Segmentation of Fault Networks Determined from Spatial Clustering of Earthquakes

Journal Article

Author(s):
Ouillon, G.; Sornette, Didier

Publication date:
2010-02

Permanent link:
https://doi.org/https://doi.org/10.3929/ethz-c-000029144

Originally published in:
Journal of Geophysical Research: Solid Earth 116, https://doi.org/10.1029/2010JB007752

JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 116, B02306, doi:10.1029/2010JB007752, 2011

Segmentation of fault networks determined from spatial clustering of earthquakes

G. Ouillon[1] and D. Sornette[2]

Received 3 June 2010; revised 14 October 2010; accepted 22 November 2010; published 9 February 2011.

[1] We present a new method of data clustering applied to earthquake catalogs, with the goal of reconstructing the seismically active part of fault networks. We first use an original method to separate clustered events from uncorrelated seismicity using the distribution of volumes of tetrahedra defined by closest neighbor events in the original and randomized seismic catalogs. The spatial disorder of the complex geometry of fault networks is then taken into account by defining faults as probabilistic anisotropic kernels. The structure of those kernels is motivated by properties of discontinuous tectonic deformation and by previous empirical observations of the geometry of faults and of earthquake clusters at many spatial and temporal scales. Combining this a priori knowledge with information theoretical arguments, we propose the Gaussian mixture approach implemented in an expectation maximization (EM) procedure. A cross-validation scheme is then used that allows the determination of the number of kernels which provides an optimal data clustering of the catalog. This three-step approach is applied to a high-quality catalog of relocated seismicity following the 1986 Mount Lewis (M_l = 5.7) event in California. It reveals that events cluster along planar patches of about 2 km^2, i.e., comparable to the size of the main event. The finite thickness of those clusters (about 290 m) suggests that events do not occur on well-defined and smooth Euclidean fault core surfaces but rather that there exist a deforming area and a damage zone surrounding faults which may be seismically active at depth. Finally, we propose a connection between our methodology and multiscale spatial analysis, based on the derivation of a spatial fractal dimension of about 1.8 for the set of hypocenters in the Mount Lewis area, consistent with recent observations on relocated catalogs.

Citation: Ouillon, G., and D. Sornette (2011), Segmentation of fault networks determined from spatial clustering of earthquakes, *J. Geophys. Res.*, *116*, B02306, doi:10.1029/2010JB007752.

1. Introduction

[2] Earthquakes are thought to be clustered over a very wide spectrum of spatial scales (see *Kagan and Knopoff* [1978, 1980], *Kagan* [1981a, 1981b, 1994], *Sornette* [1991, 2006], and *Sornette et al.* [1990] for the main concepts and observations). At the largest, worldwide scales, the dominating feature of earthquake epicenters maps is the existence of narrow corridors of activity, coinciding with boundaries of tectonic plates. Observation of earthquake distribution at smaller and smaller scales (down to the standard accuracy of the spatial locations, typically a few kilometers, or below in the case of relative locations) successively reveals smaller-scale corridors (or surfaces in 3-D), often linked to faults by geologists and seismologists.

[3] The identification of faults and of the statistical properties of their associated seismicity is a major challenge in order to delineate the potential sources of future large events that may have a societal impact [see *Plesch et al.*, 2007]. It is also necessary to clearly understand the mechanical interactions within a given fault network, in order to possibly achieve earthquake prediction or forecasting with a reasonable accuracy, both in the space and time dimensions. On more fundamental grounds, understanding the relationships between faults (or more generally earthquake generating cluster centers) at different scales can help to develop a metatheory based on the renormalization group approach [*Wilson*, 1975; *Allégre et al.*, 1982] to advance on the understanding and modeling of seismicity and rupture processes as a critical phenomenon [*Sornette*, 2006; *Bowman et al.*, 1998; *Weatherley et al.*, 2002], as the size of an upcoming event may be correlated with the range of scales over which earthquake clustering properties change collectively (thus yielding a measure of the correlation length).

[4] Fault identification from seismicity alone is a complex problem, whose success mainly depends on the seismotectonic experience and intuition of the operator, as it demands

[1]Lithophyse, Nice, France.
[2]D-MTEC and Department of Earth Sciences, ETH Zurich, Zurich, Switzerland.

0148-0227/11/2010JB007752

to incorporate data possessing dimensions of different natures (spatial location of hypocenters, rupture lengths or surfaces, magnitudes or seismic moments, and focal mechanisms of earthquakes), some of which being determined with variable and sometimes large uncertainties or ambiguity (like focal mechanism, for instance). Some properties of seismicity being highly heterogeneous even on well-identified faults, the identification thus essentially remains a subjective process, without automatization (which would require quantitative criteria to be fulfilled).

