr(Y|w_0, w_1, \sigma; X_i) = \mathcal{N}(\mu_i = w_0 + w_1 * X_i, \sigma). \tag{4.4}$$

Eq. (4.4) reads as "Y_i is distributed according to a Normal distribution with mean $\mu_i = w_0 + w_1 * X_i$ and standard deviation σ".

In Bayesian regression, we are interested in the posterior distribution of the parameters given the data, which can be estimated using Bayes' rule

$$\Pr(\boldsymbol{\Theta}|\boldsymbol{\mathcal{D}}) = \Pr(\boldsymbol{\Theta}) * \Pr(\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta}) / \Pr(\boldsymbol{\mathcal{D}}), \tag{4.5}$$

where $\boldsymbol{\Theta} = \{w_0, w_1, \sigma\}$ denotes the set of parameters of the model, $\boldsymbol{\mathcal{D}}$ denotes the data, $\Pr(\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta})$ is the likelihood function and $\Pr(\boldsymbol{\Theta})$ is the prior distribution of the parameters $\boldsymbol{\Theta}$. The denominator $\Pr(\boldsymbol{\mathcal{D}})$ is the marginal distribution of the data and can be obtained from the numerator by integrating out the parameters:

$$\Pr(\boldsymbol{\mathcal{D}}) = \int \Pr(\boldsymbol{\Theta}) * \Pr(\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta}) d\boldsymbol{\Theta}. \tag{4.6}$$

Equation (4.5) can be interpreted as updating the prior distribution of the parameters using data (via the likelihood function), resulting in the posterior distribution. It is possible to estimate the posterior distribution for a whole batch of data points. Alternatively, one can also update the model sequentially, using one data point at a time, in which case the posterior distribution given the first data point becomes the prior distribution for the second and so on. The latter procedure can be used to easily update a Bayesian model once new data becomes available.

To solve our regression example eq. (4.3), we need to specify our prior information/beliefs of the parameters w_0, w_1 and σ into a probability distribution. In conjunction with the likelihood function $\Pr(Y|\{w_0, w_1, \sigma\})$ we can estimate the posterior distribution using eq. (4.5).

4.3 Graphical Models

In the previous section we have illustrated the basic ideas behind Bayesian inference, where we are interested in the conditional distribution of the parameters given the data, $\Pr(\boldsymbol{\Theta}|\boldsymbol{\mathcal{D}})$. In simple cases this distribution may be solved analytically, but in many realistic situations the posterior distribution is high-dimensional, complex, and unavailable in closed form. This is primarily due to analytical complications in computing the marginal likelihood. Therefore, one needs to resort to approximate inference methods, such as Markov Chain Monte Carlo (MCMC) sampling of the posterior distribution (e.g. Gilks et al., 1996). MCMC constructs a Markov Chain with the stationary/limiting distribution being the posterior. Here, we will use MCMC to obtain the parameters of our GMM. In particular, we make use of the program OpenBUGS (`http://www.openbugs.info/`) (Lunn et al., 2009), where BUGS stands for 'Bayesian inference using Gibbs sampling'.

Gibbs sampling (Geman and Geman, 1984) is one particular MCMC algorithm that exploits conditional independence assumptions between parameters and (un)observables of a data generat-

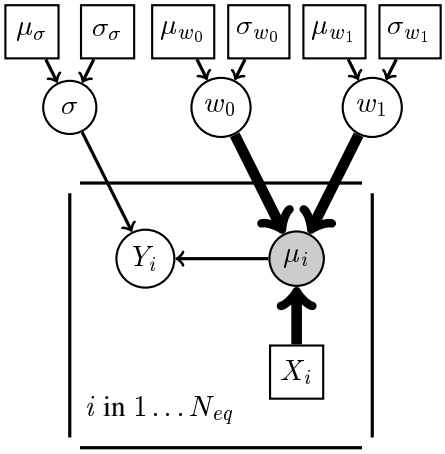

Figure 4.1: Graphical model for a simple linear model of the form $y(x, \boldsymbol{w}) = w_0 + w_1 x + \epsilon$

ing system such as the one described in eq. (2). An easy way to encode conditional (in)dependence statements are graphical models, in particular *directed acyclic graphs* (DAGs; e.g. Spiegelhalter, 1998), which we describe below. For a more detailed introduction to graphical models, see e.g. Jordan (2004), Koller et al. (2007) or Koller and Friedman (2009).

Each quantity (observable or parameter) of a model corresponds to a node in the graphical model, and arcs between the nodes show direct dependences. The graphical model for our simple regression example [eq. (4.3)] is depicted in Figure 4.1. The graph is a DAG since each edge (connection) is an arrow, and it is acyclic because there is no direct path following the arrows from one node back to itself.

In a graphical model, non-random variables or co-variates are denoted by a rectangular node. In our example, the data points X_i are considered known and are thus represented by a rectangle. Elliptical or circular nodes represent either output of mappings (here also shaded) or stochastic quantities. In Fig. 4.1 and eq. (4.4), μ_i is a function of w_0, w_1 and X_i,

$$\mu_i = w_0 + w_1 * X_i. \tag{4.7}$$

The functional dependence is denoted by a thick arrow in Figure 4.1. Thin arrows represent stochastic dependences. For example, in our case Y_i coincides with a stochastic node associated with a Normal distribution with mean μ_i and standard deviation σ. From a Bayesian perspective the nodes for w_0, w_1 and σ also represent stochastic quantities, which need to be assigned a prior distribution. If we assume a normal prior distribution for each of these parameters, the means and standard deviations of these Gaussians are represented by the nodes in the top line of Figure 4.1. These nodes are rectangles, because a fixed prior distribution is assumed. Repetitive structures (in our case the loop over the data points from $i = 1$ to $i = N$) are represented by rectangular structures, so-called 'plates'.

The advantage of a graphical model (DAG) is that it keeps details about distributions and deterministic functions hidden, but communicates the essence (i.e. the direct (in)dependences of

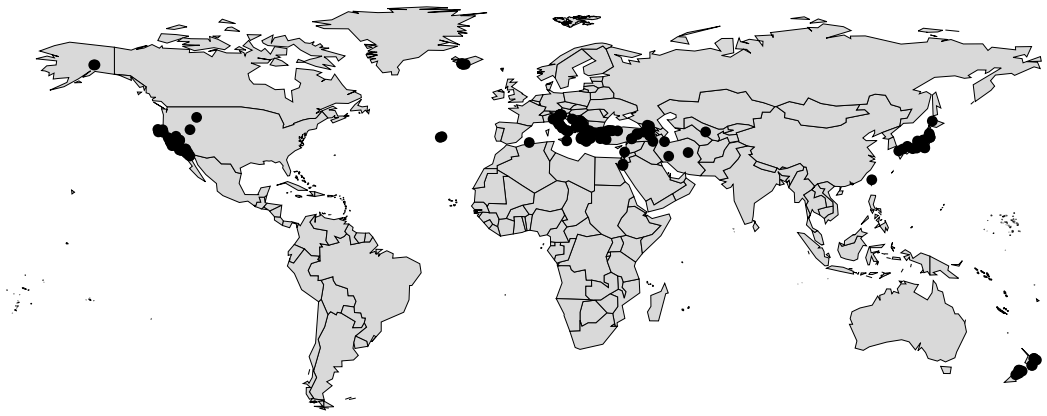

Figure 4.2: Earthquakes used in the study.

variables) of a model. This is especially useful for complex models, where we otherwise would have to resort to a large set of equations.

The DAG representation also facilitates analysis of probability models, since it encodes conditional independence statements and allows a factorization of the joint probability distribution. It can be shown (Lauritzen et al., 1990) that for any particular

$$\Pr(\boldsymbol{\Theta}|\boldsymbol{Y};\boldsymbol{X}) \propto \Pr(\boldsymbol{\Theta},\boldsymbol{Y};\boldsymbol{X}) = \prod_{V \in \boldsymbol{Y} \cup \boldsymbol{\Theta}} \Pr(V|parents\,[V];\boldsymbol{X}), \tag{4.8}$$

where $parents[v]$ specifies the parent set of the nodes from which an arrow points to V (if the parent set is empty the conditional reduces to a marginal). Hence, to specify the full joint probability distribution it is sufficient to define the local "parent-to-child" distributions along the arcs of the DAG, $P(v|parents\,[v];\boldsymbol{X})$. Gibbs sampling is a technique that makes efficient use of these properties by passing information around the DAG in accordance to the independences holding and depending on the local distributions defined.

4.4 Model Setup

In this study, our intention is to learn a Bayesian GMM. The central formula in this context is Bayes rule, eq. (4.5), which we repeat here:

$$\Pr(\boldsymbol{\Theta}|\boldsymbol{\mathcal{D}}) = \Pr(\boldsymbol{\Theta}) * \Pr(\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta}) / \Pr(\boldsymbol{\mathcal{D}}). \tag{4.9}$$

Hence, we need three ingredients: a dataset $\boldsymbol{\mathcal{D}}$, the likelihood function $\Pr(\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta})$, which is given by the data generating model, and the prior distribution of the parameters $\Pr(\boldsymbol{\Theta})$ (they in turn fully define the denominator, $\Pr(\boldsymbol{\mathcal{D}})$). In the following, we describe these parts for our GMM.

Table 4.1: Site categories based on V_S30.

Parameter	V_S30 range	No. of records
SOFT SOIL	200 - 360	2279
STIFF SOIL	360 - 660	2753
ROCK	660 - 1000	990
unknown		1935

4.4.1 Dataset

The dataset we use for constructing the Bayesian GMM is the one compiled by Allen and Wald (2009), which is a global dataset. It contains records from earthquakes in three different tectonic source types: shallow active tectonics, subduction zone and continental interiors. Here, we use only earthquakes from shallow active tectonic regimes.

The dataset of Allen and Wald (2009) comprises 10,163 records from 238 earthquakes from shallow active tectonic regimes. For details on the records compilation, we refer to Allen and Wald (2009). In this work, we use only records up to a rupture distance of 400 km from earthquakes with a moment magnitude greater than 5., which reduces the dataset to 9,872 records from 228 earthquakes.

In the dataset, information on peak ground velocity (PGV), PGA and PSA at 0.3s, 1s and 3s is available. The dataset was originally compiled to reconstruct ground-shaking from recent-historical earthquakes (Allen et al., 2009) using USGS ShakeMap methodology (Wald et al., 1999). Consequently, the target variables are taken to be the the larger horizontal component. We use only those records for which all five target quantities (PGV, PGA and PSA at 0.3s, 1s and 3s) are available, which leaves us with 7,957 records from 159 earthquakes, recorded at 2,889 unique stations. The epicenters of the earthquakes are depicted in Figure 4.2.

As predictor variables, we consider moment magnitude M_W, rupture distance R_{RUP}, the average shear wave velocity in the upper 30 m, V_S30, and the focal mechanism FM. The focal mechanism has three states, normal, strike slip, and reverse. Shear wave velocity values were estimated from topographic gradient using the approach of Wald and Allen (2007), with the minimum and maximum V_S30 values of 210 m/s and 963.9 m/s, respectively. We group V_S30 into three site categories according to Table 4.1.

For some earthquakes, there is no information on the focal mechanism. Similarly, for some stations the value of V_S30 is missing. Nevertheless, the corresponding records can be retained in the analysis, as the uncertainty of the unknown values is taken care of during the analysis. For more details, see section 4.4.2.

In Figure 4.3, we show the magnitude-distance distribution of the used records. Scatter plots between the predictor variables and PGA are depicted in Figure 4.4. We provide Tables with detailed information on the used earthquakes and records in the electronic supplement.

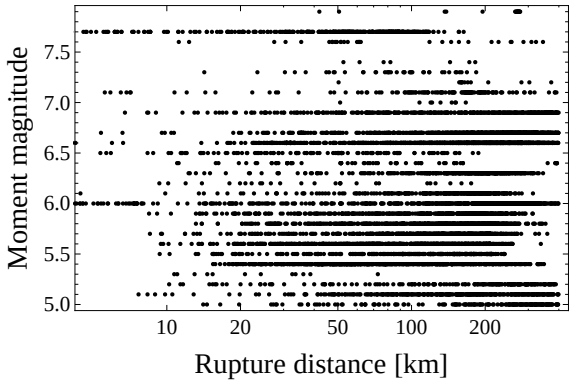

Figure 4.3: Magnitude vs. rupture distance distribution of the records that are used in this study.

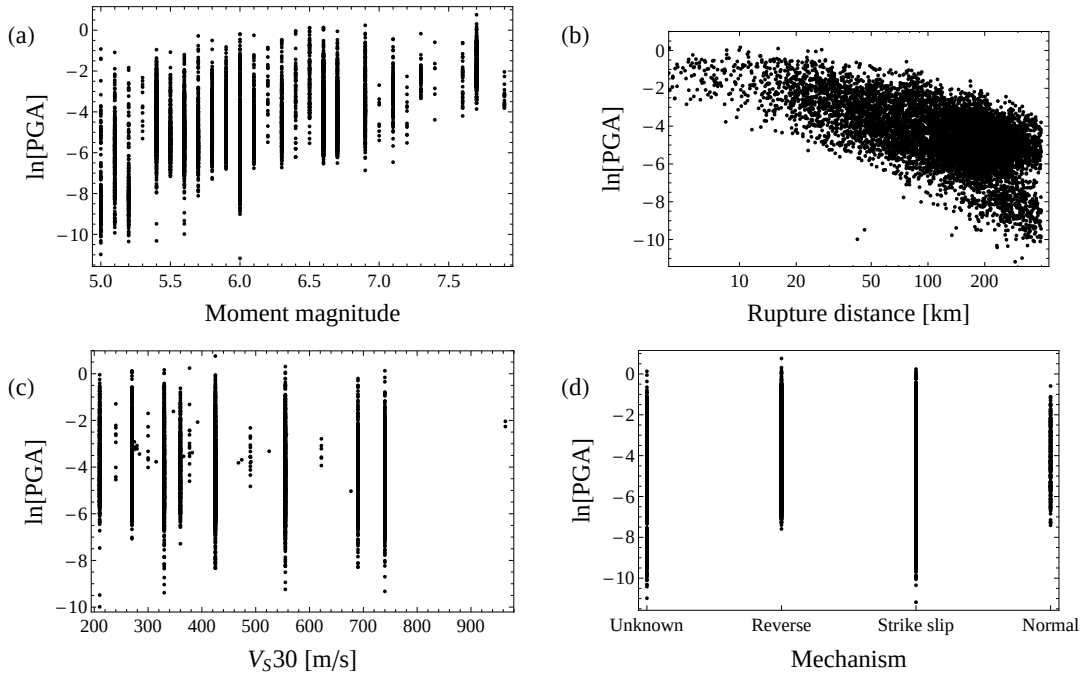

Figure 4.4: Scatter plots between PGA and (a) moment magnitude, (b) rupture distance, (c) V_S30 and (d) focal mechanism.

4.4.2 Ground Motion Model Setup

In this section, we describe the setup of the model connecting the vector of predictor variables $\boldsymbol{X} = \{M_W, R_{RUP}, V_S30, FM\}$ with the vector of target variables $\boldsymbol{Y} = \{PGA, PGV, PSA(0.3s), PSA(1s), PSA(3s)\}$. We build a multivariate model that directly estimates the joint distribution of the target vector as a function of the predictors, $\boldsymbol{Y} = f(\boldsymbol{X})$.

Almost all published GMMs make the assumption that PGA, PGV and the response spectrum are log-normally distributed, which we follow here. Thus, we introduce a new vector of random variables, $\boldsymbol{Z}$, which is the log-transformed target vector $\boldsymbol{Y}$,

$$\boldsymbol{Z} = \ln \boldsymbol{Y} \tag{4.10}$$

We then assume that $\boldsymbol{Z}$ is distributed according to a multivariate normal distribution with a vector of means $\boldsymbol{\mu}$, that are functions of the predictor variables, and a covariance matrix Σ.

In the development of a GMM, one has to take into account the correlation of records from the same earthquake. This is usually done by invoking an appropriate regression technique that allows for separation of the total variability into between-event and within-event standard deviation, such as a one-step or two-step regression (Joyner and Boore, 1993, 1994) or a random effects algorithm (Abrahamson and Youngs, 1992). Another source of variability is between-station variability, which takes into account that ground motions recorded at the same station are not independent. However, this variability is only rarely taken into account, e.g. in the work of Chen and Tsai (2002).

In the following, we use the notation introduced by Al Atik et al. (2010) to discern the different components of ground motion variability. Between-event variability is denoted by τ, while ϕ stands for within-event variability. The respective covariance matrices are denoted by upper case letters, i.e. T and Φ.

We develop our GMM as a multilevel/hierarchical model (Gelman and Hill, 2007), which can be seen as conceptually similar to a two-step regression but with the ability of easily adding extra complexity. The multiple levels allow to take into account grouped data. Hence, one level corresponds to all earthquakes, one level corresponds to all stations, and the intersection of these two levels represents the record of one earthquake recorded at one station.

In Figure 4.5, we show the GMM of this study as a graphical model. The two plates correspond to the two levels, where indices e and s denote the eth earthquake and sth station, respectively. Figure 4.5 can be thought of as a conceptual model of the data generation.

The concept of the model is as follows:

$$\boldsymbol{Z}_{es} \sim \mathcal{N}_P(\boldsymbol{\mu}_{Z,es}, \Phi) \tag{4.11}$$

$$\boldsymbol{\mu}_{Z,es} = \boldsymbol{\mu}_{\mathcal{R},es} + \boldsymbol{\mathcal{E}}_e + \boldsymbol{\mu}_{\mathcal{S},s} \tag{4.12}$$

$$\boldsymbol{\mathcal{E}}_e \sim \mathcal{N}_P(\boldsymbol{\mu}_{\mathcal{E},e}, \mathrm{T}) \tag{4.13}$$

The central node in Figure 4.5 is the vector of target variables, denoted by $\boldsymbol{Z}_{es}$, which is the eth earthquake recorded at the sth station. $\boldsymbol{Z}_{es}$ is distributed according to a multivariate normal distribution with mean vector $\boldsymbol{\mu}_{Z,es}$ and covariance matrix Φ. The mean vector $\boldsymbol{\mu}_{Z,es}$ is the sum of an event term, a station term and a record term, as can be seen in eq. (4.12). The event term $\boldsymbol{\mathcal{E}}_e$ is common to all records from the same earthquake e and is itself distributed according to a multivariate normal distribution, with mean vector $\boldsymbol{\mu}_{\mathcal{E},e}$ and covariance matrix T [eq. (4.13)]. Correspondingly, the station term $\boldsymbol{\mu}_{\mathcal{S},s}$ is common to all records recorded at the same station k. In principle, one could assume that the station term is also sampled from a multivariate normal distribution. However, a stable estimation of its covariance requires stations with multiple recordings, which are not

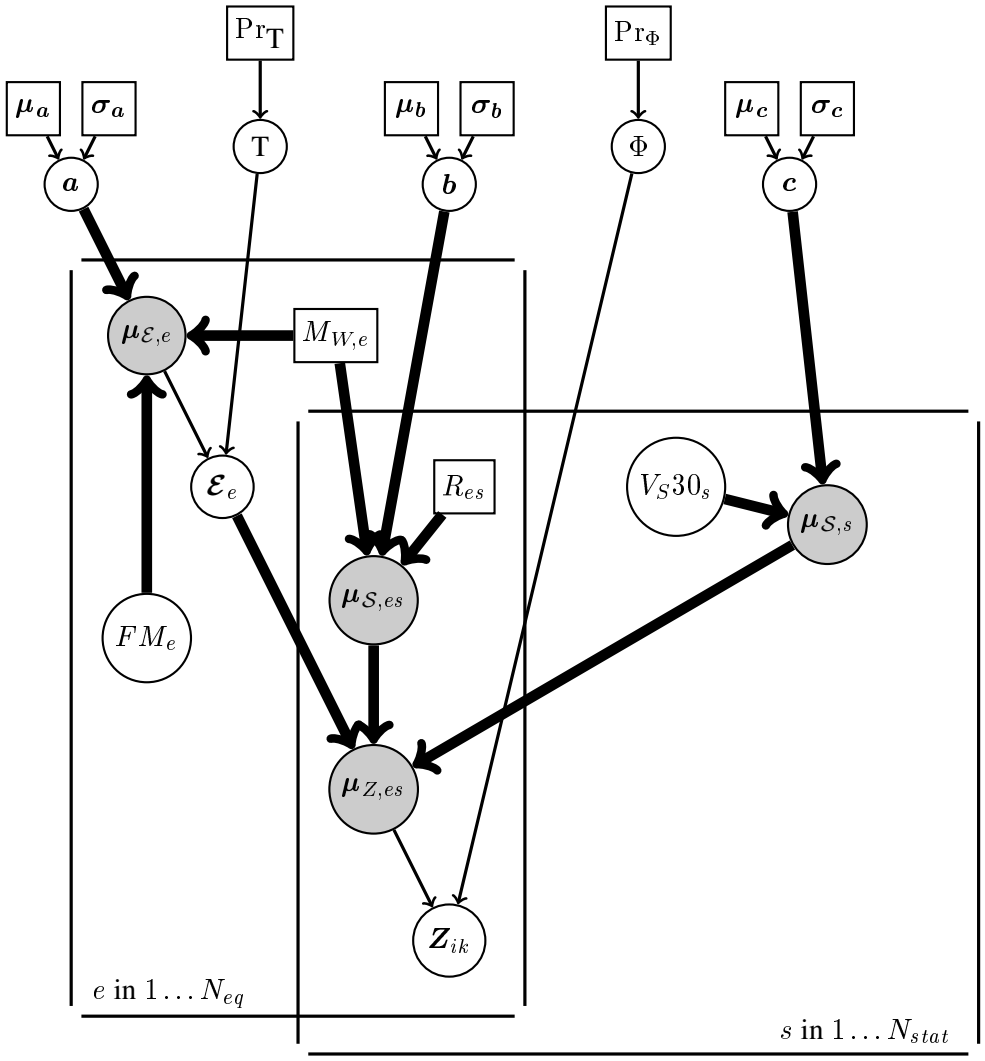

Figure 4.5: Graphical ground motion model.

abundant in our dataset. Hence, we assume that the station term is not a random variable, but a constant (strictly speaking, we assume that the components of the covariance matrix are all zero). Φ, T are the within-event and between-event covariances, respectively.

The means of the event, record and station terms are functions of parameters and the predictor variables:

$$\mu^t_{\mathcal{R},es} = f^t(R_{RUP,es}, M_{W,e}, \boldsymbol{b}^t) \tag{4.14}$$
$$\mu^t_{\mathcal{E},e} = g^t(M_{W,e}, FM_e, \boldsymbol{a}^t) \tag{4.15}$$
$$\mu^t_{\mathcal{S},s} = h^t(V_S30_s, \boldsymbol{c}^t) \tag{4.16}$$

where superscript t denotes the tth target variable, while $\boldsymbol{a}^t$, $\boldsymbol{b}^t$ and $\boldsymbol{c}^t$ are the coefficients for the individual functions. These parameters are all assumed to be independent of each other.

The parameters $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, as well as the covariances Φ and T, are displayed as stochastic nodes, since they are treated as random variables. Therefore, they are assigned a prior distribution, whose parameters are represented by the rectangular (i.e. fixed) nodes of Figure 4.5. For $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$, the prior distributions are independent univariate normal distribution with means $\boldsymbol{\mu_a}$, $\boldsymbol{\mu_b}$, $\boldsymbol{\mu_c}$ and standard deviations $\boldsymbol{\sigma_a}$, $\boldsymbol{\sigma_b}$, $\boldsymbol{\sigma_c}$, respectively. For the covariances Φ and T, the prior distributions, denoted Pr_Φ and Pr_T, are uniform distributions over their respective Cholesky decompositions (Weisstein, 2010). For more details on the prior distributions, see section 4.4.3.

We have settled on the following functional forms for f, g and h. These are based on geophysical considerations and generalization error determined by 10-fold cross-validation (e.g. Kuehn et al., 2009a; Hastie et al., 2001).

$$\begin{aligned} g^t(\boldsymbol{X}, \boldsymbol{a}^t) = {} & a^t_0 + a^t_1 * M_{W,e} + a^t_2 * (M_{W,e} - 5.5) * H(M_{W,e} - 5.5) + \\ & a^t_3 * (M_{W,e} - 6.5) * H(M_{W,e} - 6.5) + \\ & a^t_4 * F_{R,e} + a^t_5 * F_{N,e} \end{aligned} \tag{4.17}$$
$$f^t(\boldsymbol{X}, \boldsymbol{b}^t) = (b^t_0 + b^t_1 * M_{W,i}) * \ln\sqrt{R^2_{Rup,es} + (b^t_2)^2} + b^t_3 * R_{Rup,es} \tag{4.18}$$
$$h^t(\boldsymbol{X}, \boldsymbol{c}^t) = c^t_1 * S_{A,s} + c^t_2 * S_{S,s} \tag{4.19}$$

F_R and F_N are dummy variables taking the value one for reverse and normal faulting, respectively, and zero otherwise. S_A and S_S are dummy variables equaling one for stiff soil and soft soil, respectively, and zero for rock. $H(x)$ is the Heavyside-function which equals one for $x \geq 0$ and zero for $x < 0$.

