NGA East Median GMMs

Nicolas Kuehn for the TI team

PEER

17 June 2014

Part I

Continuous Distribution

Motivation

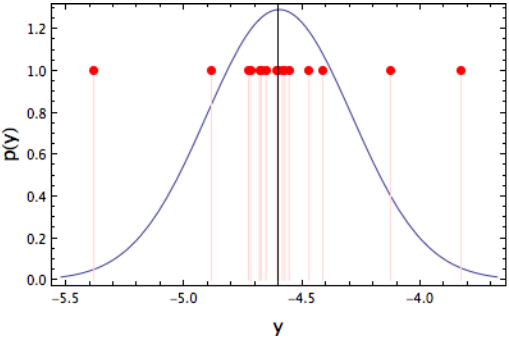
Capture *center*, *body* and *range* of median ground motion predictions.

Approach

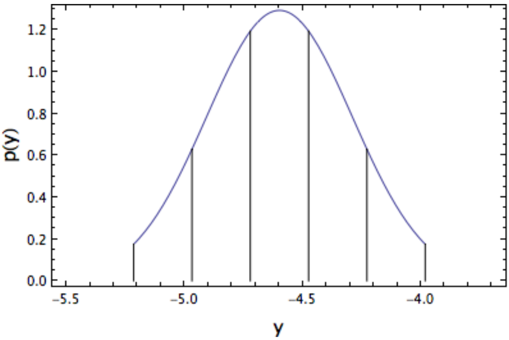
- Select (adjusted) seed GMMs.
- Estimate continuous distribution of median ground motion predictions.
- Discretize continuous distribution to get a manageable number of GMMs.
- Evaluate/Weigh selected GMMs.

- For illustration, assume a one-dimensional distribution.
- Only one M/R-scenario is relevant.
  - Estimate continuous distribution.
  - Discretize space.
  - Select representative models.
  - Assign weights.

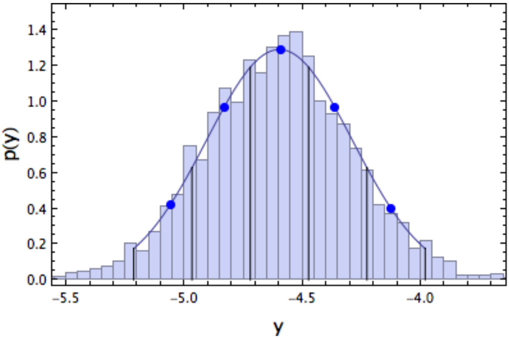
- Predictions of seed GMMs at $M = 6, R_{RUP} = 100$
- Estimated continuous normal distribution $P(Y) = N(\mu, \sigma)$



- Partitioning of 1D-normal distribution into cells to get representative models (for $P(Y)$).
- Each "model" is just a single number.



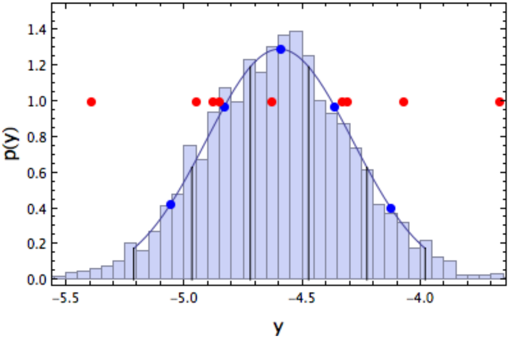
- Sample 5000 “models” from $P(Y)$
- Select samples inside each cell, calculate mean $\Rightarrow$ selected/representative model.