[5] The present paper will focus on the automatic identification of faults solely from the locations of seismic events, as recorded in a standard earthquake catalog, assuming that earthquake clusters reveal the underlying existence of such tectonic structures. This hypothesis is now universally accepted, at least for shallow events. We shall thus first provide the main definitions of clustering that are used in this work in section 2, as those definitions may differ in the pattern recognition and the geosciences fields, respectively. We refer to *Ouillon et al.* [2008] for a review of existing data clustering techniques applied to earthquakes. We advise the reader to go to the related section of *Ouillon et al.* [2008] as it underpins the argumentation developed in the present article. Section 3 presents some very general physical and morphological arguments about the nature of the tectonic strain field, which help us to choose the appropriate data clustering methodology. This method is presented in section 4. It necessitates the introduction of a probabilistic kernel that defines the spatial structure of the set of events triggered by a given fault. Section 5 presents the theoretical and empirical knowledge about earthquakes and faults in order to choose more rigorously the explicit shape of this probabilistic kernel. Section 6 uses the results presented in section 5 together with some information theoretical arguments in order to provide an analytical expression for the probabilistic kernel. The potential flaws of this kernel are outlined and discussed as well. The resulting data clustering procedure is then applied to a synthetic example in section 7 where we also introduce a cross-validation scheme that allows us to determine objectively the total number of kernels that should be used to fit the catalog. Section 8 focuses on the natural catalog of the Mount Lewis area, following the $M5.7$ event that occurred in 1986. Its analysis requires a preliminary treatment to remove events that occurred on faults that are undersampled by seismicity. The application of our three-step methodology to this natural catalog provides important information about the possible structure of fault zone segmentation and activity at depth that are discussed in section 9.

2. Standard Clustering Techniques

2.1. What Is Clustering?

[6] It is useful to make precise the definition of the term "clustering" that will be used throughout this paper. Indeed, while seismologists and statisticians study the spatial and/or temporal properties of the same set of events, both may not use the term clustering exactly in the same sense.

[7] In statistical seismology, *clustering* refers to the concentration of earthquakes into groups that form locally dependent structures in time and space. The most obvious form of space and time clustering is the pattern defined by aftershocks. Identifying the clusters then often amounts to determining the so-called main events, which are interpreted as the triggers of the following aftershocks. The procedure consisting in identifying the main shocks is indeed usually defined as *declustering* in the seismological literature, i.e., the reduction of a whole catalog to the set of the main shocks by removing the corresponding aftershocks. There are several standard declustering methods [*Gardner and Knopoff*, 1974; *Knopoff et al.*, 1982; *Reasenberg*, 1985; *Zhuang et al.*, 2002, 2004; *Marsan and Lengliné*, 2008; *Pisarenko et al.*, 2008], but recent developments suggest that the distinction between main shocks and aftershocks may be more a human construction than reflecting a genuine property of the system [*Helmstetter and Sornette*, 2003a, 2003b]. In any case, earthquake clustering considers that correlations between earthquakes stem from causal relationships between events.

[8] In contrast, statistical pattern recognition aims at determining the local similarities or correlations between events in a given data set [*Bishop*, 2007; *Duda et al.*, 2001]). Those similarities may have different natures, such as size, spatial location and so on. The identification of clusters then consists in grouping together events that share those similarities, without the consideration of any specific event that would be genetically responsible for all the others. This process is called *clustering*. Clearly, its motivation is that correlations among events stem from a common cause, i.e., the existence of a common underlying fault.

[9] As those terminologies can be rather confusing, we shall use the terms *earthquake clustering* when dealing with the phenomenology of seismic events triggering other seismic events, and *data clustering* when dealing with the identification of clusters. Our focus is to determine earthquakes that are clustered in the second sense, with each cluster being a candidate fault, i.e., a cluster of events will be conjectured to occur on the same fault. In contrast, the events involved in earthquake declustering may occur on different faults.

2.2. Main Categories of Data Clustering Techniques

[10] There are many available pattern recognition techniques for data clustering. Rather than presenting an exhaustive review, the interested reader is referred to *Bishop* [2007] and *Duda et al.* [2001], who describe a large number of data clustering techniques. Some others, which have been applied to earthquakes, are briefly discussed by *Ouillon et al.* [2008]. From now on, we shall focus exclusively on data clustering techniques in the spatial domain, except when explicitly mentioned.

[11] *Ouillon et al.* [2008] have suggested a clustering algorithm that fits an earthquake catalog with rectangular plane segments of adjustable sizes and orientations. This method is a generalization to anisotropic sets of the popular k-means method [*MacQueen*, 1967]. *Ouillon et al.* [2008] proposed an iterative version, in the sense that planes are introduced one by one in the fitting procedure until the discrepancy between the proposed fault pattern and the spatial structure of the earthquake catalog can be explained by location errors only. *Ouillon et al.* [2008] found that the final fault pattern provided by the fit is in general in rather good agreement with the fault pattern inferred by other methods (such as maps of surface fault traces). However, the assumption that fault elements are perfect rectangular plane

segments is certainly an oversimplification. One of the advantage of the method presented here is to improve on this reductionist assumption.

[12] The main drawback shared by previous clustering approaches and also, but to a lesser degree, by the anisotropic k-means of *Ouillon et al.* [2008], is that these methods are indiscriminately applied to earthquake catalogs without recognizing their specific characteristics. As a consequence, it is difficult to decide a priori which cluster solutions are to be chosen among the various constructions offered by the different clustering techniques. In this paper, we propose to incorporate the existing knowledge about earthquakes and faults into the definition of the data clustering technique to be applied to earthquake data sets. Section 3 presents a first general argumentation based on very general and qualitative physical and geological properties, while more empirical and quantitative issues are discussed in section 5. The available knowledge as well as the identification of the unknown are combined with arguments from information theory to form the basis of our approach.