We have settled for a trilinear magnitude scaling instead of a quadratic magnitude term since it results in a slightly lower generalization error and allows more control over the magnitude scaling. It also effectively decouples the large magnitude scaling from the small magnitude scaling. The "geometrical spreading term", $(b^t_0 + b^t_1 * M_{W,e}) * \ln\sqrt{R^2_{Rup,es} + (b^t_2)^2}$, is chosen because it gives a lower generalization error than when using a magnitude term inside the root. We include an "anelastic attenuation term", $b^t_3 * R_{Rup,ik}$, since the maximum distance in the dataset is 400 km (cf. Figure 4.3).

As one can see, V_S30 and the focal mechanism FM are displayed as stochastic nodes in Figure

4.5. This is due to the fact that some of these values are missing, i.e. there are some stations without an associated V_S30 value and some earthquakes with unknown focal mechanism. We treat these unknown values simply as parameters - they are assigned a prior distribution, which is updated by the likelihood resulting in a posterior distribution for each unknown V_S30 or FM value. This is a nice feature of the model, which thus provides a principled way to deal with missing data.

4.4.3 Prior Distributions

Bayesian inference works by updating the prior distributions of the parameters, $\Pr(\boldsymbol{\Theta})$, with the likelihood of the data given the current model, $\Pr(\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta})$, which results in the posterior distribution of the parameters given the data, $\Pr(\boldsymbol{\Theta}|\boldsymbol{\mathcal{D}})$ [cf. eq. (4.5)]. The specification of the prior distributions is an important step, and should be done with care. However, even though we have some general knowledge about the scaling of PGV, PGA and PSA with source, path and site effects, it is difficult to quantify our prior information into a probability distribution on, say, the parameter a_2.

By contrast, it is easier to specify a prior distribution on physical parameters such as average stress drop or quality factor Q_0. Therefore, our strategy for specifying prior distributions is as follows:

1. Specify prior distributions on the stress drop, Q_0 and the slope of the geometric attenuation.
2. Generate a synthetic dataset from these parameters using stochastic simulations (Boore, 2003).
3. Regress the function of eqs. (4.17) to (4.18) on the synthetic dataset.
4. Take the estimated coefficients and their standard errors as prior distributions for the parameters.

For the stress drop, we assume a truncated normal distribution between 1 and 10 MPa, with mean 4 MPa and standard deviation 2 MPa, loosely based on Allmann and Shearer (2008). The prior distribution for Q_0 is a truncated normal distribution between 10 and 1000, with mean 250 and standard deviation 50, roughly based on Dalton et al. (2008). We also set a prior distribution on the slope of the geometrical spreading, which is a truncated normal distribution between -1.5 and -0.8 with mean -1 and standard deviation 0.2. The stochastic simulations are carried out with the program SMSIM (Boore, 2005).

This approach can be used to determine prior distributions for the parameters a_0^t, a_1^t, a_2^t, a_3^t, b_0^t, b_1^t, b_2^t and b_3^t. For a_4^t and a_5^t, which describe the scaling of the ground motion intensity parameters with focal mechanism, we use Table III from Bommer et al. (2003). The site effects parameter μ_{c_1} and μ_{c_2} are assigned distributions based on guesses by the authors. These are loosely oriented on published GMMs as well as the work of Dobry et al. (2000). In Table 4.2 the prior distributions for all parameters are shown. To ensure that the covariance matrices Φ and T are positive definite, we place priors on the elements of their Cholesky decomposition (Weisstein 2010). The priors for the diagonal elements are $\mathcal{U}(0, 10)$ and for the non-diagonal elements $\mathcal{U}(-10, 10)$, where $\mathcal{U}(a, b)$ is a uniform distribution between a and b.

Table 4.2: Prior distributions for the parameters.

Parameter	PGV	PGA	PSA 0.3s	PSA 1s	PSA 3s
a1	$\mathcal{N}(0.746, 0.0481)$	$\mathcal{N}(1.28, 0.0454)$	$\mathcal{N}(0.997, 0.049)$	$\mathcal{N}(2.13, 0.0463)$	$\mathcal{N}(2.85, 0.0445)$
a2	$\mathcal{N}(-0.402, 0.0549)$	$\mathcal{N}(-0.388, 0.0519)$	$\mathcal{N}(-0.349, 0.0565)$	$\mathcal{N}(-1.12, 0.0538)$	$\mathcal{N}(-1.03, 0.0516)$
a3	$\mathcal{N}(-0.27, 0.0551)$	$\mathcal{N}(-0.277, 0.0522)$	$\mathcal{N}(-0.173, 0.0568)$	$\mathcal{N}(-0.206, 0.054)$	$\mathcal{N}(-0.914, 0.0518)$
a4	$\mathcal{N}(0.199, 0.1)$	$\mathcal{N}(0.122, 0.1)$	$\mathcal{N}(0.191, 0.1)$	$\mathcal{N}(0.122, 0.1)$	$\mathcal{N}(-0.105, 0.1)$
a5	$\mathcal{N}(0., 0.1)$	$\mathcal{N}(0., 0.1)$	$\mathcal{N}(0., 0.1)$	$\mathcal{N}(0., 0.1)$	$\mathcal{N}(0., 0.1)$
b0	$\mathcal{N}(-2.37, 0.0568)$	$\mathcal{N}(-2.34, 0.0529)$	$\mathcal{N}(-1.69, 0.0562)$	$\mathcal{N}(-1.24, 0.0518)$	$\mathcal{N}(-1.3, 0.0498)$
b1	$\mathcal{N}(0.172, 0.00793)$	$\mathcal{N}(0.186, 0.00747)$	$\mathcal{N}(0.0854, 0.00805)$	$\mathcal{N}(0.029, 0.00759)$	$\mathcal{N}(0.0367, 0.00729)$
b2	$\mathcal{N}(2.36, 0.419)$	$\mathcal{N}(1.8, 0.402)$	$\mathcal{N}(1.43, 0.461)$	$\mathcal{N}(0.48, 0.748)$	$\mathcal{N}(0.52, 0.666)$
b3	$\mathcal{N}(-0.00504, 0.000197)$	$\mathcal{N}(-0.00322, 0.000181)$	$\mathcal{N}(-0.00696, 0.000194)$	$\mathcal{N}(-0.00458, 0.000176)$	$\mathcal{N}(-0.00257, 0.000169)$
c1	$\mathcal{N}(0.1, 0.1)$	$\mathcal{N}(0.1, 0.1)$	$\mathcal{N}(0.1, 0.1)$	$\mathcal{N}(0.1, 0.1)$	$\mathcal{N}(0.1, 0.1)$
c2	$\mathcal{N}(0.05, 0.1)$	$\mathcal{N}(0.05, 0.1)$	$\mathcal{N}(0.05, 0.1)$	$\mathcal{N}(0.05, 0.1)$	$\mathcal{N}(0.05, 0.1)$

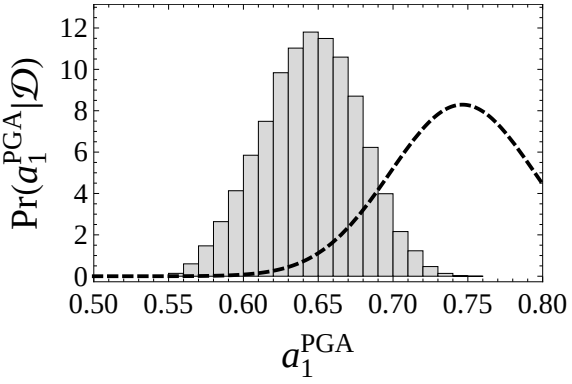

Figure 4.6: Normalized histogram of MCMC samples from the posterior distribution and prior distribution (dashed line) of the parameter a_1^1.

As described in section 4.4.2, we also treat missing V_S30 and FM values as unknown parameters, which are assigned a prior distribution. V_S30 as well as FM are categorical variables with three states, and we use a uniform prior over the three states for both of them.

4.5 Results

In the previous section, we have specified the model, the prior distribution and the dataset. Using Bayes' rule [eq. (4.5)], we can now estimate the posterior distribution of the parameters given the data, $\Pr(\boldsymbol{\Theta}|\boldsymbol{\mathcal{D}})$. As described before, the model is too complicated to estimate $\Pr(\boldsymbol{\Theta}|\boldsymbol{\mathcal{D}})$ analytically, so we resort to approximate inference, using MCMC sampling to obtain samples from the posterior distribution of each parameter. From the sequence of samples we can compute several summary statistics, such as mean values, standard deviations, quantiles and so on. The histogram of the sampled values serves as an approximation to the posterior probability density function.

For each target variable, our GMM has 12 parameters [cf. eqs. (4.17) to (4.19)]. This makes in total 60 parameters for the five target variables. In addition, there are 15 independent entries in each of the covariance matrices Φ and T. Hence, in total there are 90 parameter posterior distributions to estimate. We also compare the results with other simulations (see section 4.6), using different priors and/or no correlation, thus further increasing the number of parameters. We believe that it is important to show plots/comparisons of all posterior distributions. However, this would inflate the paper unnecessarily, so they are made available online in the electronic supplement.

In Figure 4.6, the approximated posterior distribution of the parameter a_1^1 is shown, which controls the scaling of PGA with magnitude [cf. eq. (4.17)]. The prior distribution of a_1^1 is also displayed in Figure 4.6. As one can see, prior and posterior distribution are different, thus illustrating how the former gets updated by the data. One can also see in Figure 4.6 that the posterior distribution of a_1^1, $\Pr(a_1^1|\boldsymbol{\mathcal{D}})$, is approximately normal. Hence, it can be fully described by the

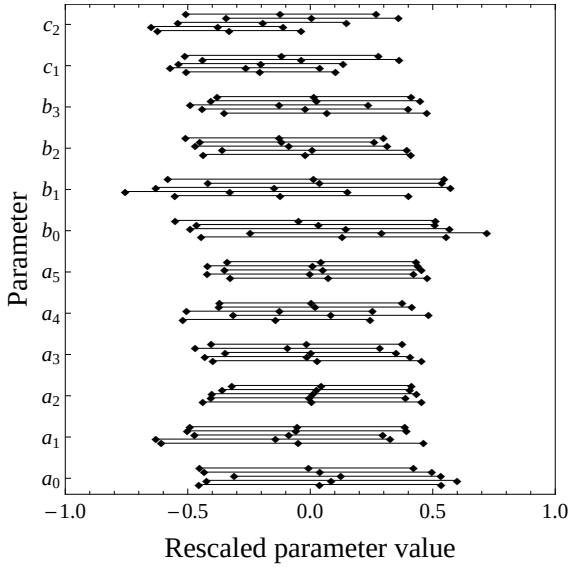

Figure 4.7: Plot of median values and 5% and 95% quantiles for posterior parameter distributions, which are rescaled to range between -1 and 1. For each parameter, five intervals are shown, corresponding to the different targets PGA (lowermost interval), PGV and PSA at 0.3s, 1s and 3s (uppermost interval).

sufficient statistics, its mean and standard deviation. In the electronic supplement, we provide plots of the posterior histogram and prior distribution for all parameters. There it can be seen that all posterior distributions can be considered approximately normal, and hence can be described by their means and standard deviations.

In Table 4.3, the means of the posterior distributions for each parameter are listed. Correspondingly, Table 4.4 contains the respective standard deviations, which allow to assess the uncertainty of each parameter. A 90% credible interval for each parameter is shown in Figure 4.7, where the 5%, 50% and 95% quantiles of each posterior distribution is plotted. Since the magnitude of each parameter is different, the posterior distributions are rescaled to range between -1 and 1 for better comparison.

The posterior distributions are unimodal and approximately symmetric (as described above, they can even be considered to be approximately normal). Hence, we can use the mean values as *maximum a posteriori* (MAP) point estimates. These can be used as a remedy to obtain point predictions. In Figure 4.8, we show the residuals between such point predictions and the data. Both between-event and within-event residuals are shown. As one can see, there is no obvious trend of the residuals with magnitude or distance. However, the event terms are slightly overpredicted. We associate this bias with the prior distribution (see also the discussion). There are only 159 data points to "pull away" the distributions of the event-related parameters ($\boldsymbol{a}$) from the prior, while there are 7,957 for the record-related ones ($\boldsymbol{b}$). This results in a higher influence of the prior distribution on the former than on the latter. Nevertheless, the bias is small, and in general the model predicts the data well.

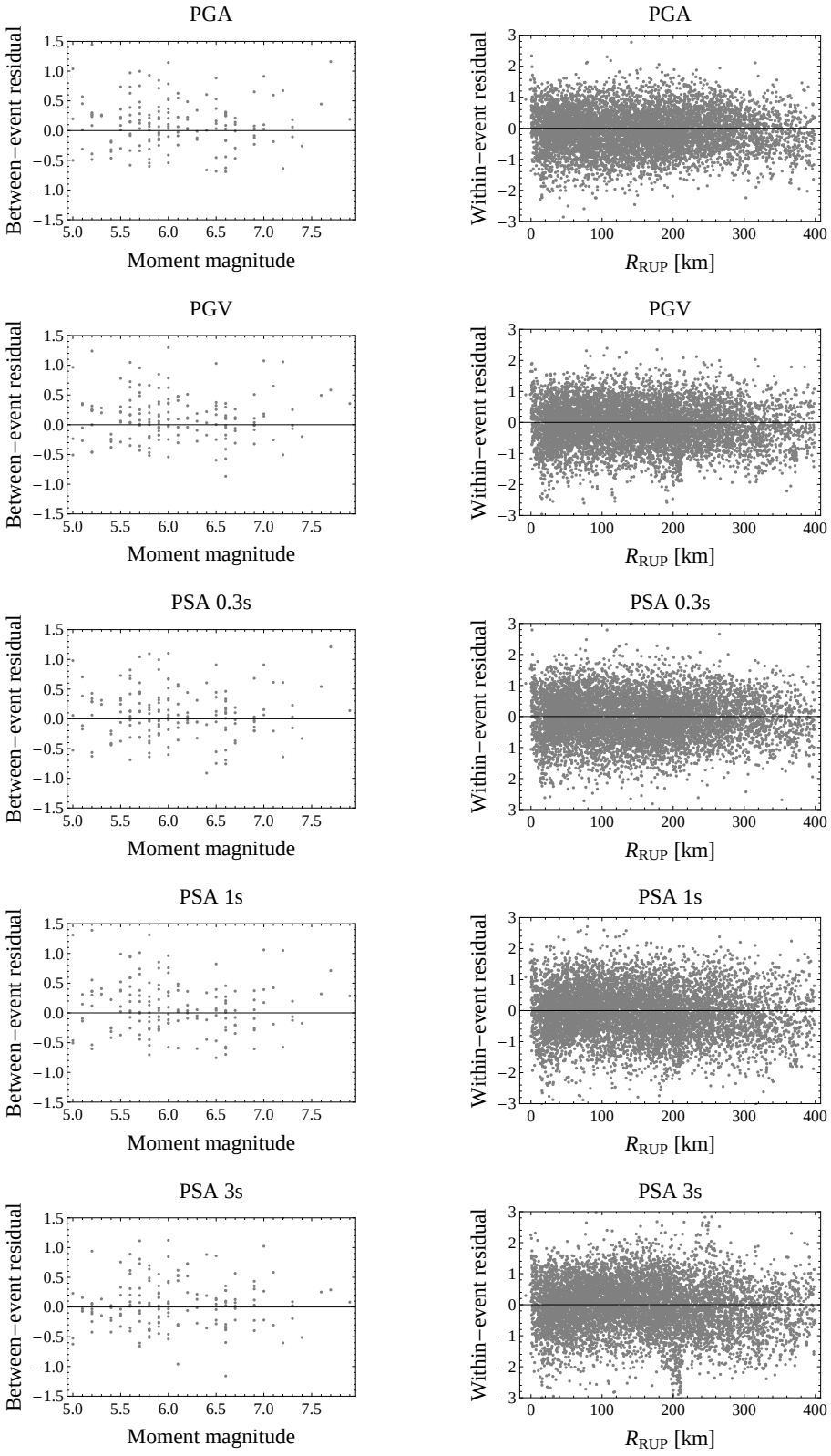

Figure 4.8: Between-event and within-event residuals, calculated with the mean values of the parameter posterior distributions. The residuals are calculated as $r = \ln \hat{Z} - \ln Z$, where $\hat{Z}$ and Z are the predicted and observed ground motion intensity value, respectively.

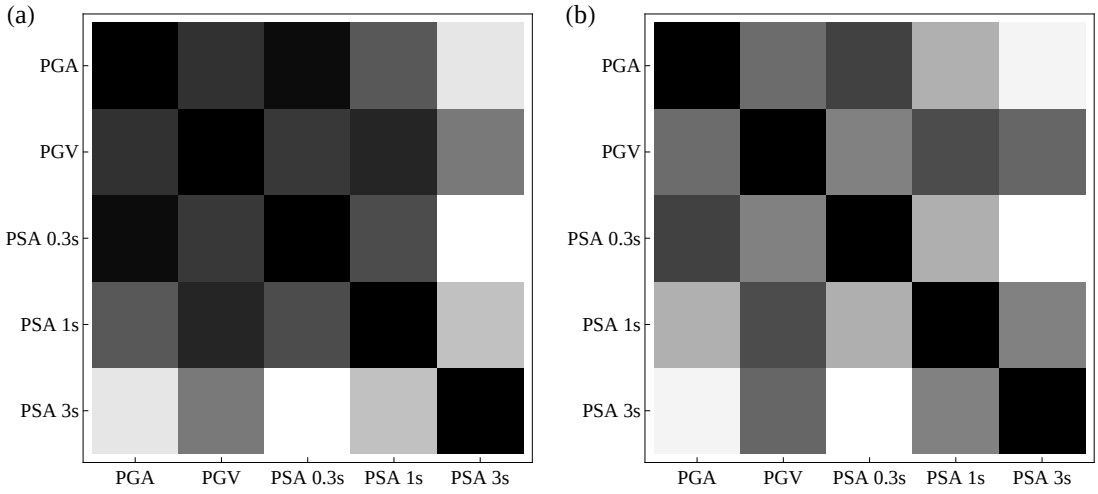

Figure 4.9: Correlation between different ground-motion predictor variables, gray shaded from white (0) to black (1). (a) Between-event correlation (b) Within-event correlation.

Since our GMM is a multivariate model, we can directly estimate the covariance structure between the target variables. As for the parameters, each entry in the covariance matrices Φ and T is associated with a posterior distribution. The histograms of these distributions are shown in the electronic supplement, where it can be seen that they can be considered normal, as is the case for the parameters. The respective means and standard deviations are listed in Tables 4.5, 4.6, 4.7 and 4.8. Using again the means as point estimates, we can calculate the correlation coefficients between the different target variables via

$$\text{corr}_{ij} = \frac{\text{cov}_{ij}}{\sigma_i \sigma_j}. \tag{4.20}$$

The between-event and within-event correlations are shown in Figure 4.9. Here, black indicates a value of one, while white corresponds to zero correlation. It is obvious from Figure 4.9 that the between-event correlation is larger than the within-event correlation.

4.6 Discussion and Conclusions

We have obtained a ground motion model for the target variables PGA, PGV and PSA at 0.3s, 1s and 3s. Since the model is a multivariate one, we can directly, during the learning phase, estimate the covariance structure, i.e. the correlations between the five targets. In particular, the model directly presents an assessment of both between-event and within-event covariance. By contrast, usually studies that investigate correlations between different ground motion intensity parameters study total residuals (though recently also between-event and within-event residuals have been

Table 4.3: Mean values of posterior distributions for the parameters.

Parameter	PGA	PGV	PSA 0.3s	PSA 1s	PSA 3s
a_0	-2.189	-6.482	-3.821	-11.771	-17.177
a_1	0.645	1.243	0.914	2.007	2.688
a_2	-0.429	-0.592	-0.312	-0.992	-1.007
a_3	-0.263	-0.349	-0.166	-0.198	-0.884
a_4	0.197	0.0754	0.151	0.0808	-0.0737
a_5	-0.0962	-0.0618	-0.00670	0.0639	-0.0114
b_0	-2.196	-2.064	-1.505	-1.181	-1.258
b_1	0.185	0.166	0.0991	0.0571	0.0509
b_2	3.857	2.319	2.569	1.530	1.802
b_3	-0.00629	-0.00279	-0.00728	-0.00390	-0.00183
c_1	0.210	0.591	0.288	0.754	0.868
c_2	0.0437	0.154	0.0915	0.227	0.209

Table 4.4: Standard deviations of posterior distributions for the parameters.

Parameter	PGV	PGA	PSA 0.3s	PSA 1s	PSA 3s
a_0	0.185	0.160	0.188	0.178	0.182
a_1	0.0322	0.0271	0.0321	0.0301	0.0306
a_2	0.0409	0.0379	0.0433	0.0434	0.0451
a_3	0.0454	0.0408	0.0472	0.0490	0.0493
a_4	0.0508	0.0435	0.0541	0.0567	0.0687
a_5	0.0564	0.0486	0.0599	0.0641	0.0757
b_0	0.0450	0.0404	0.0411	0.0398	0.0403
b_1	0.00644	0.00586	0.00586	0.00576	0.00593
b_2	0.293	0.207	0.267	0.275	0.307
b_3	0.000124	0.000100	0.000127	0.000122	0.000129
c_1	0.0259	0.0246	0.0287	0.0297	0.0312
c_2	0.0248	0.0236	0.0276	0.0286	0.0300

Table 4.5: Means of posterior distributions for the between-event covariance T.

	PGA	PGV	PSA 0.3s	PSA 1s	PSA 3s
PGA	0.233	0.202	0.242	0.208	0.156
PGV	0.202	0.212	0.212	0.221	0.203
PSA 0.3s	0.242	0.212	0.265	0.227	0.152
PSA 1s	0.208	0.221	0.227	0.267	0.187
PSA 3s	0.156	0.203	0.152	0.187	0.324

Table 4.6: Standard deviations of posterior distributions for the between-event covariance T.

	PGA	PGV	PSA 0.3s	PSA 1s	PSA 3s
PGA	0.0342	0.0314	0.0359	0.0341	0.0329
PGV	0.0314	0.0323	0.0334	0.0352	0.0354
PSA 0.3s	0.0359	0.0334	0.0391	0.0369	0.0346
PSA 1s	0.0341	0.0352	0.0369	0.0417	0.0380
PSA 3s	0.0329	0.0354	0.0346	0.0380	0.0487

Table 4.7: Means of posterior distributions for the within-event covariance Φ.

	PGA	PGV	PSA 0.3s	PSA 1s	PSA 3s
PGA	0.446	0.331	0.418	0.305	0.226
PGV	0.331	0.448	0.344	0.431	0.410
PSA 0.3s	0.418	0.344	0.552	0.340	0.235
PSA 1s	0.305	0.431	0.340	0.618	0.442
PSA 3s	0.226	0.410	0.235	0.442	0.657

Table 4.8: Standard deviations of posterior distributions for the within-event covariance Φ.

	PGA	PGV	PSA 0.3s	PSA 1s	PSA 3s
PGA	0.00716	0.00635	0.00738	0.00694	0.00672
PGV	0.00635	0.00725	0.00689	0.00779	0.00780
PSA 0.3s	0.00738	0.00689	0.00888	0.00772	0.00740
PSA 1s	0.00694	0.00779	0.00772	0.0100	0.00890
PSA 3s	0.00672	0.00780	0.00740	0.00890	0.0107

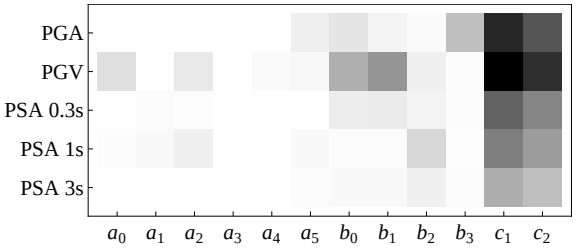

Figure 4.10: Symmetric KL-divergences between parameter posterior distributions, calculated with and without covariance between targets. The lowest value (white) is 0.00576 for a_1^{PGA}, the largest value (black) is 30.760 for c_1^{PGV}.

taken into account (Baker and Jayaram, 2008)).

