Computers & Geosciences 69 (2014) 72–77

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Manifold aligned ground motion prediction equations for regional datasets

Nikolaos Gianniotis [a,*], Nicolas Kuehn [b], Frank Scherbaum [a]

[a] University of Potsdam, Institute of Earth and Environmental Science, Germany
[b] Pacific Earthquake Engineering Research Center, University of California, Berkeley, USA

ARTICLE INFO

Article history:
Received 31 December 2013
Received in revised form
16 April 2014
Accepted 28 April 2014
Available online 6 May 2014

Keywords:
Ground-motion-model
Bagging
Ensembles
Manifold
Regional-dependence

ABSTRACT

Inferring a ground-motion prediction equation (GMPE) for a region in which only a small number of seismic events has been observed is a challenging task. A response to this data scarcity is to utilise data from other regions in the hope that there exist common patterns in the generation of ground motion that can contribute to the development of a GMPE for the region in question. This is not an unreasonable course of action since we expect regional GMPEs to be related to each other. In this work we model this relatedness by assuming that the regional GMPEs occupy a common low-dimensional manifold in the space of all possible GMPEs. As a consequence, the GMPEs are fitted in a joint manner and not independent of each other, borrowing predictive strength from each other's regional datasets. Experimentation on a real dataset shows that the manifold assumption displays better predictive performance over fitting regional GMPEs independent of each other.

1. Introduction

The goal of probabilistic seismic hazard analysis (PSHA) is the estimation of the future expected distribution of ground motion at a particular site of interest. This requires a description of possible earthquake sources affecting the site as well as the characterisation of possible ground motions generated by these earthquakes. This second aspect is usually addressed within a PSHA by using ground-motion prediction equations (GMPEs) which give an estimate of the conditional distribution of a ground-motion parameter of interest given earthquake related and site related parameters such as magnitude and source-to-site-distance or rock type. The ground-motion parameter of interest is usually peak ground acceleration (PGA) or response spectral values.

Typically, GMPEs are estimated by regression on a strong-motion dataset. For a historical overview and a list of published GMPEs, see Douglas (2011) and references therein. Due to the sparsity of strong-motions, in particular large earthquakes are rare, datasets underlying GMPEs often cover large areas—see for example the pan-European GMPEs described in Douglas et al. (2014), or the NGA-West 2 GMPEs (Abrahamson et al., 2014; Boore et al., 2013; Campbell and Bozorgnia, 2014) which are based on a global dataset. In fact, the subject of regional dependence of GMPEs is still a matter of active debate (Douglas, 2007). However, recently four of the NGA-West 2 GMPEs (Abrahamson et al., 2014; Boore et al., 2013; Campbell and Bozorgnia, 2014; Chiou and Youngs, 2014) have included regional variations in their models in the form of a different distance attenuation for different regions. Basically, the models use a common basis estimated using all the data, and then deviate from this common basis using regional data.

In this paper we take a different approach to the estimation of regional GMPEs. We start out from a large, pan-European data set (Akkar et al., 2014) which we divide into regional subsets based on geoscientific considerations. Each subset is modelled by a regional GMPE and we make the assumption that the set of all regional GMPEs reside on a common low-dimensional manifold embedded in the space of all possible GMPEs. Roughly speaking, a manifold can be thought of as a generalisation of a surface to higher dimensions, i.e. a hypersurface. Thus, the regional GMPEs are forced to share the common structure imposed by the manifold, and this helps each GMPE borrow predictive strength from the others. As experimental support, we demonstrate the proposed approach on the RESORCE dataset (Akkar et al., 2014) and show that it performs well in terms of predictive accuracy.

2. Manifold aligned GMPEs

2.1. A general ground-motion prediction equation

We assume that we are dealing with a dataset of observed ground-motion data that originate from R regions indexed by

* Corresponding author.
E-mail addresses: ngianni@geo.uni-potsdam.de (N. Gianniotis), kuehn@berkeley.edu (N. Kuehn), fs@geo.uni-potsdam.de (F. Scherbaum).