[13] At its most basic level, data clustering consists in identifying "unusual" discrepancies between a spatial distribution of data and a smooth reference distribution. The homogeneous Poisson distribution is often considered as the reference distribution, so that clusters emerge as anomalies from a statistically uniform background. This idea is very intuitive but still remains to be proven as adequate for earthquakes. In section 8.2 we discuss the statistical properties of the reference distribution that we should consider for the case of earthquakes, using simple arguments about the symmetry of the involved physics and of the rheological nature of the Earth's crust.

3. Some Preliminary Conjectures About the Probable Nature of Earthquake Clusters

[14] We propose that an appropriate data clustering procedure for earthquakes should be informed by the probable morphological nature of individual earthquakes as well as their relationship with the strain field within a tectonic plate. The time-varying strain field imposed at the boundaries of a given tectonic domain is partially accommodated within this domain through a variety of continuous rheological processes. Elastic, viscoelastic and viscoplastic deformation mechanisms provide a smooth (i.e., regular) spatial variation of strain that can, for instance, be monitored using techniques such as GPS, leveling, strain rosettes, and so on. Such data are generally sampled at a rather coarse resolution and concern only very recent periods of time, especially for the most sophisticated and accurate ones. The regularity of their spatial variation ensures that they can be interpolated with sufficient confidence both in space and time.

[15] In addition, rupture processes or surface friction instabilities contribute in general significantly to the overall strain balance. Compared with the continuous rheological processes, they correspond to mathematical singularities, labeled as *earthquakes* at small timescales and *faulting process* at longer timescales. One can typically observe only a very small number of such strain field discontinuities, as a result of instrumental limitations preventing their detection. For example, in any area, there exists a minimum magnitude below which one fails to detect earthquakes in a systematic manner. The same holds true for small faults or as a result of vegetation or urban masks at the surface. It is commonly believed that the observed singularities represent only a very small sample of a much larger set, that is, the issue of missing and censored data is acute.

[16] Statistical seismology and tectonics have to deal continuously with drastically undersampled and missing data. The impact of under sampling may be different, depending on the property to characterize. In the case of earthquakes, if we assume that their magnitude distribution obeys the Gutenberg-Richter law, this impact depends solely on the b value. We shall hereafter assume that it is equal to its observed standard value, i.e., $b = 1$. If one is interested in energy budgets or in strain fluctuations, the largest earthquakes dominate the balance. Therefore, in this case, the under sampling of small earthquakes has a relatively minor impact. If we are interested in estimating the stress fluctuations within a given spatial domain, all magnitude ranges are believed to contribute with similar amplitude to the balance. Of course, there are subtleties in the way the sizes of earthquakes control the spatial dependence of the stress fluctuations, as large events are probably associated with longer range stress correlation while small events exhibit stronger collective spatial dispersion. In that case, undersampling affects the stress balance proportionally to the magnitude detection threshold.

[17] As we are interested in the geometry of the set of event hypocenters, we have to recognize that the smallest events dominate by far in the catalog as they are much more numerous. However, they are conspicuously under sampled and actually absent below the magnitude detection threshold. For example, assume that the Gutenberg-Richter law holds down to magnitude $M = -1$ and that the magnitude detection threshold is $M = 2$ as is typical. Then, the job of determining the spatial organization of earthquakes and the fault network on which they occur has to be carried out on just 0.1% of the total number of events, while 99.9% of the earthquakes are missing in the available catalog. This would have a minor impact on statistical analyses if all clusters featured a similar spatial density of events. However, seismicity is known to behave highly intermittently both in time and space, so that the characterization of regions of low activity is definitely penalized by the undersampling. This pessimistic observation must be balanced by the fact that the recorded seismic events reveal some of the most correlated part of the singular strain field, as they propagate over larger distances. These located events occur on a subset of the fault network that is, by construction, associated with potentially dangerous earthquakes. In that sense, reconstructing the structure that supports such events certainly provides meaningful information for the quantitative estimation of seismic hazard. This conclusion is supported by *Hauksson et al.* [2006], who compared the locations of earthquakes and those of the known major fault in California. They observed that, on average, larger events stand closer to the major faults than smaller events do.

4. Decomposing a Data Set as a Mixture

4.1. Complete Versus Incomplete Data Sets

[18] The generic problem we address is schematically represented by comparing Figure 1a with Figure 1b. In