In general, our analysis shows that the behavior of between-event and within-event correlation with period is similar, which is in agreement with the findings of Baker and Jayaram (2008). However, between-event correlation is generally slightly larger than within-event correlation (cf. Figure 4.9).

The largest correlation coefficients are the ones between PGA and PSA at 0.3s, and between PGV and PSA at 1s, while there is less correlation between PGA/PGV and PSA at 3s. In general, with decreasing period there is an increase in correlation with PGA, which is expected. The response spectral value at 1s has also sometimes been used to calculate PGV is not available (cf. Newmark and Hall, 1982), which has been questioned lately (Bommer and Alarcón, 2006). Our results do not support one view or the other exclusively - e.g., there is also significant correlation between PGV and PSA at 0.3s.

To investigate the influence of learning a multivariate model on the coefficients of the model, we also compute posterior distributions of the parameters where we assume that the targets are independent, as is commonly done in the derivation of GMMs. To assess the differences between the posterior distributions with and without covariance, we calculate their respective symmetric Kullback-Leibler (KL) divergences, which is a measure of the relative information loss when one probability distribution is replaced with another (see Scherbaum et al. (2009) for the use of KL-divergences in GMM selection). The KL-divergences for the parameters are shown in Figure 4.10. The largest values are obtained for the site effects coefficients c_1 and c_2. Regarding the other parameters, there are slightly larger KL-divergences for the record-specific parameters $\boldsymbol{b}$ than for the event-specific parameters $\boldsymbol{a}$. The variances, i.e. the diagonal elements of Φ and T, are very similar to the variances computed under the independence assumption. In the electronic supplement, we provide plots of all parameter posterior distributions, both with and without covariance, for comparison.

An important point in Bayesian inference is the use of the prior distribution $\Pr(\boldsymbol{\Theta})$, which provides a principled way to incorporate prior knowledge [cf. eq. (4.5)]. In section 4.4.3, we described how we came up with the prior distributions for our parameters. These priors are also used for the calculation of posterior distributions under the assumption of target independence (Figure 4.10). To investigate the influence of the prior, we also calculate parameter posterior

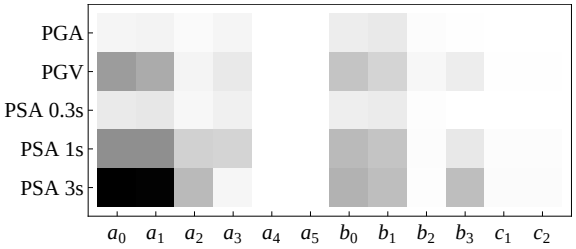

Figure 4.11: Symmetric KL-divergences between parameter posterior distributions, calculated without covariance between targets and normal/uniform prior distributions. The lowest value (white) is 0.344 for $b_3^{PSA0.3s}$, the largest value (black) is 525.254 for a_0^{PSA3s}.

distributions using a relatively flat uniform prior, assuming independence of the targets. Plots of these posterior distributions are shown in the electronic supplement. In Figure 4.11, the symmetric KL-divergences between posterior distributions with a uniform and normal prior, respectively, are shown. In contrast to the differences between the posterior distributions with and without covariance (Figure 4.10), we now get much larger values for the KL-divergences. This is due not only to differences in the mean of the distributions, but also in the spread (see electronic supplement). The largest KL-divergences are obtained for the event-specific parameters $\boldsymbol{a}$. This is due to the fact that these are essentially only determined by 159 data points (the number of earthquakes in the dataset), so the prior plays a larger role than for the record or site-specific parameters.

However, even though there is a prior dependence of the parameters, this is no a disadvantage of the model per se - in fact, it is to be expected. While the dataset of Allen and Wald (2009), which underlies our model, is quite extensive, it is by no means exhaustive. Hence, prior information can help to constrain the model. We believe that the prior we use, described in section 4.4.3, is reasonable, which in turn leads to a reasonable posterior parameter distribution. Nevertheless, model checking is very important, and we have seen in Figure 4.8 that the model predicts the data reasonably well.

The variances of the target distributions have very similar posterior distributions when computed from different priors (see electronic supplement). The distributional parameters of the covariances are listed in Tables 4.5 to 4.8. The variances of the marginal target distributions, i.e. the diagonal elements of the covariances, are relatively large. For example, the total standard deviation for PGA, calculated using the means of the respective posterior distributions (Table 4.5 and 4.7), is 0.824, compared to a value of 0.683 for the model of Akkar and Bommer (2010). A small part of this large standard deviation results from differences in how the two horizontal components are combined (Beyer and Bommer, 2006). In our study, the larger horizontal component is used, while Akkar and Bommer (2010) used the geometric mean. However, a major issue in this context are limitations in the dataset. There are measurement uncertainties associated with the predictor variables M_W, R_{RUP}, V_S30 and FM, which are larger than e.g. in the NGA dataset. For example, the V_S30 values are calculated from topographic slopes, using the method of Wald and Allen (2007). These measurement uncertainties lead to an increased ground motion variability.

Another possible source of the increased variability could be the fact that the underlying dataset is a global one, and thus combines data from different regions. Regional dependence of strong ground motions is still an open question (Douglas, 2009), but if there are differences in ground motion scaling between different regions that are not considered, these differences will map into the total ground motion variability.

We have built our GMM as a graphical model (cf. section 4.3). These provide a framework for reasoning under uncertainty, by associating each node with a probability distribution. The full joint distribution is determined by eq. (4.8). Inferences from the model can be made by using either point estimates of the parameters – such as means, medians, modes – or by sampling from the joint distribution. The factorization property of graphical models [eq. (4.8)] means that only local conditional probabilities need to be specified. This makes it very easy to extend the model, by incorporating additional nodes that might describe new aspects of the domain. Graphical models also offer great flexibility, e.g. it is possible to model directly different functional forms to incorporate epistemic model uncertainty. This is similar to a logic tree, but the weights on the functions are represented by a node and can be learned. It is also easily possible to extend the graphical model to include regional differences, where some regions might share certain coefficients.

In combination with graphical models, the Bayesian approach offers vast possibilities for reasoning under uncertainty. For one, the outcome of Bayesian inference is the conditional distribution of the parameters given the data [eq. (4.5)], which is exactly what we are interested in. In particular, strictly speaking we do not infer just one model, but a distribution of models. This makes it possible to quantify and incorporate (epistemic) uncertainties of the parameters in PSHA. The Bayesian approach also allows us to include prior knowledge in a principled way. This is especially important in the derivation of GMMs, since data is generally sparse, so other constraints need to be set.

In the future, we plan to extend our model to a general graphical hazard model. Therefore, additional functional forms should be included, as well as measurement uncertainties. This also requires additional nodes that describe the magnitude and distance distributions. The Bayesian method then allows to easily update the model once new data is acquired.

The dataset underlying our GMM ranges from moment magnitude 5-7.9 and from rupture distance 1-400 km, representing global active conditions. Its V_S30 range is 200-1000 m/s. We believe that the model is valid in this range, though it should be used with caution at the boundaries (cf. Bommer et al., 2007).

Data and Resources

The dataset used in this study is the one compiled by Allen and Wald (2009). Information about the records can be found in the electronic supplement. The MCMC sampling was done using the software OpenBUGS (`http://www.openbugs.info/w/`), version 3.06.

Electronic Supplement

The electronic supplement to this paper is available at `http://www.geo.uni-potsdam.de/mitarbeiter/Kuehn/kuehn-esupp-bayesregpsa.html`.

Acknowledgments

Trevor Allen publishes with the permission of the Chief Executive Officer of Geoscience Australia.

5

A Bayesian Hierarchical Global Ground-Motion Model To Take into Account Regional Differences

Kuehn, N. M., F. Scherbaum, C. Riggelsen, and T. I. Allen
submitted to Bulletin of the Seismological Society of America

---

In this work we present a hierarchical global ground motion model that includes regional differences in ground motion scaling in a principled way. For this purpose, we make the assumption that the scaling of ground motion intensity parameters with earthquake source, path and site parameters is similar, but not necessarily identical in different regions. In particular, we assume that models for individual regions are sampled from a global distribution of ground motion models. Thus, we set up a multi-level/hierarchical model, where the coefficients for each region are connected by so-called global hyperparameters. Via these hyperparameters, data from one region is also used to determine the coefficients in other regions, though with less weight. That way, it is possible to determine a model even for regions with a scarce amount of data. The global model is set up as a graphical model, which allows for an intuitive understanding. The coefficients are determined in a Bayesian setting using Markov Chain Monte Carlo simulation. This offers a systematic way to include prior knowledge and provides an estimate of the epistemic uncertainty of the parameters. It also allows to update the model once new data is available. The model is learned on a global dataset, divided into 10 regions. In the dataset, there are large differences in the amount of earthquakes and records between the regions. We find regional differences in scaling with large distances, while the dependance of PGA with magnitudes is not regionally different.

5.1 Introduction

Empirical ground motion models (GMMs) are a crucial ingredient for probabilistic seismic hazard analysis (PSHA). Ground motion uncertainty is one of the key factors that controls the exceedance frequency of the ground motion parameter of interest, e.g. peak ground acceleration or spectral acceleration, for a given ground motion value (e.g. Bommer and Abrahamson, 2006). There are numerous published GMMs for different seismic provinces (e.g. shallow active tectonic, stable continental interiors, and subduction zones) and different regions in the world (e.g. California, Japan or Europe). These models can differ considerably in the amount of data, the functional forms used to model the scaling of ground motions with the predictor variables, and the number and kind of predictor variables used to characterize earthquake source and site effects. For a review of published ground motion models see Douglas (2003, 2006, 2008), for a recent comparison see Douglas (2010).

An important question in the context of PSHA, especially when it comes to the applicability of a GMM, is whether ground motion scaling is regionally dependent (e.g. Douglas, 2009). This has important consequences for the possible application of a GMM, developed from data in one particular region, in another region. One way to deal with this in a PSHA is to adjust a GMM for use in a new region based on physical differences (Campbell, 2003; 2004). Recently, it has been proposed to combine data from one region with a 'reference' model from a different region which is better constrained by data (Atkinson, 2008). In this work, we present a model/approach that can be used to directly include regional differences in GMMs. For this purpose, we learn a model on a global dataset, which consists of several submodels for different regions. These submodels are not independent of each other. More technically, we assume that there is a global distribution of GMMs, specified by so-called global hyperparameters. The regional GMMs are samples from this global distribution. Hence, we assume that the coefficients for each regional model are similar, and data from one region is also used to estimate coefficients in the other regions, though with less weight. The weights are determined according to the number of data in the regions and the variability of the coefficients. For more details, see section 5.5.

Formally, a GMM is a model for the conditional distribution of a ground motion parameter Y given some earthquake and site related parameters $\boldsymbol{X}$, $\Pr(Y|\boldsymbol{X})$. Y is usually assumed to be distributed according to a log-normal distribution, whose median μ (and possibly the standard deviation σ) depends on the inputs, i.e. is a function of the predictor variables $\boldsymbol{X}$:

$$\log Y = f(\boldsymbol{X}) + \Delta. \tag{5.1}$$

Here, Δ represents the total variability of ground motion, which comprises the between-event variability, Δ_B, and the within-event variability, Δ_W. Both of these are independent, normally distributed random variables with mean zero and standard deviations τ and ϕ, respectively. The notation for the description of ground motion variability follows Al Atik et al. (2010). Eq. (5.1) is equivalent to the following notation, which emphasizes the probabilistic nature of ground motion:

$$\log Y \sim \mathcal{N}(\mu = f(\boldsymbol{x}), \sigma = \sqrt{\tau^2 + \phi^2}). \tag{5.2}$$

Eq. (5.2) reads as "$\log Y$ is distributed according to a normal distribution with mean $\mu = f(\boldsymbol{X})$

and standard deviation $\sigma = \sqrt{\tau^2 + \phi^2}$".

In PSHA, it is important to take into account epistemic uncertainty that arises from the fact that no model captures all aspects of ground motion scaling for a particular application by selecting more than one GMM. This is especially important when no GMM particularly developed for the region of interest exists. The selected GMMs are usually combined within a logic tree framework (e.g. Bommer et al., 2005). Crucial questions in this context are which models to select and how to assign the weights for the logic tree (e.g. Bommer and Scherbaum, 2005). These issues have been addressed in the past (Cotton et al., 2006; Bommer et al., 2010; Scherbaum et al., 2004, 2009, 2010) and are not the topic of this work (though they are far from being solved). Here, our concern lies rather on the epistemic uncertainty that is intrinsic to learning a GMM.

Estimating the parameters of a GMM results usually in a point estimate, i.e. a single value for each coefficient. It is obvious that these parameter estimates are not error-free, even when they are based on the best strong-motion datasets currently at hand (e.g. the NGA dataset (Power et al., 2008; Chiou et al., 2008)). When considering these uncertainties, it is important that they lend themselves to a probabilistic interpretation. This can be achieved by using a Bayesian approach, where we estimate the posterior probability of the parameters given our present state of knowledge and the current available data. The Bayesian approach also allows one to easily include prior domain knowledge, as well as to update the model once new data is available. It has been used in e.g. Ordaz et al. (1994) or Wang and Takada (2009) for the prediction of seismic ground motion. Recently, Arroyo and Ordaz (2010a,b) presented a Bayesian GMM that estimated the correlation between different ground motion intensity parameters.

Our model is developed in the framework of probabilistic graphical models (see e.g. Koller and Friedman, 2009), which provide a general framework for reasoning under uncertainty. Their graphical structure also allows for an intuitive insight into the data generating process. It is easy to extend graphical models to accommodate extra complexities by exploiting their modular structure. More technically, a joint probability distribution factorizes in a certain way for a graphical model. This property can be exploited to facilitate the analysis (see section 5.3).

5.2 Bayesian Inference

In this section, we provide a very brief, non-exhaustive introduction to the principles of Bayesian inference. A good overview of Bayesian statistics is Spiegelhalter and Rice (2009, online available at `http://www.scholarpedia.org/article/Bayesian_statistics`). There exist also numerous textbooks on the subject, e.g. Gelman et al. (2003).

Bayesian inference provides a principled way of handling epistemic uncertainties about the parameters of a model in terms of probabilities. Therefore, all information/belief we have about the states of nature/the physics of the problem are quantified in terms of a probability distribution on the parameters involved. Subsequently, this so-called "prior distribution" is updated given data, which results in the "posterior distribution". The updating process is done according to Bayes' rule,

$$\Pr(\boldsymbol{\Theta}|\boldsymbol{\mathcal{D}}) = \Pr(\boldsymbol{\Theta}) * \Pr(\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta}) / \Pr(\boldsymbol{\mathcal{D}}), \tag{5.3}$$

where $\boldsymbol{\Theta}$ is the set of parameters of the model, which are considered to be random variables, and $\boldsymbol{\mathcal{D}}$ denotes the available data. In eq. (5.3), $\Pr(\boldsymbol{\Theta})$ is the prior distribution of the parameters $\boldsymbol{\Theta}$ and $\Pr(\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta})$ denotes the likelihood function. The denominator $\Pr(\boldsymbol{\mathcal{D}})$ is the marginal distribution of the data and can be obtained from the numerator by integrating out the parameters. The result, $\Pr(\boldsymbol{\Theta}|\boldsymbol{\mathcal{D}})$, is the posterior distribution of the parameters $\boldsymbol{\Theta}$ given the data $\boldsymbol{\mathcal{D}}$.

The likelihood function is determined by the model under consideration. The prior distribution $\Pr(\boldsymbol{\theta})$ quantifies our prior degree of belief about the parameters. This can sometimes be difficult to achieve, and the choice of a particular distribution is not always easy to justify. Nevertheless, Bayesian inference provides a principled way to include prior information into an analysis by means of eq. (5.3).

In simple cases it may be possible to calculate the posterior distribution $\Pr(\boldsymbol{\theta}|Y)$ analytically. For more complicated models one needs to resort to approximate inference methods, e.g. Markov Chain Monte Carlo (MCMC) sampling of the posterior distribution, primarily due to analytical complications in computing the marginal likelihood. MCMC (e.g. Gilks et al., 1996) constructs a Markov chain such that the stationary/limiting distribution is the posterior. This is what we will use to obtain the parameters of our GMM.

5.3 Graphical models

As stated in the previous section, we estimate the posterior distribution of the parameters of our GMM by MCMC sampling. In particular, we use the program OpenBUGS (`http://www.openbugs.info/`) (Lunn et al., 2009), where BUGS stands for 'Bayesian inference using Gibbs sampling'.

Gibbs sampling (Geman and Geman, 1984) is an MCMC algorithm that exploits conditional independence assumptions between the quantities (parameters and observables) of a model. A convenient way to encode conditional (in)dependence assumptions are *directed acyclic graphs* (DAGs; e.g. Spiegelhalter, 1998), which are a special kind of graphical models. They are described below. For a more detailed introduction to graphical models, see e.g. Jordan (2004), Koller et al. (2007) or Koller and Friedman (2009).

In a graphical model, each quantity (observable, parameter, functions of both) corresponds to a node. Arcs between nodes denote direct dependence of two quantities. An example of a graphical model is shown in Figure 5.1 for a simple model of the kind

$$Y \sim \mathcal{N}(\mu_i = w_0 + w_1 X, \sigma), \tag{5.4}$$

where the parameters are $\boldsymbol{\Theta} = \{w_0, w_1, \sigma\}$ and the observables are $\boldsymbol{\mathcal{D}} = \{X_i, Y_i\}$, where $i = 1 \ldots N$ indexes individual datapoints and N is the number of data. The graphical model in Figure 5.1 is a DAG because all arcs are arrows (*directed*), and there is no direct path from node back to itself following the arrows (*acyclic*).

In a graphical model, there are different kinds of nodes for different quantities:

- Fixed nodes: These nodes are denoted by rectangles and represent quantities with set values (e.g. fixed parameters or observables).

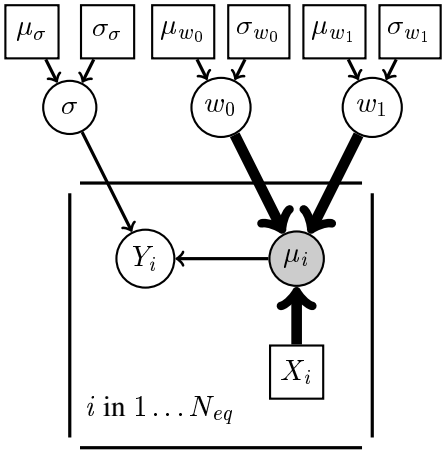

Figure 5.1: Graphical model for a simple linear model of the form $y(x, \boldsymbol{w}) = w_0 + w_1 x + \epsilon$

- Stochastic nodes: They are denoted by circles or ellipses and represent uncertain quantities which are associated with a probability distribution.
- Functional nodes: These are also denoted by circles or ellipses (here they are also shaded). They represent nodes that are functions of other quantities.

Functional dependences between parameters are represented by thick arrows, stochastic dependences by thin arrows. A node X from which an arrow points to a node Y is said to be the 'parent' of Y, while Y is called the 'child' of X.

The parameters $\boldsymbol{\Theta} = \{w_0, w_1, \sigma\}$ of our model are represented by stochastic nodes because they are uncertain parameters whose posterior distribution is to be estimated. The observed data points X_i are assumed to be error-free and is therefore represented by a rectangular node. The mean μ_i is a function of the parameters $\boldsymbol{\theta} = \{w_0, w_1, \sigma\}$ [cf. eq. (5.4)] and is therefore displayed as a shaded circular node. Y_i is represented by a stochastic node, because this is the data generating system that we assumed in eq. (5.4).

The parameters $\boldsymbol{\Theta} = \{w_0, w_1, \sigma\}$ need to be assigned a prior distribution. The parameters of these prior distributions are represented by the rectangular nodes in the top line of Figure 5.1. In principle, it would also be possible to place a prior distribution over these so-called 'hyperparameters', describing their uncertainty.

A graphical model carries the gist of the model (i.e. information about direct (in)dependencies between quantities), but conceals the nitty-gritty details (i.e. distributional or functional specifics). This is not so much a problem for the simple model of eq. (5.4), but it can be particularly useful for complex models, where a graphical model provides intuitive insight into the data generating process.

Specifying a model as a DAG also automatically encodes conditional independence assumptions and allows a factorization of the joint distribution, which makes the analysis of probabilistic models more convenient. It can be shown (Lauritzen et al., 1990) that for any particular

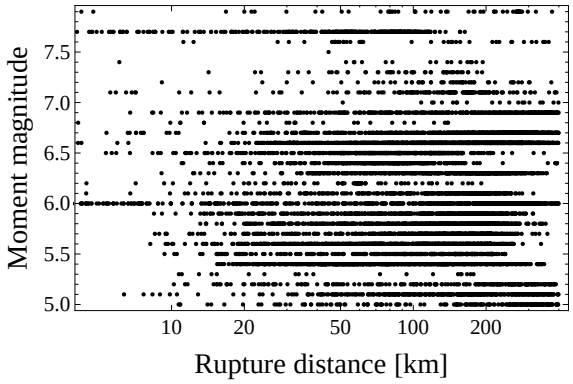

Figure 5.2: Magnitude-rupture distance distribution of the records that are used in this study.

$$\Pr(\boldsymbol{\Theta}|\boldsymbol{Y};\boldsymbol{X}) \propto \Pr(\boldsymbol{\Theta},\boldsymbol{Y};\boldsymbol{X}) = \prod_{V \epsilon \boldsymbol{Y} \cup \boldsymbol{\Theta}} \Pr(V|parents\,[V];\boldsymbol{X}), \tag{5.5}$$

where $parents[v]$ specifies the parent set of the nodes from which an arrow points to V (if the parent set is empty the conditional reduces to a marginal). Thus, it suffices to know the local distributions $\Pr(\boldsymbol{\Theta},\boldsymbol{Y};\boldsymbol{X})$ to specify the full joint distribution. Gibbs sampling makes efficient use of these properties.

5.4 Dataset

The dataset we use for constructing the global Bayesian ground motion model is the one compiled by Allen and Wald (2009). This dataset contains records from earthquakes in three different tectonic source types: shallow active tectonics, subduction zone and continental interiors. In this work, we use only earthquakes from shallow active tectonic regimes, in order to keep the model from being too complicated. However, in principle it is possible to extend the model to include also events from subduction zones and continental interiors.

The dataset of Allen and Wald (2009) contains 10,163 records from 238 earthquakes from shallow active tectonic regimes. The ground motion intensity parameters are PGA, PGV and the response spectrum at periods 0.3s, 1s and 3s. In this work, we use only PGA. For details on data compilation and processing, we refer to the original report of Allen and Wald (2009). In this work, we use only records up to a rupture distance of 400 km and above a magnitude of 5, which reduces the dataset to 9,831 records from 227 earthquakes. The magnitude-distance distribution of the used records is shown in Figure 5.2. A table with detailed information about the used earthquakes can be found in the electronic supplement.

The ground motion model we develop in this work, albeit a global one, takes into account regional differences (see section 5.5 for details). Therefore, we group the earthquakes into 10 regions. The earthquakes and regions are shown in Figure 5.3. The definition of the regions is based on geophysical considerations (e.g. stress drop variations (Allmann and Shearer, 2009)).

Figure 5.3: Location of earthquakes used in this study and definition of regions.

Table 5.1: Number of earthquakes and records per region.

Region ID	Region name	No. events	No. records
1	California	54	2433
2	Alaska	3	61
3	Middle America	3	18
4	Mid Atlantic Ridge	9	89
5	Europe	63	593
6	Iran	38	398
7	India	8	73
8	Japan and Taiwan	34	6034
9	Northeast China	1	6
10	New Zealand	14	126

However, we refrain from grouping the earthquakes into too many regions to avoid having a too small number of earthquakes in the individual regions. For example, there exist a number of ground motion models developed for different parts of Europe (e.g. Italy (or parts of Italy), Greece (e.g. Danciu and Tselentis (2007))). However, here we consider Europe and the Middle East as one region, similar to Ambraseys et al. (2005) or Akkar and Bommer (2010). The number of earthquakes and records per region is given in Table 5.1.

As one can see in Table 5.1 and Figure 5.3, there is one region with only one event (North Eastern China). Nevertheless, it is still possible to construct a model for this regions (i.e., calculate distinct parameters). In that case, data from other regions is more heavily called upon.

The predictor variables we consider are moment magnitude M_W, shortest distance to the rupture

Table 5.2: Numbers of different focal mechanism in the dataset.