http://dx.doi.org/10.1016/j.cageo.2014.04.014
0098-3004/

$r=1,\dots,R$. Each region holds $N^{(r)}$ pairs of inputs (covariates) $\boldsymbol{x}_n^{(r)}$ and outputs (responses) $y_n^{(r)}$. $\boldsymbol{x}_n^{(r)}$ is a vector $\boldsymbol{x}_n^{(r)}=[M_w, R_{hyp}, V_{S30}, \varphi]$ holding recordings of the magnitude M_w, hypocentral distance R_{hyp}, the average shear wave velocity in the upper 30 m V_{S30} and fault mechanism $\varphi \in \{0=\text{strike slip/normal}, 1=\text{reverse}\}$ of seismic records, and $y_n^{(r)}$ is the logarithmic PGA of the seismic events. Hence, the rth region is associated with a dataset $\mathcal{D}^{(r)}$ that comprises $N^{(r)}$ pairs of inputs–outputs, $\mathcal{D}^{(r)}=\{(\boldsymbol{x}_1^{(r)}, y_1^{(r)}),\dots,(\boldsymbol{x}_{N^{(r)}}^{(r)}, y_{N^{(r)}}^{(r)})\}$. The subscripts and superscripts n, r are dropped whenever we refer to logarithmic PGA y and measurements $\boldsymbol{x}$ as general quantities (i.e. not particular data items).

In order to model the functional dependency between measurements $\boldsymbol{x}$ and logarithmic PGA y, we adopt a function g with coefficients $\boldsymbol{c}$ which reads:

$$g(\boldsymbol{x};\boldsymbol{c}) = c_1 + c_2 M_w + c_3 M_w^2 + (c_4 + c_5 M_w)\log\left(\sqrt{R_{hyp}^2 + c_6^2}\right) + c_7 \ln V_{S30} + c_8\varphi. \quad (1)$$

Coefficients $\boldsymbol{c}$ reside in D-dimensional coefficient space denoted by $\mathcal{C}$. Here $D=8$ for the 8 coefficients in Eq. (1). The functional dependency in Eq. (1) is similar to the one used to develop the European GMPE in Akkar and Bommer (2010). We use this form because it is relatively simple but still able to capture the general characteristics of ground-motion scaling.

Typically, one assumes that the variability of the observed data is described by a Gaussian density with variance σ^2. This gives rise to the following negative log-likelihood for a dataset of N data items:

$$-\log \prod_{n=1}^{N} \mathcal{N}(y_n; g(\boldsymbol{x}_n;\boldsymbol{c}), \sigma^2) = \frac{1}{\sigma^2}\sum_{n=1}^{N}(y_n - g(\boldsymbol{x}_n;\boldsymbol{c}))^2 + \text{const}. \quad (2)$$

Fitting a GMPE to a set of observed data involves minimising the objective in Eq. (2) with respect to the free coefficients $\boldsymbol{c}$ and σ^2. In this work, we assume that all regions have about the same variance σ^2, and henceforth we discard it as a multiplicative constant. Minimising Eq. (2) yields the maximum likelihood estimate $\boldsymbol{c}_{ML}$. The prediction of the fitted GMPE on a unseen test input $\boldsymbol{x}^*$ is simply $g(\boldsymbol{x}^*; \boldsymbol{c}_{ML})$.

When R regions are fitted independently, we assign one GMPE per region with its own coefficients $\boldsymbol{c}^{(r)}$. Hence, R independent objectives of the type in Eq. (2) are minimised. The R independent objectives can be simply summarised as

$$\sum_{r=1}^{R}\sum_{n=1}^{N^{(r)}}(y_n^{(r)} - g(\boldsymbol{x}_n^{(r)};\boldsymbol{c}^{(r)}))^2. \quad (3)$$

This situation is illustrated in Fig. 1 which shows the coefficient space $\mathcal{C}$. Every point in $\mathcal{C}$ corresponds to a coefficient vector $\boldsymbol{c}$. Each dataset $\mathcal{D}^{(r)}$ gives rise to likelihood contours in $\mathcal{C}$ that show which coefficients are likely under Eq. (3). The likelihood contours are depicted with brighter and darker shades indicating low and high likelihood respectively. Since the R GMPEs ($R=4$ in the stylised example of Fig. 1) are treated independent of each other, coefficients $\boldsymbol{c}^{(r)}$ converge on the dark areas of their respective likelihood contours when optimising Eq. (3), i.e. they converge on their maximum likelihood estimates $\boldsymbol{c}_{ML}^{(r)}$.

2.2. Model formulation

Model $g(\boldsymbol{x};\boldsymbol{c})$ is governed by its coefficient vector $\boldsymbol{c}$ that belongs to coefficients space $\mathcal{C}$. Thus, we can say that each coefficient vector $\boldsymbol{c}$ addresses a model $g(\boldsymbol{x};\boldsymbol{c})$, and space $\mathcal{C}$ addresses the entirety of possible models $g(\boldsymbol{x};\boldsymbol{c})$. When we fit a region independent of the others, we do not impose any constraints on $\boldsymbol{c}$. Instead, we allow it to roam freely in $\mathcal{C}$, and independent of the other regions, until it converges to $\boldsymbol{c}_{ML}$ that maximises the region's likelihood.