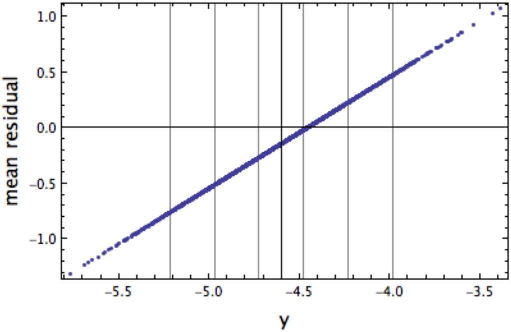
Introduction

Illustrative Example

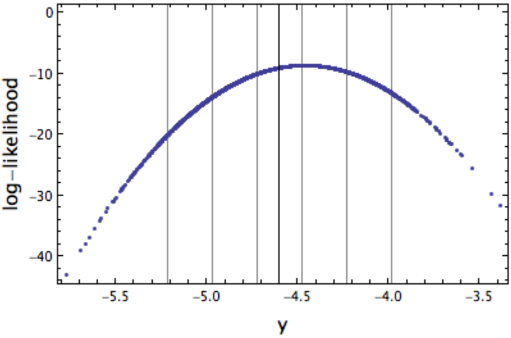
- Assume there is data available, $D = \{y_1^*, \ldots, y_{N_d}^*\}$.
- The data points are distributed according to a normal distribution $Y^* \sim N(\mu^*, \sigma^*)$, where σ^* corresponds to aleatory variability.
- $P(Y)$ describes epistemic uncertainty about μ^*.



- For each sampled model y_i, calculate residual and likelihood to D

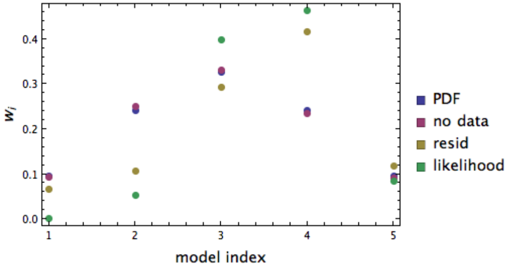


- For each sampled model y_i, calculate residual and likelihood to D



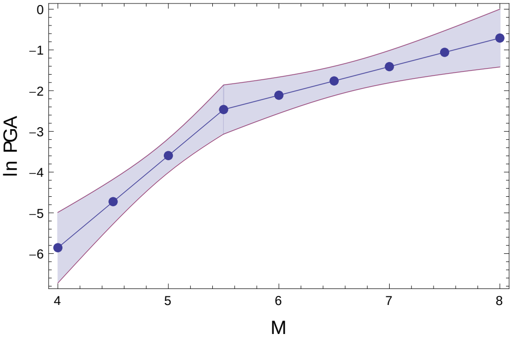
Weights can be calculated proportional to

- number of samples per cell
- integral of density $P(Y)$ over cell
- 1/mean residual per cell
- mean likelihood per cell



- In the example, only one M/R-scenario was considered: one ground-motion value Y, distribution $P(Y)$
- Go to various M/R-scenarios: multiple ground-motio values $\boldsymbol{Y} = \{Y_1, \ldots, Y_N\}$, joint distribution $P(\boldsymbol{Y})$
- Basic approach stays the same.

Ground-Motion Distribution



Assumption:

- function values $\boldsymbol{f}$ are distributed according to a (multivariate) normal distribution $\mathcal{N}_N(\boldsymbol{\mu}, \Sigma)$

Ground-Motion Distribution

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix} \sim \mathcal{N} \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix}, \begin{pmatrix} \Sigma_{1,1} & \dots & \Sigma_{1,N} \\ \vdots & \ddots & \vdots \\ \Sigma_{N,1} & \dots & \Sigma_{N,N} \end{pmatrix}$$

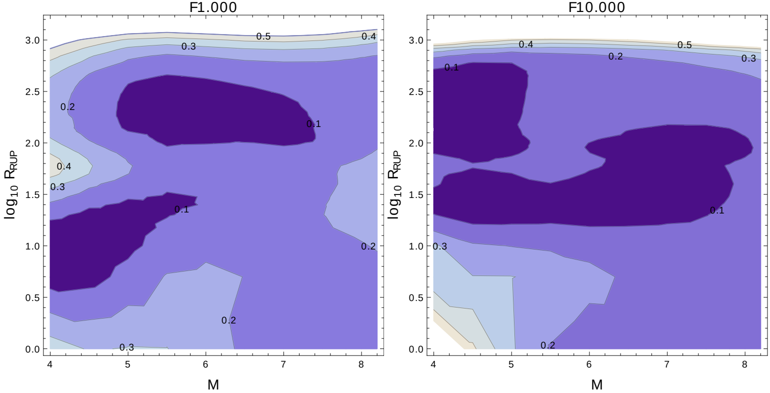
where $\Sigma_{i,j} = \rho_{i,j}\sigma_i\sigma_j$.