Focal mechanism	No. of events	No. of records
Normal	35	329
Strike slip	98	3089
Reverse	62	4540
Unknown	32	1873

Table 5.3: Number of stations and records with different V_S30 values.

V_S30-range	No. stations	No. of records
V_S30 < 360 m/s	1175	3246
360 m/s ≤ V_S30 < 660 m/s	1711	5371
660 m/s ≤ V_S30	361	808
unknown	296	406

plane R_{RUP}, average shear wave velocity in the upper 30m, V_S30, and focal mechanism. The focal mechanism has three states, normal, strike slip, and reverse. The number of events and records for each of these is given in Table 5.2. The minimum and maximum V_S30 values are 210 m/s and 963.9 m/s, respectively. We group V_S30 into three site categories according to Table 5.3.

For some earthquakes, there is no information on the focal mechanism. Similarly, for some stations the value of V_S30 is missing. Nevertheless, the corresponding records can be retained in the analysis, as the uncertainty of the unknown values is taken care of during the analysis. For more details, see section 5.5.

5.5 Ground Motion Model Setup

In this section, we describe the GMM developed in this study. The GMM is outlined both as a graphical model as well as in equations. The target variable of our GMM is horizontal PGA. Since the dataset was originally compiled to reconstruct ground-shaking from recent-historical earthquakes (Allen et al., 2009) using USGS ShakeMap methodology (Wald et al., 1999), the two horizontal components are combined by taking the larger horizontal component.

Almost all published GMMs make the assumption that ground motion is log-normally distributed. We follow this assumption and introduce a new variable, Z, which is the natural logarithm of our target variable:

$$Z = \ln PGA. \tag{5.6}$$

We compose our model as a multilevel (or hierarchical) model (Gelman and Hill, 2007). The different levels allow to take into account grouped data. There are three levels in our model: the

station level (index s), corresponding to all records recorded at the same station; the *earthquake level* (index e), comprising all records from the same earthquake; and the *region level* (index r), where all records, stations and earthquakes of the different regions are grouped. The intersection of the earthquake and station level represents the record of the eth earthquake recorded at the sth station. This setup allows to consider the correlation of records from the same earthquake or same station, which is usually taken into account by using an appropriate regression technique, such as a one-step or two-step regression (Joyner and Boore, 1993, 1994) or a random effects algorithm (Abrahamson and Youngs, 1992). There exist also techniques to deal with inter-station variability (e.g. Chen and Tsai, 2002), though it is rarely considered in published GMMs. A multilevel model can be thought of as conceptually similar to a two-step regression with the ability of easily adding extra complexity.

Figure 5.4 depicts the graphical model corresponding to our GMM. Here the three levels are represented by the three plates (rectangular shapes). The outer plate (loop over r from 1 to N_{region}) corresponds to the different regions (cf. Figure 5.3). It embraces an individual GMM with individual parameter vectors $\boldsymbol{a}_r$, $\boldsymbol{b}_r$ and $\boldsymbol{c}_r$, which denote the parameters of the earthquake, record and station level, respectively. However, even though there are distinct parameters for each region, these are connected. Each regional parameter is sampled from a corresponding global normal distribution:

$$\theta_r \sim \mathcal{N}(\mu_\theta, \sigma_\theta), \tag{5.7}$$

where $\theta_r \in \{\boldsymbol{a}_r, \boldsymbol{b}_r, \boldsymbol{c}_r\}$. Thus, the parameters in each region are connected by global hyperparameters $\boldsymbol{\mu_a}$, $\boldsymbol{\mu_b}$, $\boldsymbol{\mu_c}$, $\boldsymbol{\sigma_a}$, $\boldsymbol{\sigma_b}$ and $\boldsymbol{\sigma_c}$. The width of the global distribution is a measure of the regional differences of the parameters. The global hyperparameters μ_θ and σ_θ allow the model partially pool the data from the different regions together such that data from all regions is used to estimate the coefficients in one individual region, but with different weight. For a very brief introduction to the ideas of multilevel modeling, see the Appendix. For more details, see Gelman and Hill (2007).

As mentioned above, there is an individual GMM for each region r inside the region plate of Figure 5.4. The graphical model corresponding to the individual GMMs can be thought of as a conceptual model of the data generating process. This is explained subsequently in more detail.

The concept of the model is as follows:

$$\begin{aligned} Z_{esr} &\sim \mathcal{N}(\mu_{Z,esr}, \phi_r) && (5.8)\\ \mu_{Z,esr} &= \mu_{\mathcal{R},kr} + \mathcal{E}_{er} + \mu_{\mathcal{S},sr} && (5.9)\\ \mathcal{E}_{er} &\sim \mathcal{N}(\mu_{\mathcal{E},er}, \tau_r) && (5.10) \end{aligned}$$

The central node for the individual GMMs of each region r in Figure 5.4 is Z_{esr}, which is the eth earthquake recorded at the sth station and corresponds to an observation of the the target variable. Z_{esr} is distributed according to a normal distribution with mean $\mu_{Z,esr}$ and standard deviation ϕ_r. The observation mean value $\mu_{Z,esr}$ is the sum of an event term, a station term and a record term, as shown in eq. (5.9). The event term $\mathcal{E}_{er}$ is common to all records from the same earthquake e and is itself distributed according to a normal distribution with mean $\mu_{\mathcal{E},er}$ and standard deviation τ_r. Coorespondingly, the station term is the same for each record from the same station. In principle,

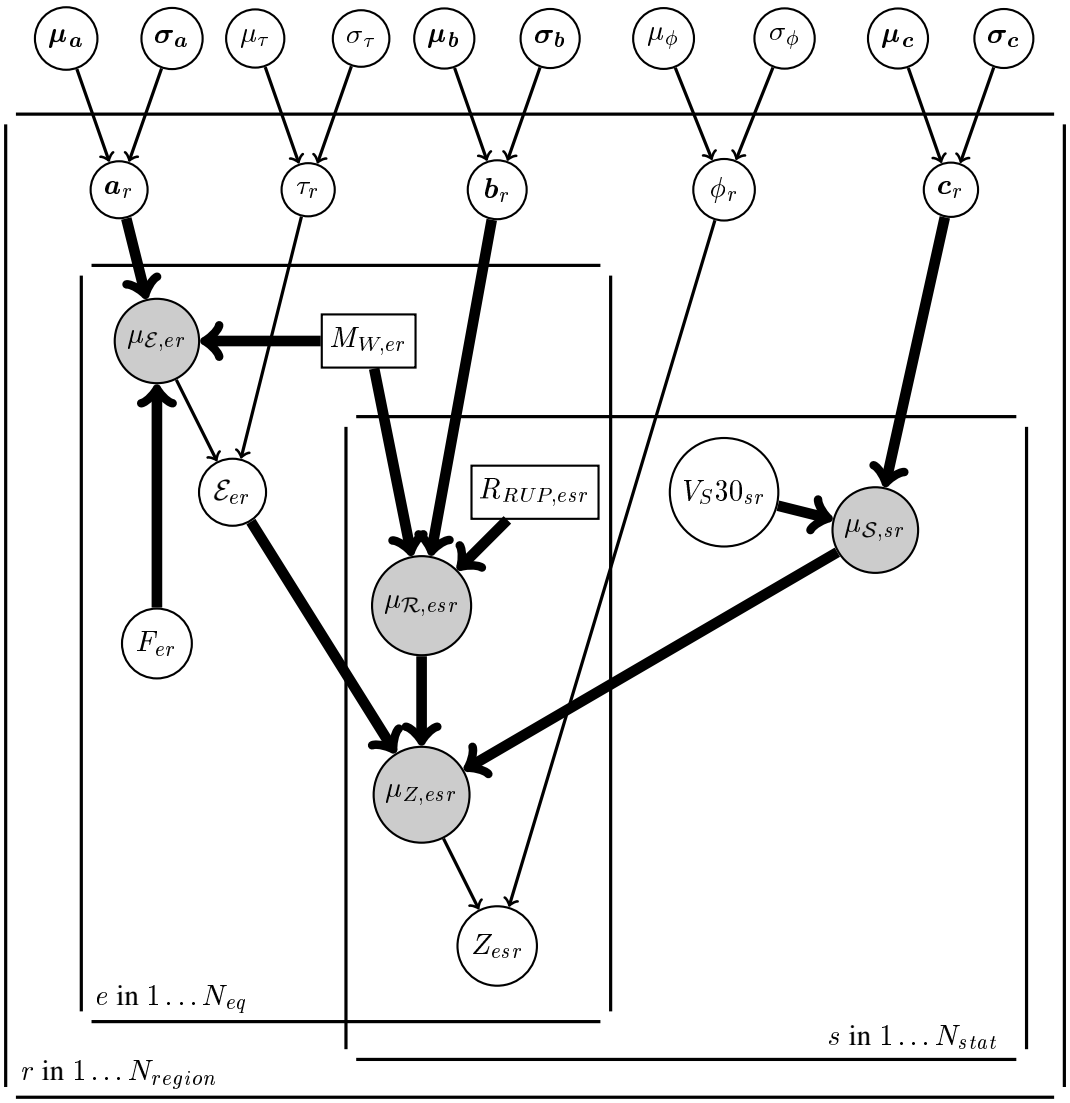

Figure 5.4: Graphical model for the global multilevel ground motion model.

it would be possible to assume that the station term is also a normally distributed random variable. However, estimation of the standard deviation of this distribution requires stations with multiple recordings, which are not abundant in our dataset. Hence, we assume that the station term is not a random variable, but a constant (strictly speaking, we assume that its standard deviation is zero). ϕ_r and τ_r are the within-event and between-event standard deviations, respectively.

The means of the event, record and station terms are functions of parameters and the predictor variables:

$$\mu_{rec,esr} = f(R_{RUP,esr}, M_{W,er}, \boldsymbol{b}_r) \tag{5.11}$$
$$\mu_{E,er} = g(M_{W,er}, F_{er}, \boldsymbol{a}_r) \tag{5.12}$$
$$\mu_{S,sr} = h(V_S30_{sr}, \boldsymbol{c}_r) \tag{5.13}$$

We have settled on the following functional forms for f, g and h. These are based on geophysical considerations and generalization error determined by 10-fold cross-validation (e.g. Kuehn et al., 2009; Hastie et al., 2001).

$$\begin{aligned} g(\boldsymbol{X}, \boldsymbol{a}_r) = {} & a_{0,r} + a_{1,r} * M_{W,ir} + a_{2,r} * (M_{W,ir} - 5.5) * H(M_{W,ir} - 5.5) + \\ & a_{3,r} * (M_{W,er} - 6.5) * H(M_{W,er} - 6.5) + \\ & a_{4,r} * F_{R,er} + a_{5,r} * F_{N,er} \end{aligned} \tag{5.14}$$
$$f(\boldsymbol{X}, \boldsymbol{b}_r) = (b_{0,r} + b_{1,r} * M_{W,er}) * \ln \sqrt{R^2_{Rup,esr} + (b_{2,r})^2} + b_{3,r} * R_{Rup,esr} \tag{5.15}$$
$$h(\boldsymbol{X}, \boldsymbol{c}_r) = c_{1,r} * S_{A,sr} + c_{2,r} * S_{S,sr} \tag{5.16}$$

F_R and F_N are dummy variables taking the value one for reverse and normal faulting, respectively, and zero otherwise. S_A and S_S are dummy variables equaling one for stiff soil and soft soil, respectively, and zero otherwise. $H(x)$ is the Heavyside-function which equals one for $x \geq 0$ and zero for $x < 0$.

We have settled for a trilinear magnitude scaling instead of a quadratic magnitude term since it results in a slightly lower generalization error and allows more control over the magnitude scaling. It also effectively decouples the large magnitude scaling from the small magnitude scaling. The term for the geometrical spreading, $(b_{0,r} + b_{1,r} * M_{W,er}) * \ln \sqrt{R^2_{Rup,esr} + b^2_{2,r}}$, is chosen because the results of the initial regression are more stable than when using a magnitude term inside the root, similar to Campbell and Bozorgnia (2008), and it gives a lower generalization error. We include an anelastic attenuation term, $b_{3,r} * R_{Rup,esr}$, since the maximum distance in the dataset is 400 km (cf. Figure 5.2).

As one can see, V_S30 and the focal mechanism F are displayed as stochastic nodes in Figure 5.4. This is due to the fact that some of these values are missing, i.e. there are some stations without an associated V_S30 value and some earthquakes with unknown focal mechanism. We treat these unknown values simply as parameters - they are assigned a prior distribution, which is updated by the likelihood resulting in a posterior distribution for each unknown V_S30 or F value. This is

Table 5.4: Prior distributions for the parameters.

Parameter	PGA
μ_{a_0}	$\mathcal{N}(-1.675, 16.75)$
μ_{a_1}	$\mathcal{N}(0.746, 7.46)$
μ_{a_2}	$\mathcal{N}(-0.402, 4.02)$
μ_{a_3}	$\mathcal{N}(-0.27, 2.7)$
μ_{a_4}	$\mathcal{N}(0.199, 1.99)$
μ_{a_5}	$\mathcal{N}(0., 1.)$
μ_{b_0}	$\mathcal{N}(-2.373, 23.73)$
μ_{b_1}	$\mathcal{N}(0.171, 1.71)$
μ_{b_2}	$\mathcal{N}(2.358, 23.58)$
μ_{b_3}	$\mathcal{N}(-0.005, 0.05)$
μ_{c_1}	$\mathcal{N}(0.1, 1.)$
μ_{c_2}	$\mathcal{N}(0.05, 0.5)$
μ_τ	$\mathcal{N}(0.3, 0.3)$
μ_ϕ	$\mathcal{N}(0.4, 0.4)$

a convenient side effect of the model, which thus provides a principled way to deal with missing data.

In Figure 5.4, prior distributions are needed for those parameters that are displayed as a stochastic node (i.e. as a circular node) that have no parents. In our case, these are the parameters of the global parameter distributions, i.e. the global hyperparameters $\boldsymbol{\mu_a}$, $\boldsymbol{\mu_b}$, $\boldsymbol{\mu_c}$, $\boldsymbol{\sigma_a}$, $\boldsymbol{\sigma_b}$ and $\boldsymbol{\sigma_c}$. The specification of the prior distributions used in the present work is explained below.

5.5.1 Prior Distributions

A key aspect of the philosophy behind Bayesian inference is the updating of the prior distribution, $\Pr(\boldsymbol{\Theta})$, with the likelihood of the data given the current model, $\Pr(\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta})$, which results in the posterior distribution of the parameters given the data, $\Pr(\boldsymbol{\Theta}|\boldsymbol{\mathcal{D}})$ (cf. section 5.2). All prior knowledge we have about the domain is quantified in the prior probability distribution $\Pr(\boldsymbol{\Theta})$. If we assume independence of the parameters, we need to specify a probability distribution for each coefficient. In our models, prior distributions are needed for the global hyperparameters, i.e. all μ_θ's and σ_θ's, and the missing values of V_S30 and F.

For the μ_θ's, we choose independent normal prior distributions. Hence, we need to specify a prior mean and standard deviation for each μ_θ. Often, there is some general knowledge about the scaling of ground motion intensity parameters with magnitude, distance and so on. However, it is difficult to quantify our prior information/belief into a probability distribution on, say, the parameter μ_{a_2}. On the other hand, it can be easier to specify a prior distribution on physical

parameters such as average stress drop or quality factor Q_0. Therefore, our strategy for specifying prior distributions is as follows:

1. Specify prior distributions on the stress drop, Q_0 and the slope of the geometric attenuation.
2. Generate a synthetic dataset from these parameters using stochastic simulations (Boore, 2003).
3. Regress the function of eqs. (5.14) to (5.15) on the synthetic dataset.
4. Take the estimated coefficients as prior means for the μ_θ's

For the stress drop, we assume a truncated normal distribution between 10 and 100 bar with mean 40 bar and standard deviation 20 bar, based on Allmann and Shearer (2008). The prior distribution for Q_0 is a truncated normal distribution between 10 and 1000, with mean 250 and standard deviation 50. We also set a prior distribution on the slope of the geometrical spreading, which is a truncated normal distribution between -1.5 and -0.8 with mean -1 and standard deviation 0.2. The above values are consistent when equivalent stochastic model parameters are determined for several GMMs from shallow active tectonic regions using the method of Scherbaum et al. (2006). The stochastic simulations are carried out with the program SMSIM (Boore, 2005).

The prior standard deviations of the μ_θ's are chosen so that the distributions are very wide. This allows the data to play a dominant role. Here, the standard deviations of the μ_θ's are chosen to yield a ration between standard deviation and mean of 10.

The above mentioned approach can be used to determine prior distributions for the parameters μ_{a_0}, μ_{a_1}, μ_{a_2}, μ_{a_3}, μ_{b_0}, μ_{b_1}, μ_{b_2} and μ_{b_3}, as well as μ_τ and μ_ϕ. For μ_{a_4} and μ_{a_5}, which describe the scaling of the ground motion intensity parameters with focal mechanism, we use Table III from Bommer et al. (2003). The site effects parameter μ_{c_1} and μ_{c_2} are assigned distributions based on guesses by the authors. These are loosely oriented on published GMMs as well as the work of Dobry et al. (2000). The means of the between-event and within standard deviations are also assigned based on published values. The standard deviations of all global distributions are assigned a broad, noninformative uniform distribution between 0 and 20. In Table 5.4 the prior distributions for the mean values are shown.

As described in section 5.5, we also treat missing V_S30 and F values as unknown parameters, which are assigned a prior distribution. V_S30 as well as F are categorical variables with three states, and we use a uniform prior over the three states for both of them.

5.6 Results

In the previous sections, we have specified the model, the prior distribution and the dataset. Using Bayes' rule, we can now estimate the posterior distribution of the parameters given the data, $\Pr(\boldsymbol{\Theta}|\boldsymbol{\mathcal{D}})$. As described before, the model is too complicated to estimate $\Pr(\boldsymbol{\Theta}|\boldsymbol{\mathcal{D}})$ analytically, so we resort to approximate inference, using MCMC sampling to obtain samples from the posterior distribution of each parameter. From the sequence of samples we can compute several summary statistics, such as means, standard deviations, quantiles and so on. The histogram of the sampled values serves as an approximation to the posterior probability density function.

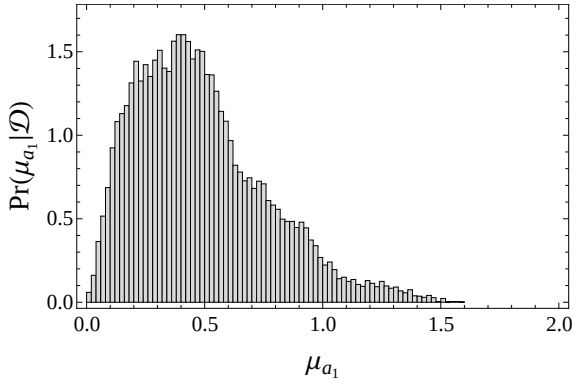

Figure 5.5: Histogram of sampled values (i.e. approximation of the posterior distribution) for the parameter μ_{a_1}.

In Figure 5.5, the approximate posterior distribution of the parameter μ_{a_1}, $\Pr(\mu_{a_1}|\mathcal{D})$, is displayed. The histogram shows clearly that μ_{a_1} can only be determined with considerable uncertainty. This uncertainty should not be neglected, thus highlighting the benefit of the Bayesian approach. Since there are too many parameters in the model to display posterior histograms for all of them, we provide these in the electronic supplement.

In Figure 5.6, we show a summary of the distributions of all parameters. Here, the 5%, 50% and 95% quantile of each parameter's posterior distribution is shown, i.e. a 90% confidence interval. The confidence intervals are shown for the means of the global distributions as well as for the parameters of each region. As one can see in Figure 5.6, the parameters that belong to the event term, $\boldsymbol{a}_r$ [eq. (5.14)], are in general associated with a higher uncertainty (illustrated by a larger confidence interval) than the record and site related parameters $\boldsymbol{b}_r$ and $\boldsymbol{c}_r$. The latter ones also display higher variability between the regions. This is discussed further in the next section.

In Figure 5.7, we show the residuals of the data with respect to the global parameters. Since the result of the analysis is a distribution for each parameter, not a single point, we calculate the residuals in the following way: For each data point, 100 parameter sets are sampled from the posterior distributions of the global parameters. With these sets of parameters, a PGA value is predicted according to eqs. (5.14) to (5.16), and the residual to the observed value is computed. Thus, there are 100 residuals per data point. The distribution of the residuals, both between- and within-event residuals, are then shown in Figure 5.7. As one can see, there is no obvious bias in the residuals.

Analogous to Figure 5.7, we show a residual distribution (again for 100 parameter sets for each data point) in Figure 5.8, but this time with respect to the regional parameters. Here, for each data point the parameters are sampled from the respective regional posterior distributions. Again, we see that there is no apparent bias in the residuals. The regional residual distributions are also narrower than the global ones in Figure 5.7, in particular for the between-event residuals.

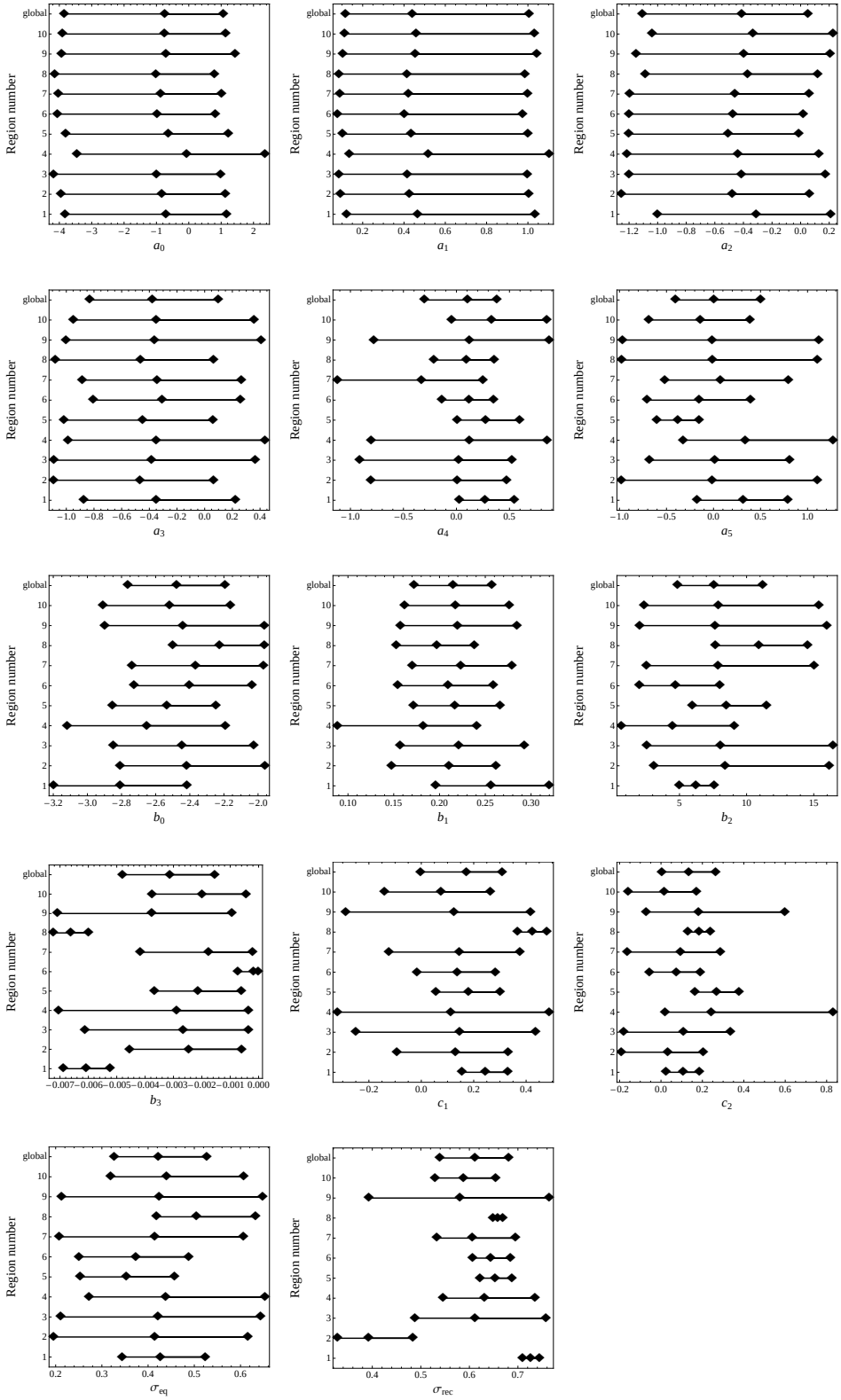

Figure 5.6: 90% confidence intervals for each parameter posterior distribution.