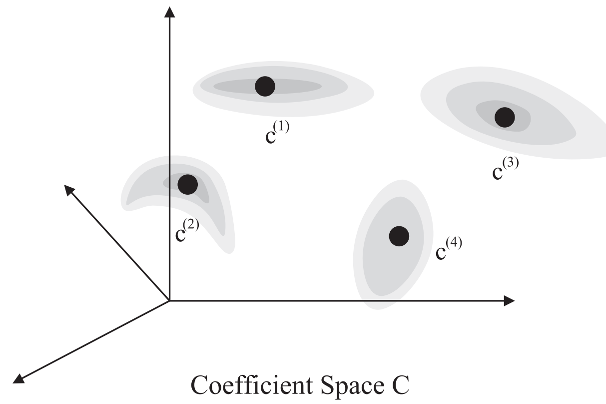

Fig. 1. Stylised depiction of coefficient space $\mathcal{C}$. Each point in $\mathcal{C}$ corresponds to a coefficient vector $\boldsymbol{c}$. Depicted are the maximum likelihood estimates $\boldsymbol{c}_{ML}$ (only for 4 regions) as points in $\mathcal{C}$ obtained by optimising the objective in Eq. (3). The contours show how likelihood for each regional model is distributed in $\mathcal{C}$.

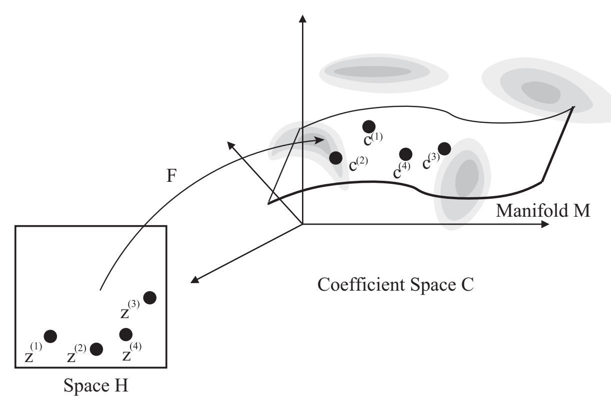

Fig. 2. Mapping F embeds the low-dimensional space $\mathcal{H}$ as a manifold in the coefficient space $\mathcal{C}$. The coefficients $\boldsymbol{c}^{(r)}$ of the regional GMPEs are the images of the parameters $\boldsymbol{z}^{(r)}$ under mapping F.

We now introduce the assumption that the regional GMPEs are organised on a manifold. For each coefficient vector $\boldsymbol{c}^{(r)}$ we postulate a Q-dimensional parameter $\boldsymbol{z}^{(r)}$ that resides in the Q-dimensional Euclidean space denoted by $\mathcal{H}$. We take $Q<D$. We also postulate a smooth mapping $F:\mathcal{H}\to\mathcal{C}$ that takes inputs $\boldsymbol{z}^{(r)}$ and maps them to coefficients $F(\boldsymbol{z}^{(r)})=\boldsymbol{c}^{(r)}$. That is, coefficients $\boldsymbol{c}^{(r)}$ are the images of the low-dimensional $\boldsymbol{z}^{(r)}$ under mapping F. Since $Q<D$, mapping F embeds $\mathcal{H}$ as a manifold $\mathcal{M}$ into $\mathcal{C}$. Hence, coefficients $\boldsymbol{c}^{(r)}$ are now constrained to reside on $\mathcal{M}$ and are no longer free to roam anywhere in $\mathcal{C}$. This is illustrated in Fig. 2. Instead of estimating each coefficient vector $\boldsymbol{c}^{(r)}$ independently, we now attempt to identify the manifold $\mathcal{M}$ that produces the best fit for all regional data.

Mapping F is parametrised by a weight vector $\boldsymbol{w}$. Specifically, we choose F to be a neural network.[1] We now write down the objective for the R regional GMPEs with coefficients $\boldsymbol{c}^{(r)}$ constrained on manifold $\mathcal{M}$:

$$\sum_{r=1}^{R}\sum_{n=1}^{N^{(r)}}(y_n^{(r)} - g(\boldsymbol{x}_n^{(r)}; F(\boldsymbol{z}^{(r)};\boldsymbol{w})))^2. \quad (4)$$

Comparing to Eq. (3), we see that the free parameters are the parameters $\boldsymbol{z}^{(r)}$ and $\boldsymbol{w}$. Parameters $\boldsymbol{z}^{(r)}$ control the position of

[1] We specify that F is a feed-forward neural network with a single hidden layer, hidden neurons that use the tanh activation function, and outputs that are linear (Bishop, 1995).