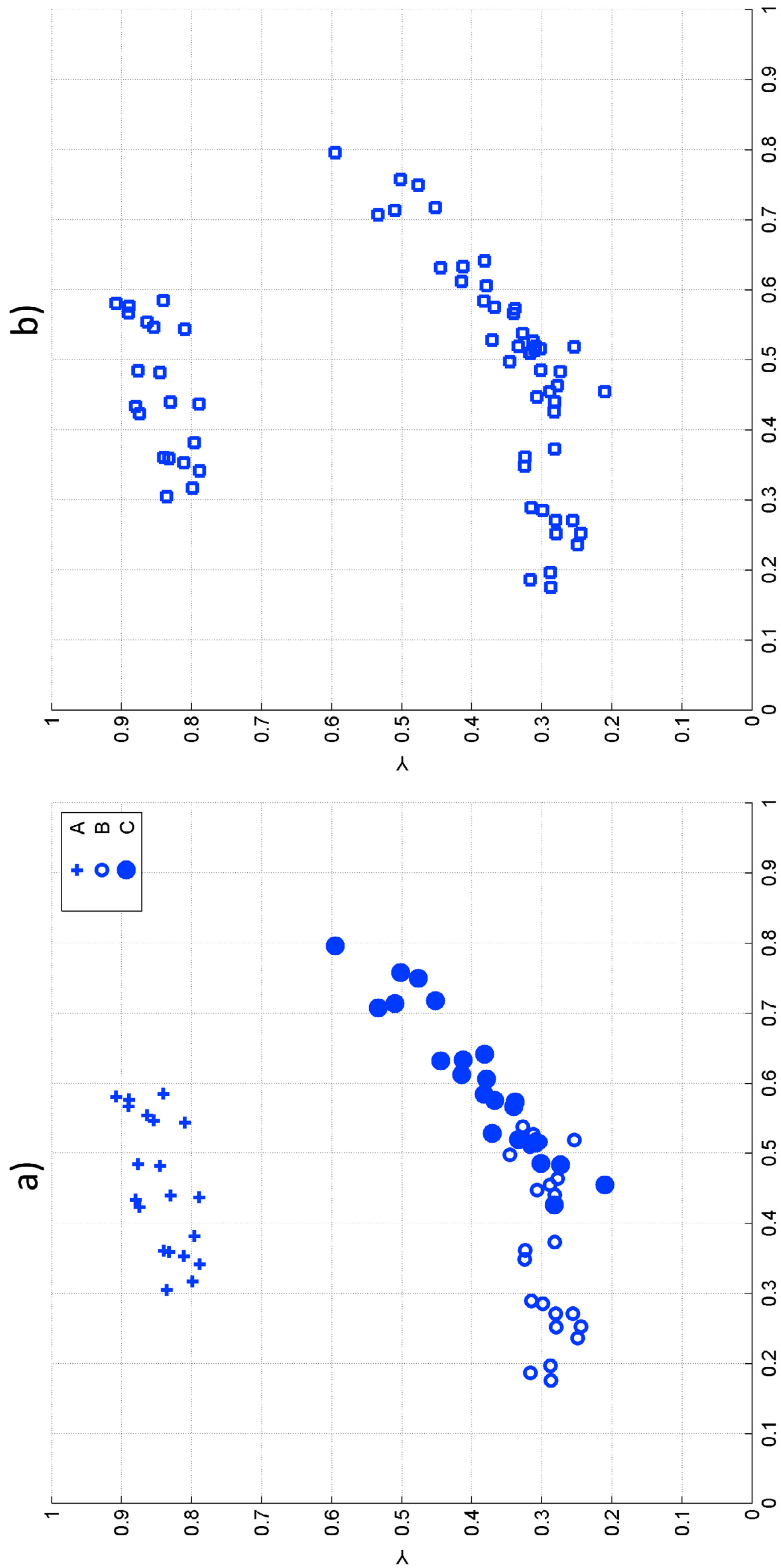

Figure 1. (a) Synthetic 2-D data set featuring three clusters. Each event is represented by a symbol (open circle, solid circle, or cross) which identifies the cluster it belongs to (A, B, or C). (b) Same data set as Figure 1a, after removal of the genetic origin of events.

Figure 1a, three synthetic event clusters are represented in a two-dimensional projection (corresponding to earthquake epicenters). The three synthetic clusters represent model earthquakes occurring on three distinct faults. The three clusters were generated as follows: (1) define three linear segments over which data points are randomly located and (2) add a random perturbation to the position of each point, thereby keeping the memory of the linear segment over which the point was initially located.

[19] Each event is represented using a different symbol according to which cluster (fault) it belongs to. Reciprocally, events with the same mark (i.e., represented by a common symbol) inform us that they should be classified within a common cluster. We thus know that all events represented by an open circle belong to the same cluster (that we label A, for instance), all events plotted using a solid circle belong to the same cluster B, while all events plotted with a cross belong to the same cluster C. Each event is thus characterized by a position (x, y) and a label (A, B, or C), thus defining a *complete data set* [see *Bishop*, 2007]. Using this complete data set, it is then very easy to provide estimates of the characteristics of the clusters A, B, and C, such as their spatial location, size, and orientation. It is then also straightforward to compute for each cluster the statistical properties of any mark its events could carry.