A full description of the distribution $P(\boldsymbol{Y})$ requires specification of μ, σ and ρ for all M/R-scenarios. This is done by estimating these terms as functions of M, R.

$$\sigma_i = f_\sigma(M_i, R_i)$$
$$\rho_{i,j} = k(\{M_i, R_i\}, \{M_j, R_j\})$$

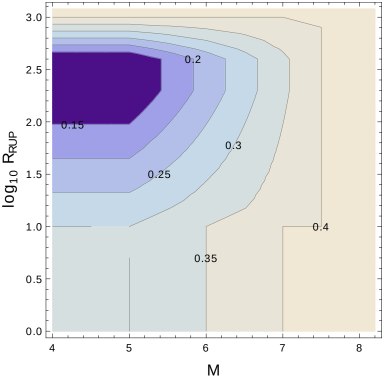
Variance Model

- f_σ is a smooth function
- σ is larger than in the West
- σ is large for large R and large M



Variance Model

- f_σ is a smooth function
- σ is larger than in the West
- σ is large for large *R* and large *M*



Correlation Model

The correlation model is a sum of a rational-quadratic term and a dot-product term. The second term models a linear relationship, the first one models nonlinearities.

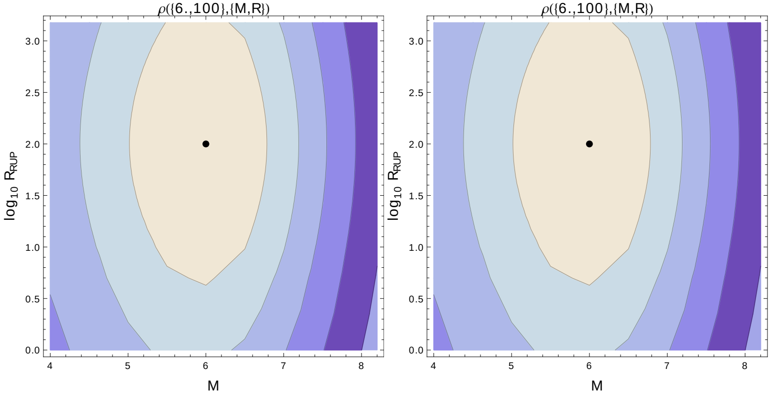
$$k(\boldsymbol{x}, \boldsymbol{x}') = \theta_1 (1 + (\boldsymbol{x} - \boldsymbol{x}')^T \begin{pmatrix} \theta_2 & 0 \\ 0 & \theta_3 \end{pmatrix} (\boldsymbol{x} - \boldsymbol{x}')/(2\theta_4))^{-\theta_4}$$
$$+ \boldsymbol{x}^T \begin{pmatrix} \theta_5 & 0 \\ 0 & \theta_6 \end{pmatrix} \boldsymbol{x}'$$

Correlation Model

$$k(\boldsymbol{x}, \boldsymbol{x}') = \theta_1 (1 + (\boldsymbol{x} - \boldsymbol{x}')^T \begin{pmatrix} \theta_2 & 0 \\ 0 & \theta_3 \end{pmatrix} (\boldsymbol{x} - \boldsymbol{x}')/(2\theta_4))^{-\theta_4} + \boldsymbol{x}^T \begin{pmatrix} \theta_5 & 0 \\ 0 & \theta_6 \end{pmatrix} \boldsymbol{x}'$$

The parameters $\boldsymbol{\theta}$ are estimated by maximizing the marginal likelihood, given mean GMM predictions.

Correlation Model



Correlation Model

- Calculate mean of seed GMMs $\boldsymbol{y}_\mu$ at different M/R-scenarios ($M = 4., 4.5, \ldots, 8., 8.2$, $R_{RUP} = 10. - 1200.$km)
- Maximize marginal likelihood of mean predictions $\boldsymbol{y}_\mu$ at M/R-values $\boldsymbol{X}$ with respect to parameters $\boldsymbol{\theta}$

$$\ln p(\boldsymbol{y}_\mu | \boldsymbol{X}, \boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{y}^T K_Y^{-1} \boldsymbol{y} - \frac{1}{2}\log|K_y| - \frac{n}{2}\ln 2\pi$$

where $K_y = K_f + \sigma^2 \boldsymbol{I}$ and $\boldsymbol{I}$ is the identity matrix. The matrix $K_f = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is the covariance evaluated at the predictor variables (the different M/R-scenarios).

