Application of semiparametric function-on-scalar regression for modeling spatio-temporal earthquake dynamics

Alexander Bauer, Fabian Scheipl, Helmut Küchenhoff and Alice-Agnes Gabriel

Abstract We present a generalized functional additive regression model for modeling spatio-temporal earthquake dynamics based on function-on-scalar regression. The flexibility of our semiparametric, spline-based approach is highlighted, both with focus on estimating complex effect structures and on accounting for spatial and temporal correlation structures in the data. For estimation based on this large data set of earthquake dynamics, a recently developed efficient algorithm for penalized likelihood-based inference is used.

Key words: Functional regression, semiparametric regression, generalized additive model, spatio-temporal analysis, geophysics

1 Introduction and data

While the physics of seismic waves is fairly well-understood, the ground movement triggered by them, which is highly relevant for earthquake hazard assessment, is difficult to predict due to its complicated dependencies with the local topography, subsurface velocity structure and earthquake source effects. Furthermore, current earthquake models rarely account for the effects
of earthquake faulting dynamics. To tackle both of these challenges, we use a generalized functional additive regression model to quantify how physical conditions at an earthquake fault as well as local topography and geology affect the surficial ground velocity measured over time during a seismic event.

We analyse simulated earthquake data, derived from a large scale computer experiment with the open source software SeisSol (Breuer et al., 2014; Pelties et al., 2014, www.seissol.org) and based on a real earthquake that took place in Northridge (USA) in 1994. Absolute ground velocities were simulated solving elastic wave equations and are defined as the (isotropic) $L^2$ norm of the ground velocities in east-west, north-south and vertical direction. Each of the simulations used a different set of physical conditions at the fault. Surficial ground velocities were then recorded at 6146 virtual seismometers, most of them located on a regular grid of 198km x 166km. For the analysis, the first 15s of the absolute ground velocity measurements from 135 simulations were used in a resolution of 2Hz. Leading zeros were discarded until the first relevant observation ($\geq 0.01$). The data used for modelling comprise $1.5 \cdot 10^7$ space-time coordinates in total.

![Typical observations by hypocentral distance](image)

**Fig. 1** Left: Area under study and the steep subjacent fault with its projection to the surface; hypo- and epicenter are marked red. The points in the lower left corner illustrate the surficial measurement station density as present over the whole area. Right: Typical observations of absolute ground velocity over time. The initial peak of the ground velocity is delayed and smaller as hypocentral distance increases.

The evaluated predictors were all constant over time. Our main interest was in five physical parameters, which were pre-set in each simulation and which consisted of three frictional resistance variables, the direction of the regional tectonic background stress and the binary soil material of the whole area ($\{\text{rock, sediment}\}$). The moment magnitude as the classical measure of earthquake size, the local topography as well as the height and the hypocentral distance of each seismometer were additionally included.
2 Semiparametric function-on-scalar regression

Following Greven and Scheipl (2017), the generalized functional additive model for function-on-scalar regression is written in the form

\[ Y_i(t) | X_i \sim F(\mu_{it}, \nu) \]

\[ g(\mu_{it}) = \beta_0(t) + \sum_{r=1}^{R} f_r(X_{ri}, t), \]  

with seismometers \( i = 1, \ldots, n \), absolute seismic ground velocity \( Y_i(t) \) recorded at specific points in time \( t \) and the number of additive effects \( R \) with associated covariates \( X_{ri} \). Conditional on the additive predictor and the known response function \( g(\cdot) \), \( Y_i(t) \) is assumed to follow some given distribution \( F \) with expectation \( \mu_{it} \) and dispersion and shape parameters \( \nu \). For our application we use a Gamma distribution with exponential response function \( g(\cdot) \).

Representing each functional additive effect \( f_r(X_{ri}, t) \) in terms of a (tensor product) P-spline basis, the semiparametric approach offers great flexibility in incorporating nonlinear uni- or multidimensional effects. For estimating the model we use the function \texttt{ppfr} from the open source \texttt{R} package \texttt{refund} (Goldsmith et al, 2016), which is based on the \texttt{gam} function for scalar additive regression from the \texttt{mgcv} package (Wood, 2017, v. 1.8-23). Additional convenience functions for visualizing and evaluating function-on-scalar models are implemented in our \texttt{R} package \texttt{FoSIntro} (Bauer, 2017), available on GitHub.

A prediction error-based approach was used for tuning basis sizes as the high-dimensional data dominate the penalization prior in the estimation. The model is fitted on almost half a million data points and comprises 18 smooth effects with a total of 633 spline coefficients and 25 smoothing parameters. To make estimation of this complex model on such large data feasible we used a recently developed highly performant algorithm for penalized likelihood-based inference from Wood et al (2016). Its major advances are the use of a highly efficient and parallelizable block-wise Cholesky decomposition and a compressed representation of the (marginal) spline basis. Using the efficient estimation algorithm lead to a substantial speed-up in computation time. Based on a Windows system with two 2.60 GHz Intel E5 CPU’s with 64GB of RAM and 16 cores model estimation could be reduced from 18:25h with the standard QR-decomposition based algorithm to 04:02 min based on the efficient algorithm, even without using parallelization. Additional use of multithreading with 10 threads for parallelization lead to a final runtime of 01:55 min.