[20] Earthquake catalogs do not belong to this category, in the sense that events that are recorded do not naturally feature any label informing us about the fault on which they occurred or that they occurred on a common fault. The typical data set of epicenters is exemplified by the set shown in Figure 1b. This is the same set as the one represented in Figure 1a. Because we do not have the luxury of knowing a priori neither the generating fault nor any information on cluster association, all events are represented by a common symbol (a square), which informs us only on the position of each event. When going from Figures 1a to 1b, we have also lost the information on how many faults are present in the system. The event locations alone define the *incomplete data set*, while the missing information (labels and number of labels) define the *latent variables* [see *Bishop*, 2007]. Our task is to process the data provided in Figure 1b to attempt recovering the original partition of Figure 1a, i.e., to explain the location of the events by estimating a solution for the latent variables. This problem is perhaps obvious for events that are plotted using crosses on Figure 1a, as they still appear as a clearly isolated cluster in Figure 1b. But even here, one could recognize two or three subclusters, suggesting the need for three faults for just this cluster. This gives a first taste of the role of location errors in the detection of relevant structures. The problem is even more difficult for the events shown as open or solid circles in Figure 1a, as the two clusters overlap.

[21] In the following, we first work out the methodology to recover the clusters and faults, given the knowledge of their number (3 in the example of Figures 1a and 1b). As this number is unknown in real seismic catalogs, we then develop a method to determine this value from the data set, which is presented in section 7. This methodology will also take into account individual earthquake location uncertainties and will evaluate the significance of the proposed data clustering solution. All the considerations and methods discussed apply to the 3-D case without difficulty.

4.2. Hard Versus Soft Assignment

[22] One can distinguish two broad classes of clustering approaches, called “hard” and “soft” assignment methods. Hard assignment characterizes the data clustering methods such as k-means [*MacQueen*, 1967]. See *Ouillon et al.* [2008] for a general presentation of the method and its extension to the representation with planes of the spatial organization of earthquakes recorded in catalogs. The k-means is an iterative and nonprobabilistic method that partitions a set of N data points into a subset of K clusters, each cluster being characterized by its barycenter and variance (or its best fitting plane and covariance matrix as in the work by *Ouillon et al.* [2008]). The k-means method aims at minimizing the global variance of the partition, and provides at least locally reasonable, if not absolute, minima. When convergence has been reached, each event belongs to the cluster defined by the closest barycenter (or plane) with probability $P = 1$, while it has a zero probability of belonging to any other cluster. This way of partitioning data certainly helps in providing a rough approximation of the partition, but clearly lacks subtlety when clusters overlap or when location uncertainty is large. For instance, for the set of events shown in Figure 1b, in the region where the two clusters A and B intersect each other, the hard assignment method is bound to incorrectly attribute with probability $P = 1$ some events to cluster A, whereas they indeed belong to cluster B, and vice versa.

[23] In contrast, soft assignment gives for each event n the set of probabilities $P_n(A)$, $P_n(B)$ and $P_n(C)$ that it belongs to A, B, or C, with $P_n(A) + P_n(B) + P_n(C) = 1$. In that approach, each cluster consists of a more or less localized probabilistic kernel so that the full data set is considered as a stochastic realization of a linear *mixture* of such probabilistic kernels. If the spatial domain over which the kernel is defined is infinite, then any point located at a large distance from a cluster has a non zero probability to belong to it. Certainty is thus banished from this view. It seems therefore well-adapted to those cases where data are located with uncertainty, as encountered in earthquake catalogs. We shall see later that there exist much deeper reasons for using a probabilistic approach to the data clustering of seismicity. An important notion underlying the mixture approach is that the events within each cluster are independent identically distributed. This means that any pattern in the set of data points is assumed to be accounted for by the property of belonging to just one cluster. Thus, within a cluster, all events are assumed to play the same role and are only related to each other through the property of belonging to this cluster. In other words, any spatial correlation among events featuring the same label is considered to be only due to the fact that they have been generated by a common localized kernel, so that events are correlated because they have the same cause. The idea of events triggering (or shadowing) some other events is thus completely foreign to this description. This notion indeed holds for the synthetic set of events shown on Figure 1a as none of those events has been generated as a function of the position of the other events.

[24] In statistical seismology, this would mean that preexisting faults generate earthquakes, but that earthquakes do not interact with each other, whether they occurred on the same fault or on different faults. A proper use of this technique would thus require first to decluster (in the seismological sense) an earthquake catalog, thus identifying all triggered events, and to keep only the set of independent events to be decomposed as a mixture. Such declustering methods exist, as mentioned above, and are usually employed to study aftershock sequences. Unfortunately, they are either dependent on a set of arbitrary parameters (such as the size of spatial and temporal windows within which two events are considered as causally correlated, and their dependence on the size of the first, triggering event), or on a preassigned mathematical model (as in the ETAS model [see *Helmstetter and Sornette*, 2002, and references therein]). In the end, the result may also depend heavily on the magnitude detection threshold [*Sornette and Werner*, 2005b]. The lower this threshold, the smaller the number of events that are qualified as independent, so that too few earthquakes could be used in a data clustering scheme.

4.3. The Expectation Maximization (EM) Algorithm

[25] The expectation maximization (or EM) algorithm is an iterative method that allows one to perform the mixture decomposition mentioned in section 4.2. In its most standard form, it requires several assumptions [*Bishop*, 2007; *Duda et al.*, 2001]:

[26] 1. The number of kernels to be used to partition the data is known and set to K.

[27] 2. All kernels used to partition the data have the same (or similar) analytical shape(s). For example, all kernels are Gaussian, yet their mean vectors and covariance matrices can differ from cluster to cluster.