Covariance Model

- Combine the variance model and correlation model
- $\rho_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$
- $\Sigma_{ij} = \sigma_i \sigma_j \rho_{ij}$

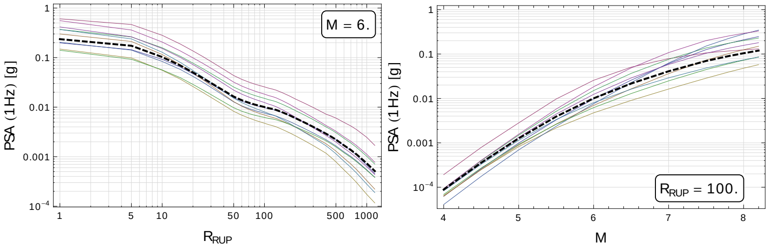
The covariance matrix can be evaluated for any M/R-value. Hence, samples can be drawn for any number of M/R-values. The distribution specified by Σ (and $\boldsymbol{\mu}$) is a distribution over functions.
For further calculations, the following M/R-scenarios are used:

- M = 4., 4.5, 5., 5.5, 6., 6.5, 7., 7.5, 7.8, 8., 8.2
- R = 0., 1., 5., 10., 15., 20., 25., 30., 40., 50., 60., 70., 80., 90., 100., 110., 120., 130., 140., 150., 175., 200., 250., 300., 350., 400., 450., 500., 600., 700., 800., 1000., 1200., 1500.
- 374 M/R-scenarios

- Each resulting GMM is a vector of ground-motion values at these M/R-scenarios.

Continuous Distribution

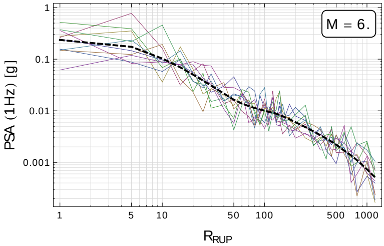
Example

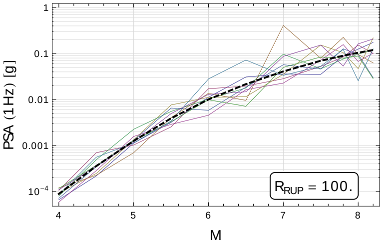


- 10 samples from continuous distribution with covariance model.

Continuous Distribution

Example

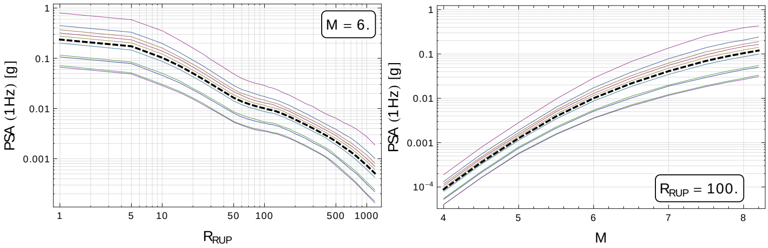




- 10 samples from continuous distribution with no correlation.

Continuous Distribution

Example



- 10 samples from continuous distribution with full correlation ($\rho_{ij} = 1$).

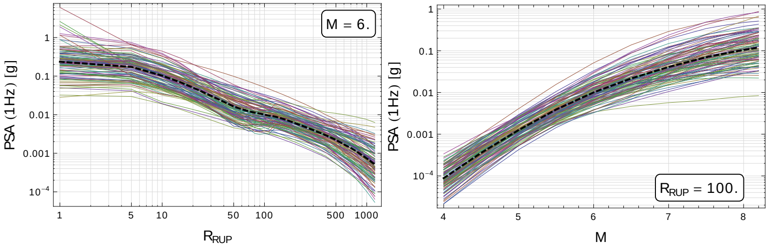
Continuous Distribution

The continuous distribution with the covariance model is a “compromise” between a distribution with full correlation (scaled backbone) and no correlation. The covariance model is closer to full correlation.

- The pointwise (marginal) distribution at each M/R-scenario $P(Y)$ is the same for all three joint distributions $P(\boldsymbol{Y})$.
- The pointwise quantiles are the same.
- The joint quantiles are different.