Local small-scale (radius 300m) and large-scale (2000m) topography at each measuring station was included via the Topographic Position Index (TPI) introduced by Weiss (2001), defined as the height difference between the seismometer and the mean height in a circular neighborhood.
3 Modelling spatio-temporal data

As each measurement at a station in one simulation was recorded over a specific time interval, we use the time domain as the natural functional domain and view our data as a sample of spatially referenced functional observations. Underlying all functional observations in our application is a regular grid of points in time. However, the chosen regression approach is also applicable for irregular grids over the functional domain.

Based on penalized spline-based estimation the semiparametric approach allows us to automatically choose the amount of flexibility needed to properly account for the response structure over the functional domain. As stated in the previous section, this also includes more complex effect structures. E.g., such regression models can be easily extended by incorporating multidimensional effects that vary over the functional domain.

Curve-specific functional random intercepts \( E_i(t) \) can be added to the predictor to account for intra-functional dependencies. Those smooth error terms can be used to incorporate possible autocorrelation and variance heterogeneity along the functional domain (Scheipl et al, 2015).

The choice of how to account for spatial structure mainly depends on whether the spatial effect is of main interest or if it should only be accounted for to reliably estimate standard errors of the remaining covariates, as is the case in our application.

While in the former case estimation of the raw spatial effect – e.g. by setting up a 2D tensor product spline basis over the coordinate space – is principally reasonable, one should be aware that applications with a spatial dimension often bear the problem that covariates themselves or their corresponding effect structures vary systematically over space, often leading to non-negligible correlation between the spatial effect and the remaining covariates. In such cases the inclusion of a spatial effect introduces bias to the estimation of other covariate effects (see e.g. Hodges and Reich, 2012).

This can also be an issue if the spatial structure is not of main interest and not estimated using a spline-based approach. E.g., also the inclusion of additional (smooth) error terms that follow a Gaussian Random Field with a specific spatial correlation structure (Scheipl et al, 2015, Supplement 3) potentially leads to biased effect estimates. Therefore, Hodges and Reich (2012) propose to use so-called “restricted spatial regression”, only including the part of the spatial effect that is orthogonal to the remaining linear predictor. In this way the introduction of bias can be generally prevented as the final spatial effect only contains information unexplained by the other covariates. For our application we will make use of this approach, adapting the method with regard to a functional response.
4 Results

A Gamma model with exponential response function \( g(\cdot) \) reduces the deviance by 70% compared to a simple intercept model. Figure 2 visualises the effects of hypocentral distance and the dynamic coefficient of friction, which show by far the strongest effect among all covariates. While the former shows a clear nonlinear and time-varying effect structure, resulting in differently shaped predictions (see e.g. the delayed peak ground motion), the latter has a linear and time-constant effect, i.e. varying coefficient of friction values only affect the overall ground motion level, but not the shape of the curves. Additional results can be found in Bauer (2016). All the estimated effects seem geophysically plausible.

![Figure 2](image)

**Fig. 2** Left: Nonlinear, time-varying effect of hypocentral distance as heatmap and 3D surface, and predictions based on varying hypocentral distances, while other covariates are held constant at realistic values. Right: Predictions based on varying values of the dynamic coefficient of friction, which has a linear, time-constant effect of \(-5.48\).

A selection of residual plots is shown in Figure 3. As noted above, these plots evaluate the final model, which at the time writing is not based on a correction strategy for spatial autocorrelation as outlined in Section 3. While the structure of the residuals plotted against the fitted values (panel 1) is still acceptable, evaluating the mean residuals over space (panel 2) shows that substantial spatial structure remains in the residuals. Across time we again observe an acceptable amount of residual structure. No systematic deviations from a constant trend at zero are present, but some extreme peak ground velocities can be observed at around five seconds (panel 3). The empirical autocovariance of the residuals (panel 4) corresponds well enough to the model assumptions: it is fairly constant along the diagonal (i.e., the variance of the residuals is fairly homogeneous over functional domain \( t \)) and drops off quickly towards zero away from it.
5 Conclusion

Functional additive regression models are a promising approach for modelling surficial ground velocity. The semiparametric, spline-based approach offers great flexibility in estimating differently shaped uni- or multidimensional effects and accounts for temporal response structures in a natural way by incorporating time as the functional domain. Approaches to account for spatial and temporal correlation structures in the data were outlined, giving an outlook on future refinements of the present model. The use of an efficient estimation algorithm was highlighted that speeded up estimation time by over 99%. Major effects are estimated for hypocentral distance and the dynamic coefficient of friction, with higher values leading to decreased ground velocities for both, resulting in valuable novel insights for geophysical research.

Acknowledgements Financial support by the German Research Foundation for the development of refund::pffr through the Emmy Noether Programme, grant GR 3793/1-1 is gratefully acknowledged

References

Hodges JS, Reich BJ (2012) Adding spatially-correlated errors can mess up the fixed effect you love. The American Statistician 64(4):325–334