[28] 3. The resulting partition is a distribution which consists in the sum of all kernels, so that the decomposition is linear.

[29] It is rather easy to relax assumptions 1 and 2 while assumption 3 is more crucial. As we plan to determine the position and size of clusters, we shall assume that the mean vector $\vec{\mu}$ and covariance matrix Σ of each kernel exist, and we shall refer to a given kernel as $F(\vec{x}|\vec{\mu}, \Sigma)$. More generally, the kernel could be expressed as a function of higher-order statistical moment tensors. Here, F is a generic kernel and is often chosen as a multivariate Gaussian. In section 6.1 we check the degree to which this is a correct assumption for earthquakes and faults.

[30] The EM algorithm assumes that at each location $\vec{x}$, the local spatial probability density of events is given by

$$p(\vec{x}) = \sum_{k=1}^{K} \pi_k F(\vec{x}|\vec{\mu}_k, \Sigma_k), \tag{1}$$

where the *mixing coefficients* π_k are such that

$$0 \leq \pi_k \leq 1, \forall k, \tag{2}$$

together with

$$\sum_{k=1}^{K} \pi_k = 1. \tag{3}$$

The π_k measure the relative weight of each kernel in the mixture. Equations (2) and (3) ensure that

$$\int_{space} p(\vec{x}) d\vec{x} = 1. \tag{4}$$

The next important quantity is the set of *responsibilities* defined as

$$\gamma(k \to n) = \frac{\pi_k F\left(\vec{x}|\vec{\mu}_k, \sum_k\right)}{\sum_{j=1}^{K} \pi_j F\left(\vec{x}|\vec{\mu}_j, \sum_j\right)}. \tag{5}$$

Thus, $\gamma\ (k \to n)$ is the probability that event n is explained by kernel k; $\gamma\ (k \to n)$ thus belongs to [0; 1]. One can then also define

$$N_k = \sum_{n=1}^{N} \gamma(k \to n), \tag{6}$$

as the expected total number of events explained by kernel k, where N is the total number of events. N_k is thus not necessary an integer. We obviously have

$$\sum_{k=1}^{K} N_k = N. \tag{7}$$

The EM algorithm proceeds as follows:

[31] 1. Choose the number of kernels, K.

[32] 2. Initialize the means $\vec{\mu}_k$, covariance matrices Σ_k and mixing coefficients π_k of the K kernels.

[33] 3. Evaluate the responsibilities using the current set of parameter values (see equation (5)).

[34] 4. Reestimate the parameters using the current responsibilities (use equation (6) for N_k):

$$\vec{\mu}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(k \to n) \vec{x}_n, \tag{8}$$

$$\sum\nolimits_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(k \to n) \left(\vec{x}_n - \vec{\mu}_k^{new}\right) \left(\vec{x}_n - \vec{\mu}_k^{new}\right)^T, \tag{9}$$

$$\pi_k^{new} = \frac{N_k}{N}. \tag{10}$$

[35] 5. Evaluate the log likelihood of this configuration:

$$\ln L = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k^{new} F\left(\vec{x}_n | \vec{\mu}_k^{new}, \sum\nolimits_k^{new}\right) \right\} \tag{11}$$

and check for the convergence of the model parameters or of the likelihood. If convergence is not achieved, return to step 3.

[36] Step 3 is called the E (for expectation) step, while step 4 is called the M (for maximization) step. In the case of non-Gaussian kernels, more sufficient statistical estimators can be computed during the maximization step. Whatever the kernel F, it can be shown that each EM steps cycle increases the log likelihood until it reaches a local or global

maximum [see *Bishop*, 2007]. One can hope to find the global maximum when starting an EM procedure with different initializations (step 2). This is a reasonable approach if K is small. The search for the global maximum is a much more difficult problem when K is large, except if the likelihood landscape has only a small number of local maxima.

[37] It is stressed in section 6.2 that the finding of a strong local maximum of the likelihood function does not necessarily imply that the EM algorithm converged to the best solution. Section 5 reviews some geological and seismological arguments that may help us constraining the best choice for the kernel F, which will, perhaps surprisingly, be found to be nothing but a multivariate Gaussian distribution (see section 6.1).

5. Geological and Seismological Constraints on the Shape of Kernel *F*

[38] We aim at extracting features labeled as "faults" from the spatial structure of an earthquake catalog. To perform this task, the EM algorithm first requires a plausible seismicity model of a fault, i.e., a mathematical kernel that describes the spatial probability density function of events triggered by a single fault. The simplicity of the EM approach mainly relies on the existence of a generic kernel F which, when properly rescaled, describes as accurately as possible what a fault is. From now on, our first assumption is that such a kernel exists. Yet, there is no reason for this kernel to be unique, as there may be different families of kernels able to describe different families of faults.

[39] The vast majority of pattern recognition studies which use an EM approach consider very simple and highly symmetric kernels such as the multivariate Gaussian, without bothering to justify such a choice other than using the argument of computational simplicity. In our opinion, the definition of this kernel should derive instead from the prior empirical or physical knowledge which is available concerning earthquake clustering on faults, as well as from their geometry or dynamics. We thus consider below the general properties that F should possess for the specific application of data clustering to the determination of faults from earthquake catalogs. These considerations will take account of the numerous, and sometimes contradicting, empirical estimations of the statistical properties of earthquake and fault catalogs. The final determination of F is presented and justified in section 6.1. It will respond to the fact that both knowledge and absence of knowledge have to be considered when defining F, thus requiring to couple the observations compiled in the present section with information theoretical arguments.