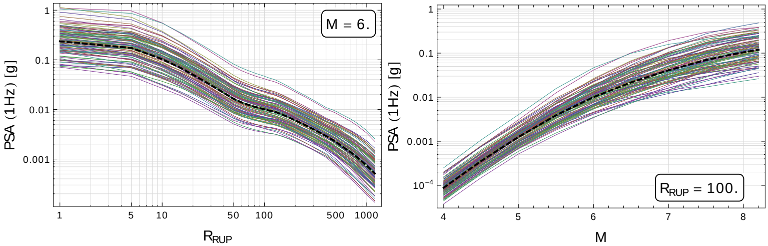
Continuous Distribution

- The correlation between different M/R-scenarios is rather strong.
- This leads to small perturbations in scaling relative to the mean model.
- To retain the characteristics of the seed GMMs, these are used as means
  $\Rightarrow P(\boldsymbol{Y})$ is a mixture model, with covariance defined by the the covariance model and means defined by the seed GMMS.



Continuous Distribution

- The correlation between different M/R-scenarios is rather strong.
- This leads to small perturbations in scaling relative to the mean model.
- To retain the characteristics of the seed GMMs, these are used as means
  $\Rightarrow P(\boldsymbol{Y})$ is a mixture model, with covariance defined by the the covariance model and means defined by the seed GMMS.



Continuous distribution

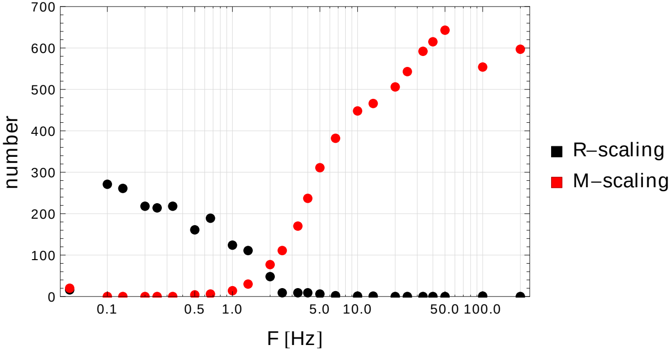
- The continuous distribution $P(\boldsymbol{Y})$ needs to be discretized to have a small, manageable number of GMMs for hazard calculations.
- $P(\boldsymbol{Y})$ is a high-dimensional distribution.
- Draw samples from $P(\boldsymbol{Y})$, combine "similar" samples into one model.
- Similarity is defined by the average GM-distance between GMMs.
- Models are visualized in 2D such that GM-distances in 2D correspond to GM-distances in HD.

5000 samples are drawn from $P(\boldsymbol{Y})$ according to:

1. Randomly select one of the seed GMMs.
2. Sample from a multivariate normal distribution with the seed GMM as mean and the covariance model.
3. Check if sample passes criteria for physicality.
   - if yes, proceed
   - if no, go to step 1
4. Add sample to list of sampled models.

Criteria for physicality:

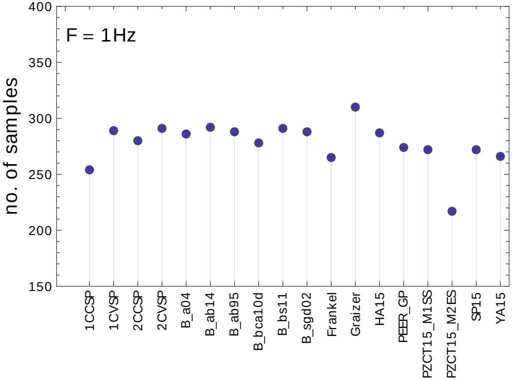
- $Y(M = 6, R) > Y(M = 5, R)$ for $R \geq 10$
- $Y(M = 7, R) > Y(M = 6, R)$ for $R \geq 10$
- The distance slope between $R_{RUP} = 10$ and $R_{RUP} = 40$ must be larger than 0.4: $\mathrm{sl}(10, 40) > 0.4$
- $\mathrm{sl}(40, 150) > -0.2$
- $\mathrm{sl}(150, 400) > \min[0.45, 0.9 \min(GMM)]$



- Number of rejected samples.

Sampling

Physicality



- Number of samples per model in set of 5000 samples

Continuous Distribution

- Joint continuous distribution over ground-motion vectors (functions) $\boldsymbol{Y}$, $P(\boldsymbol{Y})$.
- 5000 (physical) samples drawn from $P(\boldsymbol{Y})$ at 374 M/R-scenarios for each frequency.
- Appreciation of samples/GM-space is done by visualization tools.