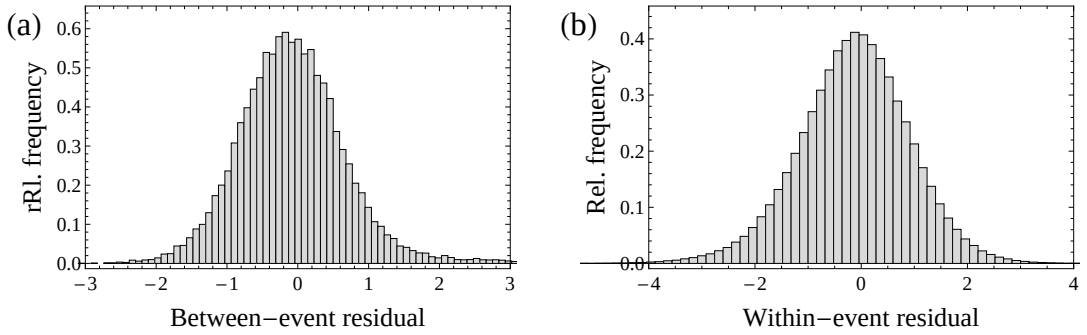

Figure 5.7: Residual distributions, calculated with the global parameters; (a) between-event residuals; (b) within-event residuals.

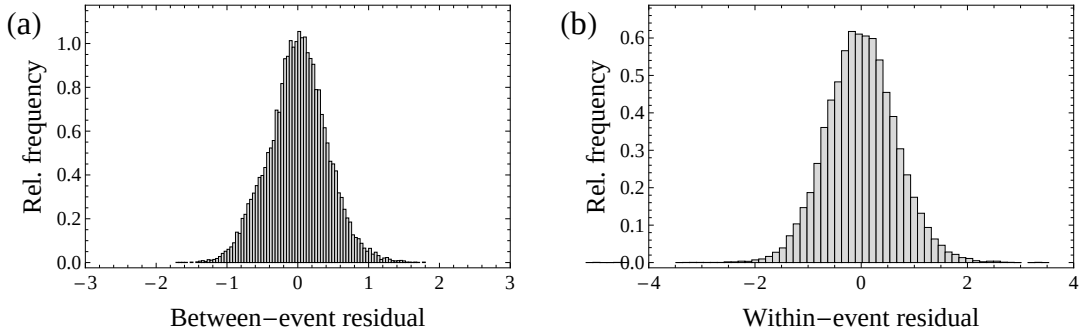

Figure 5.8: Residual distributions, calculated with the regional parameters; (a) between-event residuals; (b) within-event residuals.

The mean values of the between- and within-event residual distribution for each data point are plotted against magnitude and distance in Figure 5.9, respectively. Figure 5.9 (a) and (b) shows mean residuals with respect to the global parameters, while (c) and (d) are calculated using the residual parameters. There is no obvious trend with magnitude ((a) and (c)), while for larger distances the model underpredicts the data when using the global parameters (b). By contrast, when the residuals are calculated with the regional parameters (d), there is no trend with R_{RUP}.

In Figure 5.10, we compare the mean residuals for each of the individual regions using the regional and global parameters. Here, we see that the residuals with the regional parameters are lower over the individual regions. Thus, even though the residual distribution for the whole dataset is similar when using the global or regional parameters, within the individual regions there is a better fit with the regional parameter – not unexpected, since the regional parameters are primarily determined by data from the specific regions.

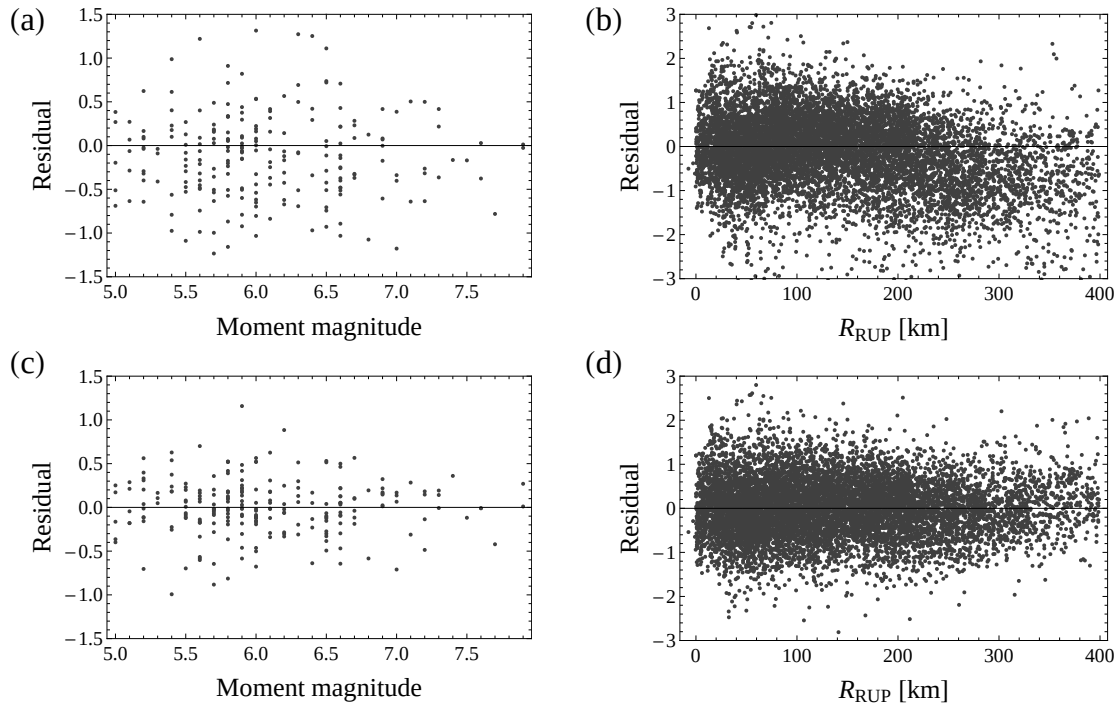

Figure 5.9: Mean residuals; (a) between-event residuals, calculated with global parameters; (b) within-event residuals, calculated with global parameters; (c) between-event residuals, calculated with regional parameters; (d) within-event residuals, calculated with regional parameters.

5.7 Discussion and Conclusions

Regional dependence of ground motion scaling is an open research question currently under debate (Douglas, 2009), whose answer has important implications for PSHA. Here, we look at this problem from a different angle. For this purpose, we have developed a ground motion model that can take regional differences in ground motion scaling into account. The coefficients in individual regions can be different, though we assume they are similar, which means that all data points from all regions are used to estimate the coefficients, though with different weights. The degree of the weights depends on the actual differences and the amount of data in the different regions [cf. the Appendix, eq. (5.18)].

The model/approach we have proposed is quite flexible. Here, we have allowed all coefficients to vary between regions. Furthermore, the variances of the global parameter distributions (e.g. σ_{a_0} and so on, cf. Figure 5.4) are determined by the data. However, these two points are not a must – we could have specified that some or all regions share coefficients, or assumed that the regional variability of, for example, the scaling of PGA with distance is fixed at some value. However, we did not feel comfortable deciding such an issue – we think that current knowledge does not support doing so. Nevertheless, the model is flexible enough to allow for that possibility.

The results of the analysis regarding regional dependence of PGA scaling with magnitude or

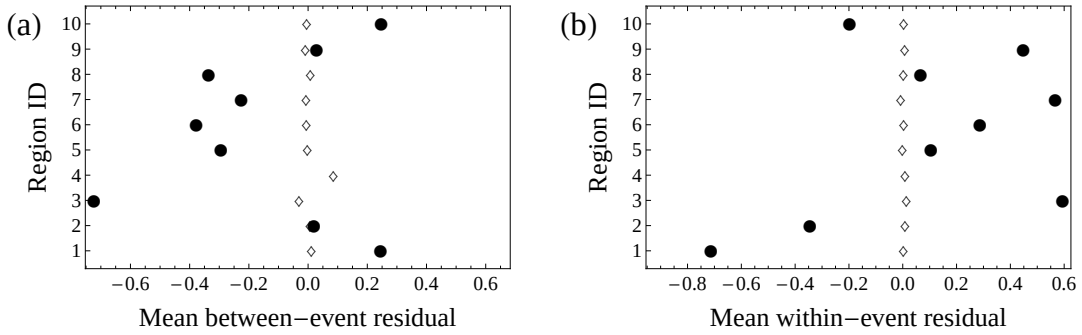

Figure 5.10: Mean residuals per region, calculated with global (•) and regional (◊) parameters: (a) between-event residuals; (b) within-event residuals.

distance are somewhat inconclusive, which is to some extent expected, as the model is not intended to answer such questions. In Figure 5.6, we have seen that there are differences in the posterior distributions of the parameters between different regions. These differences are most pronounced for the distance related parameters $\boldsymbol{b}$.

However, the interpretation of this feature as a physical characteristic might be subject to a caveat. The event related parameters are determined from 227 data points (the number of earthquakes in the dataset), while the record and site related parameters rely on 9,831 and 3,543 data points, respectively (cf. section 5.4). Therefore, the former can only be estimated with a higher degree of uncertainty. Furthermore, there are large differences in the number of records per region (cf. Table 5.1). To a lesser degree, this is also true for the number of earthquakes. This leads to the fact that parameters for regions with a large number of records (e.g. California, Europe, Japan and Taiwan; region ids 1, 5, 8, respectively) can be determined with higher precision than those for regions with a small number of records (e.g. Northeast China, region id 9). It also affects the way the partial data pooling is handled for the different regions (cf. the Appendix section 5.7). Thus, some of the apparent regional differences seen in Figure 5.6 might in part be ascribed to different sizes of the datasets.

To further investigate this issue, we look at the width of the global parameter distributions. Each parameter $\theta_r \in \{\boldsymbol{a}_r, \boldsymbol{b}_r, \boldsymbol{c}_r\}$ is sampled from a global distribution, $\theta_r \sim \mathcal{N}(\mu_\theta, \sigma_\theta)$. The parameters of this distribution, μ_θ and σ_θ, are itself associated with uncertainty, which is quantified in their posterior distribution. The width of the global parameter distribution, σ_θ, is an indicator of regional differences of the parameters. However, since the scale of the coefficients is quite different, one cannot compare the standard deviations directly. Therefore, we look at the coefficient of variation, which is defined as the standard deviation divided by the mean,

$$\mathrm{cov} = \frac{\sigma}{\mu}. \tag{5.17}$$

The higher the coefficient of variation, the larger the width of the distribution. For each sample from the posterior distribution of μ_θ and σ_θ, we can calculate cov_θ, and thus get a posterior distri-

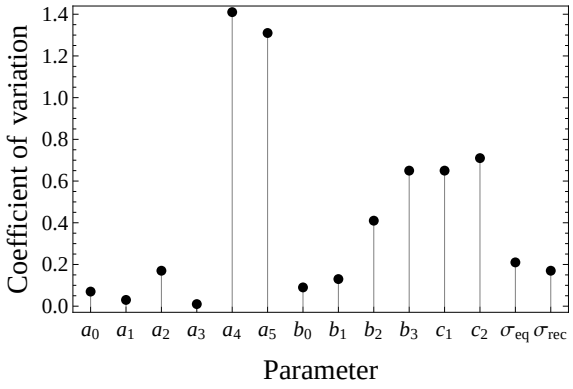

Figure 5.11: Modal values of the distribution of the coefficient of variation for the global distributions, calculated as cov = $\sigma_\theta / |\mu_\theta|$.

bution for the coefficient of variation. The modal values of these distributions are shown for each parameter in Figure 5.11.

For the distance related parameters, there are low cov-values for b_0 and b_1, while b_2 and b_3 are associated with larger values. Since these are based on a comparatively large amount of data (9,831 data points), this indicates that the latter parameters are indeed subject to stronger regional differences. This also makes physical sense: b_2 can be interpreted as a 'pseudo-depth', representing an average depth of the events in different regions. It can easily be imagined that this can be different in different regions. The parameter b_3, on the other hand, controls the anelastic attenuation with distance. It is especially important to model long-distance attenuation and is related to the quality factor Q_0. Again, it is physically plausible that this differs between regions. This can also explain the differences seen between Figures 5.9 (b) and 5.9 (d), which show the within-event residuals plotted against R_{RUP}. Using the global parameters, there is a trend visible for large distances, which diminishes when the regional parameters are used. This is an indication that for large distances, regional differences are relevant.

High cov-values are also taken for the parameters a_4 and a_5, which control the scaling for normal and reverse focal mechanism, respectively. Furthermore, also the site parameter c_1 and c_2 are associated with large cov-values. While this could be interpreted as regional differences, in this case there might be an alternative interpretation. In contrast to the other variables, there are several missing values for both the focal mechanism and V_S30. Even though we can deal with them (cf. section 5.5), this leads to an increase in uncertainty. Moreover, the distributions of both the focal mechanism and V_S30 are quite uneven over the different regions, which adds further "spread". Therefore, we think it is not feasible to interpret the parameters a_4, a_5, c_1 and c_3 in terms of regional differences.

In Figure 5.11, it can be seen that for the earthquake related parameters (except a_4 and a_5), the cov-values are low. This would indicate no or only small regional differences. However, as we have already elaborated, the number of events is small (227), and thus the parameters can only be estimated with considerable uncertainty (cf. Figure 5.6). Therefore, even if we find no or only small regional differences in the parameters a_0, a_1, a_2, and a_3, it is possible that the number of

data does not suffice to detect such differences.

The above discussion and findings illustrate an important point: It is not a GMM as a whole that is subject to regional differences or not, but certain aspects of it. Here, we find that the scaling of PGA with long distances is regional dependent, while magnitude scaling is probably not. By supplementing the NGA dataset with data from small to moderate earthquakes, Chiou et al. (2010) observe a difference in scaling with magnitude for small (¡6) magnitudes between central and southern California.

Another important aspect of our model is dealing with parameter uncertainty. The parameters of model are estimated by Bayesian regression, resulting in their posterior distribution given data, i.e. $\Pr(\boldsymbol{\Theta}|\boldsymbol{\mathcal{D}})$, which reflects the epistemic uncertainty of the parameters. This enables a full probabilistic treatment of the model in PSHA (if desired). The Bayesian approach to parameter estimation also makes it possible to incorporate prior knowledge in a principled way (cf. section 5.5.1). Most often, there is some prior domain knowledge that one can use to confine parameters. This is particular useful when the amount of data is sparse, and hence the parameters are not well constrained by data. However, the specification of the prior is not an easy task. In particular, assigning a probability, i.e. a number between 0 and 1, to a specific parameter value can be difficult.

In this work, we have used stochastic simulations (Boore, 2003) to specify the prior parameter distributions of our model (see section 5.5.1). We believe that this is a reasonable way to incorporate prior knowledge, since it is comparatively easy to specify a distribution over physical parameters such as stress drop or Q_0. This is also a good way to combine the output of simulations with regressions from empirical data.

The between-event and within-event variabilities ϕ and τ of our model are comparatively large. E.g. the total standard deviation for PGA using the global parameters is 0.742. For Europe, the total standard deviation is 0.743, compared to a value of 0.683 for the model of Akkar and Bommer (2010). Part of this difference can be attributed to how the two horizontal components are combined (Beyer and Bommer, 2006). Here, we use the larger horizontal component, while Akkar and Bommer (2010) used the geometric mean. However, the larger issue in this context is probably limitations in the dataset. There are measurement uncertainties associated with the predictor variables M_W, R_{RUP}, V_S30 and F, which are larger than e.g. in the NGA dataset. For example, the V_S30 values are calculated from topographic slopes, using the method of Wald and Allen (2007). These measurement uncertainties lead to an increased ground motion variability.

In principle, it would be possible to incorporate measurement uncertainties of the predictor variables into the model. However, we have refrained from doing so, since it would complicate an already complicated analysis, and require knowledge about the specific uncertainties of the data, which is not available in the dataset. To include measurement uncertainties, one would perform Monte Carlo simulations over the possible predictor values. In the graphical model (Figure 5.4), this would correspond to replacing the (deterministic) nodes for the predictor variables with a stochastic node, whose mean and width are the measurement and the associated uncertainty, respectively.

The above paragraph emphasizes an important aspect of graphical models – their extendability and flexibility. It is very easy to extend the model by adding nodes that might describe new aspects of the domain. Due to the factorization properties of graphical models, only local conditional probabilities need to be specified or changed. The graphical structure also provides an intuitive

insight into the model and serves as a proxy for the data generating process. In particular, it is easy to change the model by moving around nodes (easier than changing equations, which can get messy). E.g., moving one node out of the region plate means that this parameter is not regional dependent.

In the future, one can think of extending a graphical model such as depicted in Figure 5.4 to make the magnitudes distributed according to a Gutenberg-Richter distribution. The parameters a and b of the GR-relation can then be nodes in the model, capturing their uncertainty. That way, one can build a graphical model to calculate the hazard, with all associated uncertainties.

Data and Resources

The dataset used in this study is the one compiled by Allen and Wald (2009). Information about the records can be found in the electronic supplement. The MCMC sampling was done using the software OpenBUGS (`http://www.openbugs.info/w/`), version 3.06.

Electronic Supplement

The electronic supplement to this document contains information about the used earthquakes and records, as well sa plots of the histograms of the sampled posterior distributions of the parameters. It is available at `http://www.geo.uni-potsdam.de/mitarbeiter/Kuehn/kuehn-esupp-bayesreg.html`

Acknowledgements

Trevor Allen publishes with the permission of the Chief Executive Officer of Geoscience Australia.

Appendix: Multi-level Modeling

In this work, we assume that there is a GMM for each region r, which is sampled from a "global" distribution of GMMs (strictly speaking, there is a global distribution for each coefficient). This is an example of a multilevel model. Usually, if one has data from different groups, there are two natural ways to deal with it: *Complete pooling*, where all data is lumped together and inference is made on the whole dataset, and *no pooling*, where all groups are treated separately and inference is made individually for each group. The first approach can lead to problems if there are differences between the groups, while in the second approach some groups may not contain enough data to make reliable inferences.

By contrast, a multilevel model provides a compromise between these two extremes, *partial pooling*. Here, data from all groups is used for inference in the individual groups, but with different weight. As an example, imagine that we have measurements on a variable α from different groups/regions, and we want to estimate the means of this variable. It is straightforward to compute the overall mean, $\overline{\alpha}_{all}$ (the pooled estimate), as well as the unpooled estimate for each region

r, $\overline{\alpha}_r$. The multilevel (partial pooled) estimate is a weighted average of the pooled and unpooled estimates:

$$\hat{\alpha}_r^{multilevel} = \frac{\frac{n_r}{\sigma_\alpha^2}\overline{\alpha}_r + \frac{1}{\sigma_r^2}\overline{\alpha}_{all}}{\frac{n_r}{\sigma_\alpha^2} + \frac{1}{\sigma_r^2}}, \tag{5.18}$$

where n_r denotes the number of measurements in region r, σ_α^2 is the within-region variance, and σ_r^2 is the variance among the average values of the regions.

If the number of data in a region, n_r, is large, then the first term in eq. (refeq: ch4 multimean) carries more weight, and the partial pooled estimate will be close to the unpooled one. Conversely, if n_r is small, the the second term outweighs the first, and the region estimate is close to the global one.

In eq. (5.18), mean values of variables in different regions are estimated. It is straightforward to generalize it to cases where not only means, but also regression coefficients are estimated. For more details, see Gelman and Hill (2007).

6

A NAIVE BAYES CLASSIFIER FOR INTENSITIES USING PEAK GROUND VELOCITY AND ACCELERATION

Kuehn, N. M. and F. Scherbaum
Bulletin of the Seismological Society of America, in press

A naive Bayes classifier is determined to predict intensities from peak ground velocity and acceleration. It is trained on the same dataset that was used in the study of Faenza and Michelini (2010). The naive Bayes classifier directly estimates a discrete probability distribution for the ordinal intensities. Comparisons based on generalization error, estimated by cross-validation, show that the naive Bayes classifier performs better than traditionally employed regression models.

6.1 Introduction

Seismic intensities have recently gained much renewed attention. In particular, they can be used to make a first, quick assessment of potential damage after a large earthquake using the ShakeMap methodology (Wald et al., 1999a). Furthermore, they are often the only means to test the potential applicability of ground-motion models in regions where no or only very few instrumental data exists (Scherbaum et al., 2009; Delavaud et al., 2009). For both of these applications, a relation converting an instrumental ground motion parameter, such as peak ground acceleration (PGA) or peak ground velocity (PGV), into a seismic intensity I is needed.

There are several studies that provide such a relation (e.g. Atkinson and Sonley, 2000; Atkinson and Kaka, 2007; Chiaruttini and Siro, 1981; Kaka and Atkinson, 2004; Marin et al., 2004; Panza et al., 1997; Souriau, 2006; Tselentis and Danciu, 2008; Theodulidis and Papazachos, 1992; Wald

et al., 1999b). Most of these relations provide a simple regression equation of the form

$$I = a + b \log X, \tag{6.1}$$

where X is either PGA or PGV. A few studies also include additional predictor variables such as magnitude or distance (e.g. Atkinson and Kaka, 2007; Tselentis and Danciu, 2008).

It is worth remembering that macroseismic intensities are not quantitative, instrumentally measured parameters, but depend on human judgment and may also be mixed with information about building quality. Thus, they carry a large amount of uncertainty. Therefore, from a purely physical/seismological perspective, PGA, PGV or the response spectrum are better parameters to describe ground shaking, whereas seismic intensities have their main value in providing information about historical earthquakes (i.e. that occurred before the advent of seismometric data acquisition). On the other hand, as discussed in Wald et al (1999a), ShakeMaps of seismic intensities might aid interpretability in terms of rapid damage/loss estimation. Hence, converting instrumental ground motion intensity parameters into macroseismic intensities are beneficial for certain applications, and new relations, incorporating new data/methods, are useful.

Recently, Faenza and Michelini (2010) have presented new relations between PGA, PGV and intensities for Italy. They also provide an excellent, extensive review and discussion of different methodologies in this context, and provide an example of the application of the results in ShakeMap.

Relations like eq. (6.1) have been applied successfully in the context of ShakeMaps. However, in eq. (6.1), the target variable I is treated as continuous, while it is inherently a discrete, ordinal variable. Therefore, dependent on the value of X, a model like eq. (6.1) can predict an estimate of seismic intensity of, say, $I = 6.75$. Such a value, however, is not meaningful in terms of a discrete variable like seismic intensity, and it either has to be rounded to the next integer value, or an interpretation like "the intensity is 75% VII and 25% VI" has to be applied. Neither of these interpretations is completely satisfactory though. Furthermore, uncertainties in the intensity estimates are usually taken into account via a normal distribution, which is also continuous. For example, given a PGV-value of 5 cm/s, the model of Faenza and Michelini (2010) predicts an estimated intensity value of 6.75 with a standard deviation of 0.26, which is hard to interpret in terms of a discrete variable.

Here, we present a method that directly estimates intensities as a discrete variable: Naive Bayes classification (e.g. Mitchell, 1997, chapter 6). In this context, Bayes' rule is used to estimate the (discrete) conditional distribution $\Pr(I|X)$. We explain naive Bayes classification in section 6.2.

Our approach is similar to the one taken by Ebel and Wald (2003), who propose a Bayesian method to estimate the conditional distribution of an instrumental ground motion parameter X given modified Mercalli intensity, $\Pr(X|I)$. The method of Ebel and Wald requires estimates of $P(I|X)$, which they obtain using a strategy that is similar to a naive Bayes classifier, even though they do not call it that way.

The paper of Ebel and Wald (2003) has its main emphasis on the estimation of continuous instrumental ground motion parameters from seismic intensities, using Bayesian updating. Here, on the other hand, we focus on predicting seismic intensities from instrumental ground motion parameters by a naive Bayes classifier. In this context, it is important to point out that the name

Bayes classifier originates from the use of Bayes' rule in the analysis, but does not pertain to any statistical philosophy. On the other hand, the method of Ebel and Wald (2003) to estimate $\Pr(X|I)$ can be considered "Bayesian" in that it requires the specification of a prior probability $\Pr(X)$, which is updated by $\Pr(I|X)$ (see eq. (1) of Ebel and Wald (2003)). However, the parameters in Ebel and Wald (2003) are determined using maximum likelihood.

The goal of the present note is not to make a full-fledged investigation into the use of naive Bayes classification for predicting intensities, but rather to make simple comparisons between naive Bayes methods and regression models like eq. (6.1). Therefore, we use the dataset of Faenza and Michelini (2010), since it is freely available and is the basis of one of the most recent relations connecting seismic intensities and PGA/PGV.