5.1. Present Knowledge About Kernel *F*

[40] Notwithstanding the large amount of work and significant progress obtained on the structural and physical description of earthquakes and faults, there is no consensus on such a kernel since no theoretical model is successful in predicting the mechanical behavior of the brittle crust over the many different timescales necessary to take into account when going from earthquakes to faults. There are several noteworthy attempts to develop numerical simulations that model the progressive development and organization of fault networks by the cumulative effect of earthquake activity [see, e.g., *Miltenberger et al.*, 1993; *Sornette et al.*, 1994, 1995; *Lyakhovsky et al.*, 2001], but none of them has been seriously validated on natural data. We thus have to examine how empirical observations can help better constrain and define the kernel F.

[41] The ideal situation would consist in observing a single and isolated fault and record its seismic activity over a long period of time. We could then provide a fit to the observed spatial distribution of events, look at its dependency on various parameters such as the applied stress field, estimate the influence of major lithological interfaces such as the free surface or the Conrad or Moho discontinuities on the shape of the resulting cluster, and so on.

[42] Unfortunately, such direct observation is an illusion, as any tectonic domain features a large collection of active faults that mechanically interact, connect, or even cross each other. Fault networks (as observed by field geologists) can then feature very complex structures with variable strikes and dips, while those quantities may also vary spatially within the same fault zone. There is thus no hope to use the simple one-body approach.

[43] In sections 5.2–5.5, we first focus on the representation of faults by seismologists and structural geologists, as well as on the probable influence of the timescale of observations on these definitions. We shall then examine the short-term and long-term statistical properties of individual faults and ruptures. The case of fault networks and earthquake catalogs as a whole will also be briefly reviewed.

5.2. Influence of the Timescale of Observations

[44] The nature of the object defined as a fault may change with the timescale of observation. Consider first an earthquake as viewed by a seismologist. The typical timescale of a detectable event (say, with magnitude $M \geq 1$) ranges from $\simeq 0.01$ s to $\simeq 5$ min and corresponds to the duration of the rupture propagation. The geometry of the rupture process associated with such an event is generally considered as a perfect plane or a slightly curved surface, defined as the boundary between a footwall and a hanging wall (see Figure 2a for the case of an earthquake with thrust focal mechanism). Segmentation features along the earthquake rupture are sometimes mapped but their reliability is questionable, as they may reveal interaction processes with the free surface, and may therefore be not representative of rupture at seismogenic depths. At the shortest timescales of an earthquake rupture, a fault thus seems to be a rather simple object whose identification and quantification is made rather straightforward by direct observation and/or sophisticated geophysical techniques. Figure 2a also shows the location of hypothetical events located on such a fault. They should thus define perfectly planar clusters, so that F should look like a planar segment, corresponding to a flat kernel. Potential irregularities of the fault plane geometry as well as location errors would constrain F to have a nonzero yet small thickness. If the fault was curved or clearly segmented, we could approximate it using several planar segments. According to a seismologist, a fault would thus be close to a completely deterministic Euclidean structure which shape could be analytically determined. This view is in favor of a hard assignment methodology of clustering, yet one would have to take location errors into account. This can be done approximately as by *Ouillon et al.* [2008],

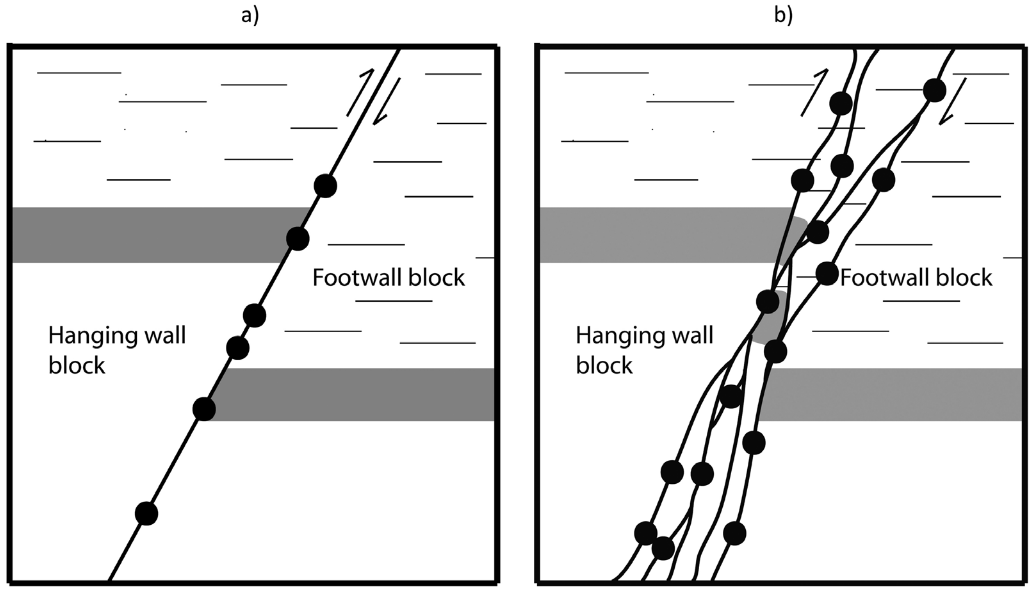

Figure 2. (a) A fault as usually modeled by seismologists (dark oblique straight line), together with its potential seismicity (black circles). (b) Schematic view of a fault as observed by a structural geologist (dark oblique lines), together with its potential seismicity (black circles).