In our paper, we often use the word "learn". We use this term in a rather broad sense, where "learning a model" means building the model and estimating its parameters. This is a reference to the machine learning community, where learning is used in this sense.

A few words on notation: Upper case symbols (e.g. X, Y, I) denote random variables. Vectors are represented in bold face (e.g. $\boldsymbol{X}$), and subscripts refer to each random variable or feature of a vector (e.g. X_i is a feature of $\boldsymbol{X}$). Lower case symbols denote values of a random variable (i.e., $X_i = X_{ij}$ refers to the random variable X_i taking on its jth possible value). We use the notation $\#(z)$ to denote the number of elements that satisfy property z (e.g. $\#(Y = y_i)$ is the number of instances in the data where Y takes on the value y_i).

6.2 Naive Bayes Classification

Suppose we have a variable Y that is categorical, i.e. has discrete instantiations, and that depends on some other variables $\boldsymbol{X} = \{X_1 \dots X_N\}$. A naive Bayes classifier predicts the the conditional distribution of Y given $\boldsymbol{X}$ using Bayes rule

$$\Pr(Y|\boldsymbol{X}) = \frac{\Pr(\boldsymbol{X}|Y)\Pr(Y)}{\Pr(X)} \tag{6.2}$$

First, we assume that all X_j are categorical as well. We will later generalize to continuous X_j. In classification, we want to predict the class value of Y given some information on the X_js, i.e. we are interested in estimating $\Pr(Y = y_i|X_1 \dots X_N)$. We can either work with the full distribution, or select the most probable value as the best estimate of Y.

With Bayes' rule, we can write $\Pr(Y = y_i|\boldsymbol{X})$ as

$$\Pr(Y = y_i|\boldsymbol{X} = \boldsymbol{x_k}) = \frac{\Pr(\boldsymbol{X} = \boldsymbol{x_k}|Y = y_i)\Pr(Y = y_i)}{\sum_j \Pr(\boldsymbol{X} = \boldsymbol{x_k}|Y = y_j)\Pr(Y = y_j)}, \tag{6.3}$$

where summation over j covers the whole event space, and such the y_j form a partition of the event space.

One can learn $\Pr(\boldsymbol{X}|Y)$ and $\Pr(Y)$ from the training data using the relative frequencies in the

dataset:

$$\Pr(Y = y_i) = \frac{\#(Y = y_i)}{\#(Y)}, \tag{6.4}$$

where $\#(Y)$ is the total number of data. However, it is impractical to learn $\Pr(\boldsymbol{X}|Y)$, since we need to estimate all the probabilities $\Pr(X_1 = x_{1k} \dots X_N = x_{Nk}|Y = y_i)$, which means that we need to have a sufficient number of data for each of these cases. One way to circumvent this problem is to make the assumption that the X_j are conditionally independent given Y, i.e.

$$\Pr(X_1 \dots X_N|Y) = \Pr(X_1|Y) \dots \Pr(X_N|Y) = \prod_{j=1}^{N} \Pr(X_j|Y). \tag{6.5}$$

This assumption greatly reduces the number of parameters to learn, since now we only have to estimate the probabilities $\Pr(X_j = x_{jk}|Y = y_i)$, which can be estimated from the relative frequencies in the data:

$$\Pr(X_j = x_{jk}|Y = y_i) = \frac{\#(X_j = x_{jk} \wedge Y = y_i)}{\#(Y = y_i)}, \tag{6.6}$$

where $\wedge$ means "logical and". The assumption of conditional independence of the predictor variables is often not very realistic from a physical perspective, but it works surprisingly well in many cases. In this context, it is important to keep in mind that the intention of the naive Bayes classifier is not to be a generative, physical model of the data generating process, but to predict Y given $\boldsymbol{X}$. Even if the assumption of conditional independence does not represent the physics of the problem, it is often sufficient for prediction. For example, in our problem PGV and PGA are not independent of each other, but for the purpose of predicting intensities it suffices to assume that they are. The name naive Bayes comes from this *naive* assumption.

As a side note, in a regression with more than one predictor variable, these are also usually assumed independent. Contrary to the naive Bayes classifier, however, where we state the assumption of independence explicitly, it is only implicit in regression.

In the case of continuous predictor variables X_j, Bayes' rule can still be used to estimate $\Pr(Y = y_i|\boldsymbol{X})$, but now the conditional distributions $\Pr(X_j|Y)$ cannot be specified using eq. (6.6). Instead, one can make the assumption that for each possible value y_i of Y, the continuous variable X_j follows a parametric distribution, e.g. a normal or log-normal distribution. Then, the learning task is to estimate the parameters of that distribution. To reduce the number of parameters, one can also assume that some parameters are the same for each y_i, e.g. that all standard deviations are the same. When it is not possible to make the assumption of a parametric distribution for continuous inputs, one can try to use a non-parametric method such as Kernel density estimation (e.g. Hastie et al., 2001, chapter 6.6), or discretize the continuous variables.

We stress again that the name Bayes classifier comes from the use of Bayes' theorem [eq. (6.3)]. It should not be confused with Bayesian inference. There is nothing inherently Bayesian about the method as it is outlined above, all parameters are estimated by maximum likelihood. It would nevertheless be possible to estimate the parameters using Bayesian inference.

6.3 Naive Bayes Classifiers Connecting PGA, PGV and seismic intensities

In this section, we learn naive Bayes classifiers relating seismic intensities to PGA and PGV. For our analysis, we use the same dataset as Faenza and Michelini (2010). They discuss in detail the data assemblage and the properties of the dataset. In total, the dataset comprises 266 intensities from 66 Italian earthquakes in 12 intensity classes from 2 to 8. Note that there are also intermediate intensities with values .5. Since these are not physically meaningful, we treat them such that they belong to both the class with full integer value above and below with a respective weight of 0.5. The intensity scale of the dataset is the Mercalli-Cancani-Sieberg scale (Sieberg, 1930). Hereafter, a seismic intensity is denoted by I. As predictor variables we consider PGA (in cm/s^2) and PGV (in cm/s). The minimum and maximum PGA values are 0.29 cm/s^2 and 569.55 cm/s^2, respectively. PGV ranges from 0.01 cm/s to 34.39 cm/s.

In our case, the predictor variables PGA and PGV are continuous variables. However, many studies have found one can assume a log-normal distribution of the ground motion intensity parameter for each intensity class, i.e. $\ln X$ is normally distributed given each intensity class:

$$\Pr(\ln X | I = k) = \mathcal{N}(\mu_{\ln X,k}, \sigma_{\ln X,k}), \tag{6.7}$$

where X is either PGV or PGA. Faenza and Michelini (2010) have shown that this assumption is well justified for their dataset. Hence, we need to estimate the mean values and standard deviations of $\ln(PGA)$ and $\ln(PGV)$ for each intensity class

$$\mu_{\ln X,k} = E\left[\ln(X) | I = k\right] \tag{6.8}$$

$$\sigma_{\ln X,k} = E\left[(\ln(X) - \mu_{\ln X,k})^2 | I = k\right], \tag{6.9}$$

where X is either PGA or PGV. We also need the prior distribution $\Pr(I)$, which can be calculated from the relative frequencies in the dataset using eq. (6.4). In addition to an individual standard deviation for each intensity class as estimated by eq. (6.9), we also estimate a common standard deviation for all intensity classes (but different for PGA and PGV). We do that since the number of intensities is small for some intensity classes, which does not allow a robust estimation of an individual standard deviation. The common standard deviation is determined by

$$\sigma = \frac{\sum_k \sum_{j=1}^{N_k} (\ln x_{jk} - \mu_{\ln X,k})^2}{\#(Y) - 1}, \tag{6.10}$$

where N_k denotes the number of data points with $Y = y_k$, i.e. $N_k = \#(Y = y_k)$.

The mean values, standard deviations and relative frequencies are given in Table 6.1. From the information presented in Table 6.1, it is straightforward to compute $\Pr(I|PGA, PGV)$. Therefore, first the probabilities $\Pr(PGA|I)$ and $\Pr(PGV|I)$ are calculated, which are then converted into $\Pr(I|PGA, PGV)$ using Bayes' rule and the information provided in Table 6.1 (cf. section 6.2 and eqs. (6.2) and (6.3)). It is equally simple to compute the conditional distribution of I given just one predictor variable, i.e. $\Pr(I|X)$, where X can be either PGA or PGV. It is also straightforward to calculate the distribution of $\ln PGA$ or $\ln PGV$ given I, which are just normal distributions with means and standard deviations as in Table 6.1.

Table 6.1: Means and standard deviations of $\ln(PGA)$ and $\ln(PGV)$, as well relative frequencies for each intensity class. The common standard deviation of $\ln(PGA)$ for all intensity classes is 0.89, the one of $\ln(PGV)$ is 0.87.

I_k	$\mu_{\ln PGA,k}$	$\sigma_{\ln PGA,k}$	$\mu_{\ln PGV,k}$	$\sigma_{\ln PGV,k}$	$\Pr(I_k)$
2	-0.04	0.62	-3.16	0.65	$\frac{10}{266}$
3	1.36	0.95	-1.92	1.00	$\frac{9.5}{266}$
4	2.23	1.18	-1.09	1.08	$\frac{38.5}{266}$
5	3.33	0.98	0.08	0.96	$\frac{99}{266}$
6	3.99	0.85	1.05	0.85	$\frac{76}{266}$
7	4.42	0.89	1.67	1.02	$\frac{24.5}{266}$
8	5.17	0.70	2.43	0.67	$\frac{8.5}{266}$

To assess the performance of different classifiers, we use leave-one-out cross-validation to estimate the generalization error (e.g. Hastie et al., 2001, chapter 7; Kuehn et al., 2009a) of the classifiers. The generalization error is a measure of the error that is made when predicting unseen data. Therefore, we drop one record from the dataset, learn the classifiers for the rest of dataset, classify I for the left out data point, and calculate the residual to the actual value. This is done for all data points, and the resulting residuals are averaged to give an estimate of the generalization error. We also perform leave-one-out cross-validation for regression, using the following functions:

$$I = a + b \ln PGA, \tag{6.11}$$

$$I = a + b \ln PGV, \tag{6.12}$$

$$I = a + b \ln PGA + c \ln PGV \tag{6.13}$$

The parameters of the regression models are learned by averaging the logarithmic ground motion values for each intensity class, following common practice (see e.g. Faenza and Michelini, 2010). Similar to the naive Bayes classifier, PGA and PGV are (implicitly) assumed to be independent in eq. (6.13).

The intensity predictions of the different models are made in the following way: For the naive Bayes classifiers, we simply take the most probable intensity value, i.e.

$$I \leftarrow \arg\max_{i_k} \frac{\Pr(I = i_k) \prod_i \Pr(X_i | I = i_k)}{\sum_j \Pr(I = i_j) \prod_i \Pr(X_i | I = i_j)}. \tag{6.14}$$

In eq. (6.14), the first term of the numerator is the "prior probability" of intensity class i_k, the second term represents $\Pr(\boldsymbol{X}|I = i_k)$ under the assumption of conditional independence (see eq. (6.5)), while the denominator is the marginal distribution of $\boldsymbol{X}$. Hence, the argument of the $\arg\max$ function is the full conditional distribution $\Pr(I|\boldsymbol{X})$, where $\boldsymbol{X}$ is either $\{PGA\}$, $\{PGA\}$ or $\{PGA, PGV\}$. The modal value of this distribution is the predicted intensity value. From the

Table 6.2: Generalization errors for different classifiers/regression models, calculated with the 0-1 loss $\mathcal{L}(I, \hat{I}(X))$. $\mathrm{NB}_{X,sd}$ is a naive Bayes classifier with the same standard deviation for all intensity classes.

Model	GE_{0-1}
$\mathrm{NB}_{PGA,sd}$	0.52
$\mathrm{NB}_{PGV,sd}$	0.54
$\mathrm{NB}_{PGA/PGV,sd}$	0.53
NB_{PGA}	0.52
NB_{PGV}	0.53
$\mathrm{NB}_{PGA/PGV}$	0.54
PGA regression	0.67
PGV regression	0.63
PGA & PGV regression	0.72

regression equations, intensity values are predicted by rounding the result of eqs. (6.11) to (6.13) to the nearest integer class value.

To calculate the residual between the predicted intensity $\hat{I}(X)$ and the test value I, we use the so-called 0-1 loss function $\mathcal{L}(I, \hat{I}(X))$. It takes the value 0 if the intensity is correctly classified and 1 if not, i.e.

$$\mathcal{L}(I, \hat{I}(X)) = \begin{cases} 0, & \text{if } I = \hat{I}(X) \\ 1, & \text{if } I \neq \hat{I}(X) \end{cases} . \tag{6.15}$$

The generalization errors for the different classifiers/regressions are displayed in Table 6.2. As one can see, the naive Bayes classifiers perform consistently better than regression, shown by a lower generalization error. Table 6.2 also shows that both PGV and PGA are predictors of similar quality when using the naive Bayes classifier, while for the regression PGV performs better (similar to Boatwright et al.(2001)). We also see that a combined predictor of PGV and PGA does not improve the predictive performance of either naive Bayes or regression. This can be interpreted that the information content of PGA and PGV with respect to seismic intensity is similar. In the case of the naive Bayes classifiers, there is also no difference between the ones with a common standard deviation and those with a different one for each intensity class. This indicates that both classifiers generalize equally well to unseen data. However, we believe that it is preferable to use a common standard deviation, since for some intensity classes there are only a few data points, which might render the estimation of the standard deviations unstable.

In Figure 6.1, we show predictions of I given PGV. Here, for each PGV value the full conditional distribution $\Pr(I|PGV)$ is shown, color-coded from light (low $\Pr(I|PGV)$) to dark (high $\Pr(I|PGV)$) colors. For each value of PGV on the x-axis, the corresponding color-coded values of $\Pr(I|PGV)$ along the vertical (I) axis sum up to unity. For comparison, we also plot the data points as well as the geometric means of the PGV values for each intensity class. The latter are

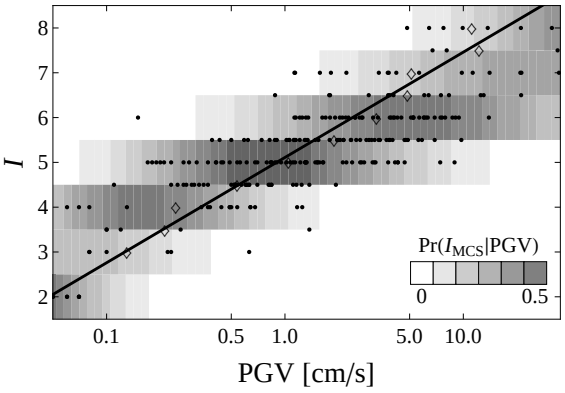

Figure 6.1: Comparison of regression model of Faenza and Michelini (2010) (straight line) and a naive Bayes classifier, between 0.05 cm/s and 35 cm/s. For the naive Bayes classifier, the full distribution is plotted, color coded by the value of $\Pr(I|PGV)$. The data points are plotted as black dots, the geometric means of PGV for each intensity as diamonds. The dataset contains seismic intensities with values .5, which are taken to belong to to the classes above and below with weight 0.5. These points are plotted with their original value.

used for the regression of the model of Faenza and Michelini (2010), which is also shown in Figure 6.1. As one can see, the most likely I predicted by the naive Bayes classifier (i.e. the intensity class with the highest $\Pr(I|PGV)$, corresponding to the darkest color for each intensity class) correlates reasonably well with the model of Faenza and Michelini (2010).

Figure 6.1 also shows the large scatter in the data, both for a given PGV value as well as for a given intensity value. This is very well represented by the naive Bayes classifier, which returns a relatively broad distribution $\Pr(I|PGV)$. The large scatter in intensity values for a particular PGV value indicates that it is important to treat I probabilistically, i.e. use the full distribution. This is facilitated by a naive Bayes classifier.

6.4 Discussion and Conclusions

We have presented naive Bayes classification to predict intensities from ground motion intensity parameters (PGA and PGV) as an alternative to traditional regression models. A naive Bayes classifier predicts the distribution of a discrete variable given some predictor variables using Bayes' rule, making the naive assumption that the predictor variables are conditionally independent given the target. This assumption greatly reduces the number of parameters to learn and is, albeit not realistic from a physical perspective, often sufficient for prediction. In our case, the assumption of conditional independence only applies if we use both PGA and PGV as predictors (and it applies to regression as well). From a purely physical perspective, this assumption is not justified, since there is correlation between PGA and PGV, but analysis of the generalization error (see Table 6.2) shows that the naive Bayes classifier nevertheless outperforms regression when it comes to prediction of seismic intensities from PGA and PGV. The naive Bayes classifier, however, is not suitable

as a physical model for the data generating process.

We have built a naive Bayes classifier to estimate $\Pr(I = k|PGV, PGA)$, making the assumption that the conditional distribution of PGA and PGV, respectively, given an intensity class, is log-normal. The analysis of the generalization error, estimated via leave-one-out cross-validation, shows that the naive Bayes performs better than regression when it comes to predicting unseen data. The generalization error also shows that PGV and PGA individually can both predict I similarly well, while the joint use of them does not lead to an improvement in prediction. Incidentally, we believe that this is due to the high correlation between PGA and PGV, which means that one can be used as a surrogate for the other.

A particular appealing feature of the naive Bayes classifier is that it provides a direct estimate of the discrete intensity distribution $\Pr(I|PGV, PGA)$. Compared to regression, there is no rounding or interpolation necessary, meaning that directly integer values are estimated. Since Bayes' rule [eq. (6.2)] is used for the estimation of $\Pr(I|PGV, PGA)$, an estimate of $\Pr(X|I)$ is required, where X is either PGA or PGV. Thus, the model can be just as easily used to predict the ground motion parameters given I.

We have learned two naive Bayes classifiers, one with a common standard deviation of the distribution of the ground motion intensity parameters over the different intensity classes, and one with different standard deviations. Even though both classifiers have a similar generalization error, we believe that it is better to use the former, since it provides a more stable estimate of the standard deviation. For some intensity classes, there are only 3 or 5 records, which makes it difficult to obtain a precise estimate of the standard deviation. Other possibilities exist, e.g. one could estimate a common standard deviation for adjacent intensity classes, which is done in Ebel and Wald (2003). Nevertheless, we think that the assumption of a common standard deviation over all intensity classes is reasonable.

In contrast to a regression model, which is unbounded, the naive Bayes classifier can only predict intensity values which occur in the underlying dataset. In principle, one could extrapolate a regression model to ground motion intensity values that lie beyond the extreme values found in the dataset to predict higher/lower intensity values (e.g. intensity values greater than 8 for the current dataset). This is not possible with a naive Bayes classifier. However, it is questionable if this is a disadvantage, since extrapolation of a model outside the parameter boundaries of its underlying dataset can be dangerous (see e.g. Bommer et al. (2007), for a discussion of extrapolating ground motion prediction equations).

The naive Bayes classifier that was learned in this study is trained on a dataset consisting of macroseismic intensities (of the Mercalli-Cancani-Sieberg scale) and PGA/PGV values from Italy, which is the same dataset used in Faenza and Michelini (2010) (see Data and Resources section). The reason why we have chosen it was because of the good documentation of the selection and preprocessing steps. We do not claim that this automatically justifies the application of their or our model in other regions which is an issue which requires careful consideration of a number of arguments (e.g. Cotton et al., 2006; Bommer et al., 2010). Certain ground shaking levels are bound to cause damage everywhere in the world, but since macroseismic intensity is a somewhat qualitative parameter that may include information on building quality, exact values/distributions might change from region to region. On the other hand, if the goal is to predict the most probable intensity value, the model may well be applicable in other regions, since this task is probably less

sensitive to regional influences. As said before, we leave this issue up to the user.

In this short note, we have considered PGA and PGV as predictor variables for I. Of course, the model can be extended to include also other variables such as magnitude or distance (e.g. Tselentis and Danciu, 2008). In that case, however, it is not as easy as in the case of PGA and PGV to assume a parametric distribution for each intensity class. Thus, either these variables need to be discretized, or some other method such as a Kernel density estimation needs to be employed. Such an analysis, however, is beyond the scope of this article.

Data and Resources

The dataset used in this study is the one compiled by Faenza and Michelini (2010), which is available in their electronic supplement under `http://www3.interscience.wiley.com/journal/123266793/suppinfo`.

Acknowledgments

We acknowledge that this paper was helped by the discussions in the Pegasos Refinement Project workshops. We thank the reviewers Fleur Strasser and Karen Assatourians and the editor Arthur McGarr for their comments which helped to clarify and improve the manuscript.

7

GENERAL CONCLUSIONS AND PERSPECTIVES

In this work, we have looked at uncertainty in GMMs. During that process, we also investigated some other questions that are of interest in the context of GMMs and PSHA, such as correlation between ground motion intensity parameters or regional differences in ground motion scaling. A considerable amount of uncertainty that is associated with GMMs pertains to their functional form $f(\boldsymbol{X})$ (cf. eq. (1.2)). Often, $f(\boldsymbol{X})$ is determined based on physical considerations and the analysis of residuals. In chapter 2, we have taken a new stance and based $f(\boldsymbol{X})$ on its predictive capability over the generating dataset. Therefore, we introduced the concept of generalization error and cross-validation. The idea here is that for PSHA, the primary goal of a GMM is not to to be a model of the physical processes in the ground motion domain, but to accurately predict future expected ground motions. Therefore, a GMM should be oriented along the lines of its predictive power. In this context, see also Breiman (2001a).

Based on the above considerations, a regression model is learned based on the NGA dataset which is optimized for its predictive capability. The model is rather complex (having many parameters), but is not overfit. We have calculated an equivalent stochastic model, which is physically interpretable (and also plausible, compared with already published models for western North America). Thus, the method we proposed is a convenient way to optimize a regression model for predictive power and checking that it makes physical sense.

A real physical interpretation is possible only for the equivalent stochastic model, since the parameters of the regression model are not tied to any physical meaning. However, partial dependence plots can reveal several characteristics of the model/data (cf. Figure 2.4, eq. (2.7) and Friedman (2001)). For example, in the partial dependence plot showing the scaling of PGA with distance there is a 'bump' visible in the range between R_{JB} = $50km$ and R_{JB} = $90km$. This 'bump' can be associated with the so-called Moho-bounce. This effect is not modeled in the NGA

models, but our analysis shows that it is supported by the data. Hence, the flexible, generalization-error optimized model shows which features are actually inherent in (or supported by) the data and can thus be helpful in choosing a functional form that models these features, thereby reducing uncertainty about $f(\boldsymbol{X})$.

On the other hand, the partial dependence plots also show data ranges which are problematic. In particular, this holds for the magnitude and the depth to the top of the rupture. Since there are typically fewer earthquakes than records in a strong motion dataset, these two variables are less well sampled than e.g. distance, and thus the scaling of ground motion with them is less clear defined by data. This manifests itself in 'rougher' partial dependence plots for the earthquake related variables. The overall scaling makes sense, but in some ranges is overly complicated. This reflects that the underlying dataset – at least for the earthquake related variables magnitude and depth to the top of the rupture – is not a representative sample of the true underlying distribution. Thus, for these variables the model can only provide guidance on the general form of ground motion scaling.

In chapter 2, we have optimized the model with respect to generalization error for the moment magnitude, Joyner-Boore distance, V_S30 and depth to the top of the rupture – the faulting style is included in the model, but is not adapted. One could also include other variables, such as directivity parameters or sediment depth, to investigate their (functional) relation to ground motion. One could also use other basis functions than polynomials, such as splines. Non-parametric regression methods such as MARS (multivariate adaptive regression splines, Friedman (1991)) or random forests (Breiman, 2001b) also may provide viable insights.

Along the same lines as the flexible regression model, we used Bayesian networks (BNs) to investigate what can be learned (purely) from data. The BN is a representation of the joint distribution of PGA and the (potential) predictors $\boldsymbol{X}$, $\Pr(PGA, \boldsymbol{X})$. Here, results are slightly complicated due to the need of discretizing the data, but again we find that there are problems in the underlying dataset – several data ranges are not well sampled. However, in ranges with good data coverage the BN gives reasonable results. One example of a possible not well represented data range is the scaling of PGA with very large magnitudes ($M_W > 7.5$. Here, a decrease of PGA with increasing magnitude is observed (so-called oversaturation). This is also seen in the NGA models (Abrahamson and Silva, 2008; Boore and Atkinson, 2008; Campbell and Bozorgnia, 2008), but it was decided not model this effect due to a lack of scientific consensus on that matter.

The BN is particularly well suited to investigate the set of possible predictor variables $\boldsymbol{X}$. Learning the structure of the BN means learning conditional independences, and hence with complete information, PGA is only influenced by the variables that are directly connected to it. We find that only five parameters – magnitude, Joyner-Boore distance, azimuth, style of faulting and depth to a shear wave horizon of 2.5 km/s, $Z_{2.5}$ – are directly connected to PGA in the final model, where the connection style of faulting ← PGA is included as expert knowledge. In particular, the effect of V_S30 on PGA is mediated by $Z_{2.5}$. Hence, once $Z_{2.5}$ is known, V_S30 does not provide any further information for the prediction of PGA. Thus, it is sufficient to know the five parameters for estimation of $\Pr(PGA|\boldsymbol{X})$. This may have implications for future studies, as these five variables should be included in new GMMs. However, this requires that information about them is present in strong motion datasets, which is currently not the case for all datasets.

Both the regression model as well as the BN provide valuable insight into the functional form

$f(\boldsymbol{X})$ and/or the set of predictors $\boldsymbol{X}$ of a GMM. However, both analyses also reveal that the underlying dataset has limitations, and that a completely data-driven (assumption free) approach is unwarranted. Furthermore, often there is good reason to make assumptions – they may be based on firm beliefs about the physics of the process. Assumptions can be made on specific forms of scaling of ground motions with the predictor variables (i.e. on the functional form of $f(\boldsymbol{X})$, but also on the parameters (e.g. to ensure monotonic scaling with magnitude).

Here is where the Bayesian approach to inference comes into play. Bayesian inference makes it possible to quantify prior beliefs, which are subsequently updated using data (cf. section 4.2). Thus, even though assumptions are made/quantified, these can be 'overridden' by data.

Two Bayesian GMMs were developed in this work. Prior beliefs were not quantified on the parameters of the GMMs themselves, but on physical parameters such as stress drop and Q_0. Using stochastic simulations, a synthetic dataset was created, on which prior distributions on the parameters are estimated by regression. We find that this is a good way to incorporate physical knowledge into a GMM.

In the Bayesian GMMs, we have specified the functional form based on physical considerations, since we have found that some constraints must be made regarding the scaling. However, as we have elaborated on chapter 2, the capability of a model to generalize should be an important factor. Therefore, we use generalization error based on cross-validation to choose between different forms of scaling – e.g., different forms of the magnitude scaling (quadratic, linear, tri-linear) can be found in the literature.

We find that the Bayesian approach works well for estimating the parameters of a GMM. The models we learned fit the data well, but also include (via the prior) our prior beliefs. Uncertainty on the parameters is quantified in terms of a probability distribution (the posterior $\Pr(\boldsymbol{\Theta}|\boldsymbol{\mathcal{D}})$). This is useful since it simplifies a full probabilistic treatment, which is desirable for PSHA.

In this work, we have developed two Bayesian GMMs. Both are similar in that they use the same basis model, but they investigate different aspects which are important in th context of GMMs and PSHA. The first model directly estimates the correlation between different ground motion intensity parameters during learning. This is important, since neglecting this correlation may distort the results of a PSHA. Here, we find no large difference in the posterior distribution of the parameters estimated with and without correlation, but this might not hold in general. Hence, learning a GMM and ignoring correlation between ground motion intensity parameters can lead to distortions.

We find that there is strong correlation between PGA, PGV and the response spectrum at three periods. The strength of the correlation depends on the period difference. There is both considerable between-event and within-event correlation, with the former being larger.

The second model takes into account potential regional differences in ground motion scaling between different regions. Here, the results are somewhat inconclusive, as we do not find the "smoking gun" evidencing the existence or absence of regional differences. There probably is no such thing as a smoking gun, as regional dependance of ground motion scaling is to all appearances not binary – it is not a GMM per se that is regional dependent, but aspects of a model. For example, we observe regional differences for the large distance scaling, while they appear to be negligible for the other aspects of the model. Similarly, Chiou et al. (2010) find that small magnitude scaling is different between southern and central California.

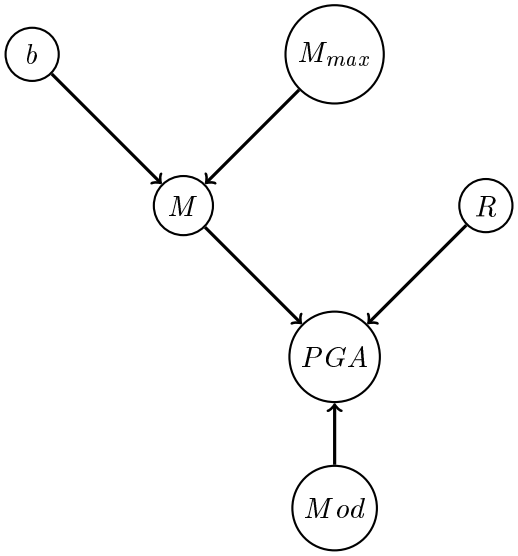

Figure 7.1: Concept of a graphical hazard model.

Even if we do not find evidence for regional differences for most aspects of ground motion scaling, that does not mean that we can rule them out, since the results are associated with large uncertainties. However, the Bayesian approach to inference is particularly apt to deal with that problem – we do not need to specify a model one way or the other, but can allow for the possibility of regional differences. Initially, all parameters can vary regionally, with large uncertainties associated with the degree of the regional differences. As new data becomes available, the uncertainties decrease, and regional variability in the model persists only for those parameters for which true regional differences exist.

Most of the models presented in this thesis are developed as probabilistic graphical models, with the BN being a special kind of these models. We find that graphical models are a convenient tool for reasoning under uncertainty. On the one hand, their graphical structure provides an easy and intuitive insight into the model, the data generating structure and the dependencies between variables. It also makes it easy to enhance them with additional complexities. Furthermore, due to their ability to encode conditional independencies between parameters they are an ideal instrument for setting up/analysing probabilistic models.

The flexibility of graphical models is demonstrated by the two Bayesian GMMs of chapters 4 and 5, which enhance the same base model to account for different complexities. The BN of chapter 3 is again a different kind of graphical model. Due to the factorization properties of graphical models, they are easy to enhance with additional nodes. For example, one can add nodes describing the magnitude distribution (e.g. a, b-value and maximum magnitude of a Gutenberg-Richter distribution, together with their associated uncertainty) without changing the other local probability distributions. That way, one can arrive at graphical model that describes all steps and uncertainties of a PSHA. This is conceptually shown in Figure 7.1. Here, the magnitude distribution depends on the values of the b- and M_{max} node, each of which can be assigned uncertainty

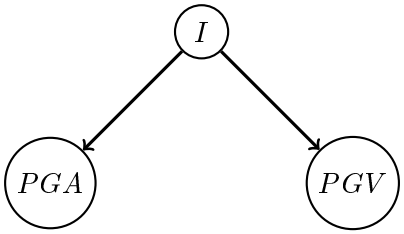

Figure 7.2: Graphical model of the naive Bayes classifier of chapter 6.

to. The distribution of the ground motion parameter Y depends on the Mod-node, which describes which GMM should be used to calculate $\Pr(Y|M, R)$. The conditional distribution of the ground motion parameter of interest can then be obtained using sampling or employing directly the fast inference algorithms of BNs. The conceptual model shown in Figure 7.1 can be enhanced in different ways to include more uncertainties/submodels/variables.

The model developed in chapter 6, a naive Bayes classifier connecting seismic intensities I, PGA and PGV, can also be represented by a graphical model, event though it is not explicitly mentioned in chapter 6. The graphical model which is equivalent to the naive Bayes classifier is shown in Figure 7.2. Here, the joint probability of I, PGA and PGV is encoded as

$$\Pr(I, PGA, PGV) = \Pr(I) \Pr(PGA|I) \Pr(PGV|I), \tag{7.1}$$

all of which are learned from data. From eq. (7.1) it is straightforward to compute $\Pr(I|PGA, PGV)$, which can be used in the generation of ShakeMaps (Wald et al., 1999a) or for the selection of GMMs in regions where data is sparse (Scherbaum et al., 2009; Delavaud et al., 2009). We find that the naive Bayes classifier performs better than commonly employed linear regression models, where performance is assessed via the 0-1 loss and generalization error (cf. section 6.3).

The naive Bayes classifier also has a conceptual advantage over regression models – it treats seismic intensity as a discrete rather than continuous variable. This leads to a better representation of uncertainty, quantified by $\Pr(I|PGA, PGV)$. Thus, the naive Bayes classifier lends itself to a convenient way of a fully probabilistic treatment of the conversion between instrumental ground motion parameters and seismic intensities.

Bibliography

Abrahamson, N. A. (2000). State of the practice of seismic hazard evaluation. *GeoEng 2000*. Melbourne, Australia, 2000.

Abrahamson, N. A. and J. J. Bommer (2005). Probability and Uncertainty in Seismic Hazard Analysis, *Earthquake Spectra* **21**, 603-607.

Abrahamson, N. A. and R. R. Youngs (1992). A Stable Algorithm for Regression Analyses Using the Random Effects Model, *Bull. Seism. Soc. Am.* **82**, 505-510.

Abrahamson, N. A. and W. J. Silva (1997). Empirical response spectral attenuation relations for shallow crustal earthquakes, *Seismol. Res. Lett* **68**, 9-23.

Abrahamson, N. A. and W. J. Silva (2008). Summary of the Abrahamson & Silva NGA Ground-Motion Relations, *Earthquake Spectra* **24**, 67-97.

Abrahamson, N. A. and R. R. Youngs (1992). A Stable Algorithm for Regression Analyses Using the Random Effects Model, *Bull. Seism. Soc. Am.* **82**, 505-510.

Ahmad, I., M. H. El Naggar and A. M. Khan (2008). Neural Network Based Attenuation of Strong Motion Peaks in Europe, *J. Earthq. Eng.* **12**, 663-680.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control AC***19**, 716-723.

Akkar, S. and J. J. Bommer (2010). Empirical Equations for the Prediction of PGA, PGV, and Spectral Accelerations in Europe, the Mediterranean Region, and the Middle East, *Seism. Res. Let.* **81**, 195-206.

Akkar, S. and J. J. Bommer (2007a). Empirical Prediction Equations for Peak Ground Velocity Derived from Strong-Motion Records from Europe and the Middle East, *Bull. Seism. Soc. Am.* **97**, 511-530.

Akkar, S. amd J. J. Bommer (2007b). Prediction of elastic displacement response spectra in Europe and the Middle East, *Earthq. Engng. Struct. Dyn.*, **36**, 1275-1301.

Al Atik, L., N. Abrahamson, F. Cotton, F. Scherbaum, J. Bommer, and N. Kuehn (2010). The Variability of Ground-Motion Prediction Models and its Components, *submitted to Seism.*

Res. Let.

Allen T. I., D. J. Wald, P. S. Earle, K. D. Marano, A. J. Hotovec, K. Lin and M. Hearne (2009). An Atlas of ShakeMaps and population exposure catalog for earthquake loss modeling. *Bull. Earthq. Eng.* **7**(3), 701-718, doi:710.1007/s10518-10009-19120-y

Allen, T. I. and D. J. Wald (2009). Evaluation of Ground-Motion Modeling Techniques for Use in Global ShakeMap - A Critique of Instrumental Ground-Motion Prediction Equations, Peak Ground Motion to Macroseismic Intensity Conversions, and Macroseismic Intensity Predictions in Different Tectonic Settings, U.S. Geological Survey Open-File Report 2009-1047, 114 p.

Allison, P. D. (2002). *Missing Data*, Sage Publications.

Allmann, B. P. and P. M. Shearer (2009). Global variations of stress drop for moderate to large earthquakes, *J. Geophys. Res.* **114**, B01310, doi:10.1029/2008JB005821.

Ambraseys, N. N., J. Douglas, S. K. Sarma and P. M. Smit (2005). Equations for the Estimation of Strong Ground Motions from Shallow Crustal Earthquakes Using Data from Europe and the Middle East: Horizontal Peak Ground Acceleration and Spectral Acceleration, *Bull. Earthq. Eng.* **3**, 1-53.

Anderson, J. G. (2000). Expected Shape of Regressions for Ground-Motion Parameters on Rock, *Bull. Seism. Soc. Am.* **90**, S43-S52.

Anderson, J. G. and J. N. Brune (1999). Probabilistic seismic hazard assessment without the ergodic assumption, *Seism. Res. Let.* **70**, 19-28.

Anderson, J. G. and Y. Lei (1994). Nonparametric Description of Peak Acceleration as a Function of Magnitude, Distance, and Site in Guerrero, Mexico, *Bull. Seism. Soc. Am.* **84**, 1003-1017.

Arroyo, D. and M. Ordaz (2010a). Multivariate Bayesian Regression Analysis Applied to Ground-Motion Prediction Equations, Part 1: Theory and Synthetic Example, *Bull. Seism. Soc. Am.* **100**, 1551-1567.

Arroyo, D. and M. Ordaz (2010b). Multivariate Bayesian Regression Analysis Applied to Ground-Motion Prediction Equations, Part 2: Numerical Example with Actual Data, *Bull. Seism. Soc. Am.* **100**, 1568-1577.

Atkinson, G. M. (2008). Ground-Motion Prediction Equations for Eastern North America from a Referenced Empirical Approach: Implications for Epistemic Uncertainty, *Bull. Seism. Soc. Am.* **98**, 1304-1318.

Atkinson, G. M. and W. J. Silva (2000). Stochastic Modeling of California Ground Motions, *Bull. Seism. Soc. Am.* **90**, 255-274.

Atkinson, G.M. and E. Sonley (2000). Empirical relationships between modified Mercalli intensity and response spectra, *Bull. Seism. Soc. Am.* **90**, 537-544.

Atkinson, G. M. and D. M. Boore (2006). Earthquake Ground-Motion Prediction Equations for Eastern North America, *Bull. Seism. Soc. Am.* **96**, 2181-2205.

Atkinson, G.M. and S.I. Kaka (2007). Relationships between Felt Intensity and Instrumental Ground Motion in the Central United States and California, *Bull. Seism. Soc. Am.* **97**, 497-510.

Baker, J. W. (2007). Correlation of ground motion intensity parameters used for predicting struc-

tural and geotechnical response, *Proceedings of the 10th International Conference on Applications of Statistics and Probability in Civil Engineering (ICASP10)*.

Baker, J. W. and C. A. Cornell (2006). Correlation of response spectral values for multicomponent ground motions, *Bull. Seism. Soc. Am.* **96**, 215-227.

Baker, J. W. and N. Jayaram (2008). Correlation of Spectral Acceleration Values from NGA Ground Motion Models, *Earthquake Spectra* **24**, 299-317.

Bayraktarli, Y. Y., U. Yazgan, A. Dazio and M. H. Faber (2006). Capabilities of the Bayesian Probabilistic Network Approach For Earthquake Risk Management, *First European Conference on Earthquake Engineering and Seismology*, Geneva 2006, Paper 1458.

Beyer, K. and J. J. Bommer (2006). Relationships between Median Values and between Aleatory Variabilities for Different Definitions of the Horizontal Component of Motion, *Bull. Seism. Soc. Am.* **96**, 1512-1522.

Blaser, L., M. Ohrnberger, C. Riggelsen and F. Scherbaum (2009). Bayesian Belief Network for Tsunami Warning Decision Support, I: Sossai, C., Chemello, G. (eds.) ECSQARU 2009, LNAI 5590, pp. 757-768.

Boatwright, J., K. Thywissen and L.C. Seekins (2001). Correlation of Ground Motion and Intensity for the 17 January 1994 Northridge, California, Earthquake, *Bull. Seism. Soc. Am.* **91**, 739-752.

Bommer, J. J., F. Scherbaum, H. Bungum, F. Cotton, F. Sabetta, and N. A. Abrahamson (2005). On the use of logic trees for ground-motion prediction equations in seismic-hazard analysis, *Bull. Seism. Soc. Am.* **95**, 377-389.

Bommer, J. J. and F. Scherbaum (2005). Capturing and limiting ground-motion uncertainty in seismic hazard assessment, in *Directions in Strong Motion Instrumentation*, In Gülkan, P. and J. G. Anderson (Eds.), NATO Science Series, Springer.

Bommer, J. J. and J. E. Alarcón (2006). The Prediction and Use of Peak Ground Velocity, *J. Earthq. Eng.* **10**, 1-32.

Bommer, J. J. and N. A. Abrahamson (2006). Why Do Modern Probabilistic Seismic-Hazard Analyses Often Lead to Increased Hazard Estimates?, *Bull. Seism. Soc. Am.* **96**, 1967-1977.

Bommer, J. J., P. J. Stafford, J. E. Alarcón, and S. Akkar (2007). The influence of magnitude range on empirical ground-motion prediction, *Bull. Seism. Soc. Am.* **97**, 2152-2170.

Bommer, J. J., J. Douglas, F. Scherbaum, F. Cotton, H. Bungum, and D. Fäh (2010). On the Selection of Ground-Motion Prediction Equations for Seismic Hazard Analysis, *Seism. Res. Let.*, in press.

Bommer, J. J., J. Douglas, and F. O. Strasser (2003). Style-of-Faulting in Ground Motion Prediction Equations, *Bull. Earthq. Eng.* **1** 171-203.

Bommer, J. J., J. Douglas, and F. O. Strasser (2003). Style-of-Faulting in Ground Motion Prediction Equations, *Bull. Earthq. Eng.* **1** 171-203.

Boore, D. M. (1983). Stochastic Simulation of High-Frequency Ground Motions Based on Seismological Models of the Radiated Spectra, *Bull. Seism. Soc. Am.* **73**, 1865-1894.

Boore, D. M. (2003). Simulation of Ground Motion Using the Stochastic Method, *Pure Appl. Geophys.* **160**, 635-676.

Boore, D.M. (2003). Simulation of Ground Motion Using the Stochastic Method, *Pure Appl. Geophys.* **160**, 635-676.

Boore, D.M. (2005). SMSIM - Fortran Programs for Simulating Ground Motions from Earthquakes: Version 2.3, A modified version of USGS OFR 00 - 509.

Boore, D. M. and G. M. Atkinson (2008). Ground-Motion Prediction Equations for the Average Horizontal Component of PGA, PGV, and 5%-Damped PSA at Spectral Periods between 0.01 s and 10.0 s, *Earthquake Spectra* **24**, 99-138.

Boore, D. M., J. Watson-Lamprey and N. A. Abrahamson (2006). Orientation-Independent Measures of Ground Motion, *Bull. Seism. Soc. America* **96**, 1502-1511.

Boore, D. M., W. B. Joyner and T. E. Fumal (1997), Equations for Estimating Horizontal Response Spectra and Peak Acceleration from Western North American Earthquakes: A Summary of Recent Work, *Seism. Res. Let.* **68**, 128-153.

Breiman, L. (2001a). Statistical Modeling: The Two Cultures, *Statist. Sci.* **3**, 199-231.

Breiman, L. (2001b). Random Forests, *Machine Learning* **45**, 5-32.

Burger, R. W., P. G. Somerville, J. S. Barker, R. B. Herrmann and D. V. Helmberger (1987). The Effect of Crustal Structure on Strong Ground Motion Attenuation Relations in Eastern North America, *Bull. Seism. Soc. Am.* **77**, 420-439.

Burnham, K. P. and D. R. Anderson (2002). *Model Selection and Multi-model Inference: A Practical Information-Theoretic Approach,* 2nd ed, Springer.

Campbell, K. (1991). An Empirical Analysis of Peak Horizontal Acceleration For The Loma Prieta, California, Earthquake of 18 October 1989, *Bull. Seism. Soc. Am.* **81**, 1838-1858.

Campbell, K. (2003). Prediction of Strong Ground Motion Using the Hybrid Empirical Method and Its Use in the Development of Ground-Motion (Attenuation) Relations in Eastern North America, *Bull. Seism. Soc. Am.* **93**, 1012-1033.

Campbell, K. (2004). Erratum to 'Prediction of Strong Ground Motion Using the Hybrid Empirical Method and Its Use in the Development of Ground-Motion (Attenuation) Relations in Eastern North America', *Bull. Seism. Soc. Am.* **93**, 1012-1033.

Campbell, K. and Y. Bozorgnia (2008). NGA Ground Motion Model for the Geometric Mean Horizontal Component of PGA, PGV, PGD and 5% Damped Linear Elastic Response Spectra for Periods Ranging from 0.01 to 10 s, *Earthquake Spectra* **24**, 139-171.

Castellaro, S., F. Mulargia and P. L. Rossi (2008). V_S30: Proxy for Seismic Amplification? *Seismol. Res. Letters* **79**, 540-543.

Castelo, R. and Kocka, T. (2003). On Inclusion-Driven Learning of Bayesian Networks, *J. of Machine Learning Research* **4**, 527-574.

Chen, Y. H. and C. C. P. Tsai (2002). A New Method for Estimation of the Attenuation Relationship with Variance Components, *Bull. Seism. Soc. Am.* **92**, 1984-1991.

Chen, Y. H. and C. C. P. Tsai (2002). A New Method for Estimation of the Attenuation Relationship with Variance Components, *Bull. Seism. Soc. Am.* **92**, 1984-1991.

Chiaruttini, C. and S. Siro (1981). The correlation of peak ground horizontal acceleration with magnitude, distance and seismic intensity for Friuli and Ancona, Italy, and the Alpide Belt, *Bull. Seism. Soc. Am.* **71**, 1993-2009.

Chiou, B. S.-J. and R. R. Youngs (2008). An NGA Model for the Average Horizontal Component of Peak Ground Motion and Response Spectra, *Earthquake Spectra*, **24** 173-215.

Chiou, B., R. Darragh, N. Gregor, and W. Silva (2008). NGA Project Strong-Motion Database, *Earthquake Spectra*, **24**, 23-44.

Chiou, B., R. Youngs, N. Abrahamson, and K. Addo (2010). Ground-Motion Attenuation Model for Small-To-Moderate Shallow Crustal Earthquakes in California and Its Implications on Regionalization of Ground-Motion Prediction Models, *Earthquake Spectra, in press*.

Choi, Y. and J. P. Stewart (2005). Nonlinear Site Amplification as Function of 30 m Shear Wave Velocity, *Earthquake Spectra* **21**, 1-30.

Coppersmith, K. J., and R. R. Youngs (1986). Capturing uncertainty in probabilistic seismic hazard assessments within intraplate tectonic environments, in *Proceedings of the Third U.S. National Conference on Earthquake Engineering*, Vol. 1, 301-312.

Cotton, F., F. Scherbaum, J.J. Bommer, and H. Bungum (2006). Criteria for selecting and adjusting ground-motion models for specific target regions: application to Central Europe and rock sites, *J. Seism.* **10**, 137-156.

Cotton, F., G. Pousse, F. Bonilla and F. Scherbaum (2008). On the Discrepancy of Recent European Ground-Motion Observations and Predictions from Empirical Models, *Bull. Seism. Soc. Am. 98*, 2244-2261.

Dalton, C. A., G. Ekström and A. M. Dziewonski (2008). The global attenuation structure of the upper mantle, *J. Geophys. Res.* **113**, B09303, doi:10.1029/2007JB005429.

Danciu, L. and G-A. Tselentis (2007). Engineering Ground-Motion Parameters Attenuation Relationships for Greece, *Bull. Seism. Soc. Am.* **97**, 162-183.

Delavaud, E., F. Scherbaum, N. Kuehn, and C. Riggelsen (2009). Information-Theoretic Ground-Motion Model Selection for Seismic Hazard Analysis: An Applicability Study Using Californian Data, *Bull. Seism. Soc. Am.* **99**, 3248-3263.

Dobry, R., R. D. Borcherdt, C. B. Crouse, I. M. Idriss, W. B. Joyner, G. R. Martin, M. S. Power, E. E. Rinne, and R. B. Seed (2000). New Site Coefficients and Site Classification System Used in Recent Building Seismic Code Provisions, *Earthquake Spectra* **16**, 41-67.

Douglas, J. (2003). Earthquake ground motion estimation using strong-motion records: a review of equations for the estimation of peak ground acceleration and response spectral ordinates, *Earth-Science Reviews* **61**, 43-104.

Douglas, J. (2004). Ground Motion Estimation Equations 1964-2003. *Research Report 04-001-SM,* Department of Civil and Environmental Engineering, Imperial College London.

Douglas, J. (2006). Errata of and additions to 'Ground Motion Estimation Equations 1964-2003', BRGM/RP-54603-FR.

Douglas, J. (2008). Further errata of and additions to Ground motion estimation equations 1964-2003, BRGM/RP-56187-FR.

Douglas, J. (2009). Investigating possible regional dependence in strong ground motions, 2nd Euro-Mediterranean Workshop on Accelerometric Data Exchange and Archiving, Ankara, 10-12 November 2009.

Douglas, J. (2010). Consistency of ground-motion predictions from the past four decades, *Bull.*

Earthq. Eng., online first.

Douglas, J. and Smit, P. M. (2001). How Accurate Can Strong Ground Motion Attenuation Relations Be?, *Bull. Seism. Soc. Am. 91*, 1917-1923.

Ebel, J.E. and D.J. Wald (2003). Bayesian Estimations of Peak Ground Acceleration and 5% Damped Spectral Acceleration from Modified Mercalli Intensity Data, *Earthquake Spectra* **19**, 511-529.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife, *Annals of Statistics* **7**, 1-26.

Faenza, L. and A. Michelini (2010). Regression analysis of MCS intensity and ground motion parameters in Italy and its application in ShakeMap, *Geophys. J. Int.* **180**, 1138-1152.

Fayyad, U. M. and K. B. Irani (1993). Multi interval discretization of continuous valued attributes for classification learning, in *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, 1022-1027.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine, *Annals of Statistics* **29**, 1189-1232.

Friedman, J. (1991). Multivariate adaptive regression splines (with discussion), *Annals of Statistics* **19**, 1-141.

Gallipoli, M. R. and M. Mucciarelli (2009). Comparison of Site Classification from V_{S30}, V_{S10}, and HVSR in Italy, *Bull. Seism. Soc. Am.* **99**, 340-351.

Gelman, A. and J. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 3rd ed., Cambridge University Press.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2003). *Bayesian Data Analysis*, 2nd edition, Chapman & Hall/CRC, Boca Raton, FL.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-741.

Getoor, L., N. Friedman, D. Koller, A. Pfeffer and B. Taskar (2007). Probabilistic Relational Models, in *Statistical Relational Learning*, L. Getoor and B. Taskar (eds)., MIT Press.

Gilks, W. R., S. Richardson and D. J. Spiegelhalter, Editors, Markov chain Monte Carlo in practice, Chapman & Hall/CRC, Boca Raton, FL (1996).

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Reading, MA.

Graizer, V. and E. Kalkan (2007). Ground Motion Attenuation Model for Peak Horizontal Acceleration from Shallow Crustal Earthquakes, *Earthquake Spectra 23*, 585-613.

Hanks, T. C. and R. K. McGuire (1981). The Character of High-Frequency Strong Ground Motion, *Bull. Seism. Soc. Am.* **71**, 2071-2095.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning,* Springer, New York.

Heckerman, D., C. Meek and D. Koller (2007). Probablistic Entitiy-relationship Models, PRMs, and Plate Models, in *Statistical Relational Learning*, L. Getoor and B. Taskar (eds)., MIT Press.

Idriss, I. M. (2008). An NGA Empirical Model for Estimating the Horizontal Spectral Values

Generated By Shallow Crustal Earthquakes, *Earthquake Spectra* **24**, 217-242.

Jensen, F. V. and T. D. Nielsen (2001). *Bayesian Networks and Decision Graphs,* 2nd edition, Springer, New York.

Jordan, M. I. (2004). Graphical models, *Statistical Science* **19**, 140-155.

Joyner, W. B. and D. M. Boore (1993). Methods for regression analysis of strong-motion data, *Bull. Seism. Soc. Am.* **83**, 469-487.

Joyner, W. B. and D. M. Boore (1994). Errata to 'Methods for regression analysis of strong-motion data', *Bull. Seism. Soc. Am.* **84**, 955-956.

Kaka, S.I. and G.M. Atkinson (2004). Relationships between Felt Intensity and Instrumental Ground Motion in the Central United States and California, *Bull. Seism. Soc. Am.* **94**, 1728-1736.

Kohonen, T. (2001). *Self-Organizing Maps*, Springer Verlag, New York.

Koller, D. and N. Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques,* Cambridge, MA: MIT Press.

Koller, D., Friedman, N., Getoor, L., and Taskar, B. (2007). Graphical models in a nutshell. In L. Getoor and B. Taskar (Eds.), *Introduction to statistical relational learning*. Cambridge, MA: MIT Press.

Kuehn, N. M., F. Scherbaum, and C. Riggelsen (2009a). Deriving Empirical Ground-Motion Models: Balancing Data Constraints and Physical Assumptions to Optimize Prediction Capability, *Bull. Seism. Soc. Am.* **99**, 2335-2347.

Kuehn, N. M., C. Riggelsen, and F. Scherbaum (2009b). Facilitating Probabilistic Seismic Hazard Analysis Using Bayesian Networks, *7th Workshop on Bayes Applications, UAI/ICML/COLT 2009*.

Kuehn, N. M., C. Riggelsen and F. Scherbaum (2009c). Sensitivity of Seismic Hazard Estimates to Earthquake and Site-Parameters Investigated by Bayesian Networks [abstract], *Seism. Res. Letters* **80**, 277.

Kulkarni, R. B., R. R. Youngs, and K. J. Coppersmith (1984). Assessment of confidence intervals for results of seismic hazard analysis, in *Proceedings of the Eighth World Conference on Earthquake Engineering*, San Francisco, Vol. 1, 263270.

Lauritzen S. L., A. P. Dawid, B. N. Larsen, and H. G. Leimer (1990). Independence properties of directed Markov fields. *Networks* **20**, 491-505.

Lee, W. H. K., T. C. Shin, K. W. Kuo, K. C. Chen and C. F. Wu (2001). CWB Free-Field Strong-Motion Data from the 21 September Chi-Chi, Taiwan, Earthquake, *Bull. Seism. Soc. Am.* **91**, 1370-1376.

Lunn, D., D. Spiegelhalter, A. Thomas, and N. Best (2009). The BUGS project: Evolution, critique and future directions, *Statistics in Medicine* **28**, 3049-3067.

Marin, S., J.-P. Avouac, M. Nicolas, and A. Schlupp (2004). A Probabilistic Approach to Seismic Hazard in Metropolitan France, *Bull. Seism. Soc. Am.* **94**, 2137-2163.

McGuire, R. K. (2008). Probabilistic seismic hazard analysis : Early history, *Earth. Eng & Struc. Dyn.* **37**, 329-338.

McGuire, R. K. and T. C. Hanks (1980). RMS Accelerations and Spectral Amplitudes of Strong

Ground Motion During the San Fernando, California Earthquake, *Bull. Seism. Soc. Am.* **70**, 1907-1919.

Mitchell, T. (1997). *Machine Learning*, McGraw-Hill.

Mosteller, F. and J. Tukey (1977). *Data Analysis and Regression,* Addison-Wesley, Redding, MA.

Musson, R. M. W. (2009). Ground motion and probabilistic hazard, *Bull. Earthq. Eng.* **7**, 575-589.

Newmark, N. M. and W. J. Hall (1982). *Earthquake Spectra and Design*, Earthquake Engineering Research Institute, El Cerrito, California.

Nikovski, D. (2000). Constructing Bayesian networks for medical diagnosis from incomplete and partially correct statistics. IEEE Transactions on Knowledge and Data Engineering **12**(4), 509-516.

Ordaz, M., A. Arciniega, and S. K. Singh (1994). Bayesian Attenuation Regressions: an Application to Mexico City, *Geophy. J. Int.* **117**, 335-344.

Panza, G.F., R. Cazzaro and F. Vaccari (1997). Correlation between macroseismic intensities and seismic ground motion parameters, *Ann. Geofis.* **40**, 1371-1382.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems,* Morgan Kaufman Publishers, San Mateo, California.

Pearl, J. and S. Russell (2000). *Bayesian Networks*. UCLA Cognitive Systems Laboratory, Technical Report (R-277).

Power, M., B. Chiou, N. Abrahamson, Y. Bozorgnia, T. Shantz, and C. Roblee (2008). An Overview of the NGA Project, *Earthquake Spectra*, **24** 3-21.

Reiter, L. (1990). *Earthquake Hazard Analysis: Issues and Insight*, Columbia University Press, New York.

Rhoades, D. A. (1997). Estimation of attenuation relations for strong motion data allowing for individual earthquake magnitude uncertainties, *Bull. Seism. Soc. Am.* **87**, 1674-1678.

Riggelsen, C. (2006). Learning bayesian networks from incomplete data: An efficient method for generating approximate predictive distributions. In: Jonker, W., Petković, M. (eds.) SDM 2006. LNCS, vol. 4165, Springer, Heidelberg (2006).

Riggelsen, C. (2008). Learning Bayesian Networks: A MAP Criterion for Joint Selection of Model Structure and Parameter, ICDM, pp.522-529, 2008 Eighth IEEE International Conference on Data Mining, 2008.

Rubin, D. (1976). Inference and Missing Data, *Biometrika* **63**, 591-592.

Sabetta, F., A. Lucantoni, H. Bungum and J.J. Bommer (2005). Sensitivity of PSHA results to ground motion prediction relations and logic-tree weights. *Soil Dyn. and Earthq. Eng.* **25**, 317-329.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis, *IEEE Trans. Comput.* **C-18**, 401409.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London.

Scherbaum, F., E. Delavaud, and C. Riggelsen (2009). Model Selection in Seismic Hazard Analysis: an Information-Theoretic Perspective, *Bull. Seism. Soc. Am.* **99**, 3234-3247.

Scherbaum, F., F. Cotton and H. Staedtke (2006). The Estimation of Minimum-Misfit Stochastic Models from Empirical Ground-Motion Prediction Equations, *Bull. Seismol. Soc. Am.* **96**,

427-445.

Scherbaum, F., F. Cotton, and P. Smit (2004a). On the use of response spectral-reference data for the selection of ground-motion models for seismic hazard analysis: the case of rock motion, *Bull. Seism. Soc. Am.* **94**, 2164-2185.

Scherbaum, F., J. Schmedes and F. Cotton (2004b). On the Conversion of Source-to-Site distance measures for extended earthquake source models, *Bull. Seismol. Soc. Am.* **94**, 1053-1069.

Scherbaum, F., N. M. Kuehn, M. Ohrnberger, and A. Koehler (2010). Exploring the Proximity of Ground-Motion Models Using High-Dimensional Visualization Techniques, *Earthquake Spectra*, *in press*.

Schwarz (1978). Estimating the Dimension of a Model, *Annals of Statistics* **6**, 461-464.

Sieberg, A. (1930). Geologie der Erdbeben, *Handbuch der Geophysik*, **2,4**, 552-555.

Somerville, P. G. and Yoshimura, J. (1990). The Influence of Critical Moho Reflections on Strong Ground Motions Recorded in San Francisco and Oakland during the 1989 Loma Prieta Earthquake, *Geophys. Res. Lett.* **17**, 1203-1206.

Souriau, A. (2006). Quantifying felt events: A joint analysis of intensities, accelerations and dominant frequencies, *J. Seism.* **10**, 23-38.

Spiegelhalter, D. J (1998). Bayesian graphical modelling: a case-study in monitoring health outcomes. *Applied Statistics* **47**, 115-133.

Spiegelhalter, D. and K. Rice (2009), Scholarpedia, 4(8):5230.

Spudich, P. and B. S.-J. Chiou (2008). Directivity in NGA Earthquake Ground Motions: Analysis using Isochrone Theory, *Earthquake Spectra* **24**, 279-298.

Stafford, P. J., F. O. Strasser and J. J. Bommer (2008). An Evaluation of the Applicability of the NGA models to Ground-Motion Prediction in the Euro-Mediterranean Region, *Bull. Earthq. Eng.* **6**, 149-177.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *J. Roy. Statist. Soc.* **B 36**, 111-147.

Strasser, F. O., N. A. Abrahamson and J. J. Bommer (2009). Sigma: Issues, Insights and Challenges, *Seismol. Res. Lett* **80**, 41-56.

Straub, D. (2005). Natural hazards risk assessment using Bayesian networks, in *Safety and Reliability of Engineering Systems and Structures (Proc. ICOSSAR 05, Rome)*, Augusti et al. (eds), Millpress.

Tavakoli, B. and S. Pezeshk (2007). A New Approach to Estimate a Mixed Model-Based Ground Motion Prediction Equation, *Earthquake Spectra 22*, 665-684.

Theodulidis, N.P., and B.C. Papazachos (1992). Dependence of strong ground motion on magnitude-distance, site geology and macroseismic intensity for shallow earthquakes in Greece: I, peak horizontal acceleration, velocity and displacement, *Soil Dyn. Earthq. Eng.* **11**, 387-402.

Toro. G. R. (2006). The Effects of Ground-Motion Uncertainty on Seismic Hazard Results: Examples and Approximate Results, *Annual Meeting of the Seismological Society of America*, San Francisco.

Tselentis, G-A. and L. Danciu (2008). Empirical Relationships between Modified Mercalli Intensity and Engineering Ground-Motion Parameters in Greece, *Bull. Seism. Soc. Am.* **98**,

1863-1875.

Wald D. J. and T. I. Allen (2007). Topographic slope as a proxy for seismic site conditions and amplification, *Bull. Seism. Soc. Am.* **97**(5),1379-1395.

Wald, D.J., V. Quitoriano, T.H. Heaton, H. Kanamori, C.W. Scrivner, and B.C. Worden (1999a). TriNet "ShakeMaps": Rapid generation of peak ground-motion and intensity maps for earthquakes in southern California: *Earthquake Spectra* **15**, 537-556.

Wald, D., V. Quitoriano, T.H. Heaton, and H. Kanamori (1999b). Relationships between peak ground acceleration, peak ground velocity and Modified Mercalli Intensity in California, *Earthquake Spectra* **15**, 557-564.

Walling, M. (2009). Non-ergodic probabilistic seismic hazard analysis and spatial simulation of variation in ground motion, *PhD Thesis*, University of California, Berkeley.

Wang, M. and T. Takada (2009). A Bayesian Framework for Prediction of Seismic Ground Motion, *Bull. Seism. Soc. Am.* **99**, 2348-2364.

Weisstein, E. W. (cited 01/2010). "Cholesky Decomposition." From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/CholeskyDecomposition.html

Zhao, J. X., J. Zhang, A. Asano, Y. Ohno, T. Oouchi, H. Ogawa, K. Irikura, H. K. Thio, P. G. Somerville, Y. Fukushima (2006). Attenuation Relations of Strong Ground Motion in Japan Using Site Classification Based on Predominant Period, *Bull. Seism. Soc. Am.* **96**, 898-913.

SUMMARY

The goal of a probabilistic seismic hazard analysis (PSHA) is the estimation of the expected rate of exceedance of a particular ground motion level A, $\nu(Y > A)$, where Y is a ground motion intensity parameter such as peak ground acceleration or the response spectrum. This is achieved by combining probabilistic assumptions about earthquake occurrences (spatial and temporal), magnitudes and generated ground motions, which yields a so-called hazard curve that relates a ground motion level with its expected rate of exceedance. Any such analysis is accompanied by large uncertainties, both aleatory and epistemic in nature. It has been noted (Toro, 2006) that uncertainties in the estimation of ground motions, i.e. so-called ground motion models (GMMs), have the largest effect on the results of PSHA, in particular for very low rates of exceedance, which are important for critical facilities such as nuclear power plants.

In this work, several investigations are carried out with respect to GMMs. These investigations address different issues in the development and estimation of GMMs.

First, a polynomial GMM is developed, where the order of the polynomials is determined based on generalization capability. The model is rather complex and non-physical, but is optimized for predictive power. Partial dependence plots reveal the characteristical scaling of the ground motion parameter with the predictor variables. They also show ranges which are not well sampled by data. The polynomial model is converted into a physical stochastic model to make it physically interpretable. The results are in good agreement with other published models.

Going a step further, a Bayesian network is learned on roughly the same dataset as the polynomial model. The Bayesian network provides a multivariate model for the ground motion domain, where direct (in)dependencies between quantities are estimated. It can both be used as a powerful tool for reasoning under uncertainty, as well as to investigate which parameters are directly relevant for predicting ground motions. In particular, V_S30 is not directly connected to PGA in the Bayesian network, but is mediated through the depth to the 2.5 km/s shear wave horizon. This is an indication that V_S30 might not be the parameter characterizing site effects with the highest predictive power for ground motions. The Bayesian network is in reasonable agreement with regression models in regions of good data coverage.

Two Bayesian GMMs are developed to investigate parameter uncertainty. Prior distributions of the coefficients are determined by setting prior distributions on physical parameters, simulating a synthetic ground motion dataset and determining the coefficients by regression on the synthetic dataset. This provides a way to combine both physical knowledge (simulations) and data-driven models in a principled way. The parameters related to source scaling (magnitude and style-of-faulting dependence) are generally associated with higher uncertainty than the ones related to path and site scaling. This is not surprising, since the latter ones are based on much more data – there are generally more records than earthquakes in a strong motion dataset.

The two Bayesian GMMs are similar in that they both comprise the same base model, which is expanded in different ways: The first model estimates directly the covariance (both between-event and within-event) between different ground motion intensity values during learning, which are usually considered independently. Strong correlations are found, the strength depending on the difference in period of the response spectrum. The between-event correlation is larger than the within-event correlation.

The second model takes into account possible regional differences in ground motion scaling. Therefore, the global dataset is split into 10 regions, each of which is represented by an individual GMM. These regional GMMs are assumed to be sampled from a global distribution of GMMs, which are parameterized by global hyperparameters. Data from all regions is used to estimate the parameters of the regional GMMs, though autochthonous data is assigned more weight. This procedure makes it possible to estimate an individual GMM in regions with a sparse amount of data. The analysis is not supposed to prove or disprove regional differences in ground motion scaling, but merely to present a methodology to take them into account in the light of large uncertainties. Consequently, results regarding regional differences are inconclusive. It seems to be that regional differences in magnitude scaling of PGA are small – the model is developed for magnitudes between 5 and 7.9 – while differences regarding anelastic attenuation appear to be genuine. However, results may be obfuscated due to large differences in the amount of data between different regions.

Finally, a naive Bayes classifier to predict seismic intensities from peak ground acceleration or peak ground velocity is learned, based on an Italian dataset. Such a model is useful for the rapid generation of so-called ShakeMaps or for the selection of GMMs in regions where instrumental ground motion data is sparse. The naive Bayes classifier performs better than commonly employed regression models, judged by generalization error under 0-1 loss. It also provides a better representation of the uncertainty by estimating a (discrete) conditional distribution of intensity given the instrumental ground motion variables, $\Pr(I|PGA, PGV)$, which makes a fully probabilistic treatment of the conversion possible.

ALLGEMEINVERSTÄNDLICHE ZUSAMMENFASSUNG

Das Ziel einer seismischen Gefährdungsanalyse besteht darin, die erwartete Überschreitensrate eines bestimmten Bodenbewegungswertes A, $\nu(Y > A)$, zu bestimmen. In diesem Zusammenhang bezeichnet Y einen Bodenbewegunsparameter von ingenieurseismologischem Interesse wie z.B. maximale Bodenbeschleunigung (peak ground acceleraiotn, PGA) oder das Antwortspektrum. $\nu(Y > A)$ wird bestimmt, indem probabilistische Modelle für das räumliche und zeitliche Auftreten von Erdbeben, ihre Magnituden sowie der auftretenden Bodenbwegungen kombiniert werden. Das Ergebnis ist eine sogenannte Hazardkurve, die einen bestimmten Bodenbewegungswert mit seiner erwarteten Überschreitensrate in Beziehung setzt. Seismische Gefährdungsanalysen sind von großen, sowohl aleatorischen als auch epistemischen Unsicherheiten betroffen. Insbesondere Unsicherheiten bei der Bestimmung der erwarteten Bodenbewegung haben den größten Einfluß auf das Ergebnis einer seismischen Gefährdungsanalyse.

Die erwartete Bodenbewegung wird durch sogenannte Bodenbewegunsmodelle (ground motion models, GMMs) quantifiziert, welche die Abhängigkeit der Bodenbewegung von relevanten Einflußgrößen modelliert – z.B. Magnitude oder Distanz. In dieser Arbeit werden Untersuchungen zu GMMs angestellt, die verschiedene Aspekte in Bezug auf die Enwtwicklung von GMMs und damit verbundener Unsicherheiten beleuchten.

Zunächst wird ein GMM entwickelt, das die Abhängigkeit der Bodenbewegung von den Einflußgrößen in Form von Polynomen modelliert. Die Ordnung der Polynome wird durch die Fähigkeit zu generalisieren bestimmt. Das sich daraus ergebene Modell ist verhältnismäßig komplex und nicht-physikalisch, jedoch optimiert in Bezug auf seine Vorhersagefähigkeit. Es kann Charakteristika der Abhängigkeit des Bodenbewegunsparameters von den Einflußgrößen aufzeigen, genau wie solche Datenbereiche, die nicht gut von Daten abgedeckt sind. Zur besseren Interpretierbarkeit wird aus dem Polynommodell ein äquivalentes physikalisches, stochastisches Modell invertiert. Dieses ist physikalisch plausibel.

Um die Frage zu untersuchen, welche Parameter wichtige Einflußgrößen für Bodenbewegung sind, wird ein multivariates Modell für die Verbundwahrscheinlichkeit aller möglicherweise rel-

evanten Parameter entwickelt, ein sogenanntes Bayesisches Netz. In diesem werden direkte (Un)Abhängigkeiten zwischen diesen bestimmt. Das Netz kann auch als Werkzeug für "Urteilen mit Unsicherheit" dienen. Im Bayesischen Netz ist V_S30 nicht direkt mit PGA verbunden, sondern ihr Einfluß wird vermittelt durch $Z_{2.5}$, die Tiefe zu einem Scherwellenhorizont von 2.5 km/s. Dies ist ein Hinweis darauf, daß V_S30 vermutlich nicht die Standortvariable mit der größten Vorhersagekraft für die Bestimmung von PGA ist.

Um Parameterunsicherheit von GMMs zu untersuchen, werden zwei Bayesische Regressionsmodelle entwickelt. Die A-Priori-Verteilungen für die Parameter werden aus einem synthetischen Datensatz bestimmt, der auf einem stochastischen Modell basiert. Auf diese Art und Weise können physikalisches Wissen (über Simulationen) und datengetriebene Modelle solide miteinander verknüpft werden. Die Parameter, die mit dem Erdbebenherd in Verbindung stehen (Skalierung mit Magnitude und Herdmechanismus), sind i.a. mit größerer Unsicherheit behaftet als jene, die die Distanz- und Standortabhängigkeit beschreiben. Dies ist nicht überraschend, da die letzteren von mehr Daten bestimmt werden – es gibt i.a. mehr Einzelaufzeichnungen als Erdbeben in einem Datensatz.

Die beiden Bayesischen GMMs sind ähnlich – sie basieren beide auf dem gleichen Grundmodell, das auf verschiedene Art erweitert wird: Im ersten Modell wird direkt die Kovarianz (Intra- und Inter-Event) zusammen mit den Parametern bestimmt. Starke Korrelationen zwischen den einzelnen Bodenbewegunsparametern sind erkennbar, wobei die Stärke abhhängig vom Unterschied in der Periode des Antwortspektrums ist. Die Inter-Event Korrelation ist größer als die Intra-Event Korrelation.

Das zweite Modell berücksichtigt regionale Unterschiede in der Skalierung von Bodenbewegung. Dafür wird der globale Datensatz in Regionen aufgeteilt, für die jeweils ein eigenes Modell bestimmt wird. Die Parameter der regionalen GMMs werden mithilfe von Daten aus allen Regionen bestimmt, wobei autochthone Daten mehr Gewicht bekommen. Auf diese Art und Weise ist es möglich, ein eigenes GMM auch für Regionen zu bestimmen, in denen nur wenige Daten vorliegen. Für die Magnitudenskalierung wird keine regionale Abhängigkeit gefunden, wohingegen die Skalierung mit großen Distanzen wohl tatsächlich regional unterschiedlich ist. Allerdings können die Ergebnisse durch große Unterschiede in der Anzahl von Daten zwischen verschiedenen Regionen verwischt werden.

Zu guter Letzt wird ein naiver Bayes-Klassifikator gelernt, der die Vorhersage von seismischen Intensitäten aus maximaler Bodenbeschleunigung und -geschwindigkeit ermöglicht. Solch ein Modell ist nützlich für die Erzeugung von sogenannten ShakeMaps, oder um GMMs in Regionen mit wenig instrumentellen Daten (wie z.B. Mitteleuropa) auszuwählen. Der naive Bayes-Klassifikator führt zu besseren Ergebnissen als normalerweise benutzte Regressionsmodelle, beurteilt über den Generalisierungsfehler unter 0-1 Verlust. Der naive Bayes-Klassifikator bildet auch besser die Unsicherheit dieser Konversion ab, indem direkt eine (diskrete) bedingte Wahrscheinlichkeit von Intensitäten, gegeben die instrumentellen Bodenbewegungsparameter, bestimmt wird. Dies macht eine voll probabilistische Behandlung der Umwandlung möglich.