where it was assumed that the thickness of a cluster has to be smaller than the average earthquake location errors, or more elegantly, as in the work by *Wang et al.* [2009], by using the expected square distance. The expected square distance is computed by using all distances from a plane to an earthquake location weighted by their full probability density function (pdf). In contrast, the Euclidean distance used by *Ouillon et al.* [2008] takes account only of the location of the hypocenter (i.e., the most probable location).

[45] Consider now a fault as viewed by a structural geologist. In this case, the relevant timescales typically range between $\simeq 10^5$ years and $\simeq 100$ Myr. The rupture geometry appears as a more diffuse damage zone (with maximum intensity at its core), as very schematically depicted in Figure 2b. In this case, no analytical description of its geometry is known, as many small-scale structures (often qualified as subfaults) define a complex set of numerous interwoven and branching planes which are impossible to identify individually. The only solution is thus to consider this collection of subplanes collectively, so that one has to deal with the fault zone as a whole. Such a complexity favors a stochastic description because, along the fault, the deterministic geometry varies, yet the fault's statistical properties (such as fracture density) may be conserved. Figure 2b also shows a set of hypothetical hypocenters occurring on such a structure. This set does not define a plane but a rather fuzzy structure which has to be considered collectively as well. Some of the events may nucleate on different subfaults, but they may partly share their complete rupture paths. The rupture geometry of a single event may be still simply approximated by a smooth surface, yet all possible rupture surfaces within a fault zone cannot be isolated individually and have to be collectively considered as defining the same entity. For the same reasons as above, this set of events has to be described using statistical tools, and this last example may help to better understand the need for using a kernel to describe the seismicity of such a tectonic structure. This suggests that for intrinsic reasons, the kernel F cannot be flat, and its thickness results from both the fuzzy structure of the fault zone and the earthquakes location uncertainties. Note that this model is compatible with the wide distribution of focal mechanisms of earthquakes thought to occur on the same fault [*Kagan*, 1992].

[46] Recent observations within fault zones provide a slightly different picture (see *Ben-Zion and Sammis* [2003] for a review). It seems that major fault zones indeed appear as a diffuse damage zone (typically of a few hundreds of meters wide) within a more intact rock matrix, but the majority of slip across the fault seem to be accommodated within a very thin and finely crushed gouge zone called the principal fracture surface. Due to the mechanics of rock fragmentation within fault zones as well as indicated by field observations, this principal fracture surface develops rather early within the damage zone which then becomes mainly inactive, as most events nucleate and propagate along the principal fracture surface. This localization of slip is favored by strain weakening mechanisms, so that, on the long term, major faults can be considered as purely Euclidean surfaces. This view is in agreement with works on strike-slip fault segmentation [see *de Joussineau and Aydin*, 2009] which show that the density of fault steps decreases as total displacement offset increases across the fault. No rigorous result is available about the structure of the active part of fault zones at depth, if indeed some activity occurs there. *Robertson et al.* [1995] computed a fractal dimension in the range D = 1.8–2.0 for relocated sets of hypocenters of aftershock sequences southern California. Unfortunately, the observed self-similar scaling holds for scales beyond 1 km, i.e., larger than the typical width of the damage zone. We shall thus assume that in the long-term view of faulting, earthquakes may occur within the damage zone surrounding the principal fracture surface.

[47] As typical earthquake catalogs span only a few decades, it would seem that the seismologist's viewpoint should be the most relevant to constrain the shape of the kernel F. However, taking into account the empirical evidence and theoretical arguments that the growth of a fault is a self-similar or self-affine process in space as well as in time [*Scholz et al.*, 1993; *Scholz*, 2002], we are led to recognize that the timescales involved in the statistical analyses of instrumental catalogs stand halfway between the seismologist's and the geologist's viewpoints. It follows that the kernel F should thus feature characteristics that are common to both descriptions, and be as general as possible for all other properties. For example, F should define an anisotropic kernel that allows one to identify clusters with different lengths, widths, strikes and dips. But their thicknesses (defined as the size of the cluster perpendicularly to its best fitting plane) should not be constrained to be zero so that we leave the possibility to detect faults as seismologists or structural geologists model them.

[48] We face another problem along the time dimension, as faults grow with time (and earthquakes are thought to be part of this process), so that a given mathematical kernel may be valid to describe the seismicity triggered by a given fault only for a limited span of time. Thus, it may be possible that each fault would be described by a different kernel because they are at different stages of maturity. In order to take account of that possibility, the EM algorithm should be modified so that the mixture consists of the superimposition of kernels possibly belonging to different families (Gaussian, exponential, power law, etc.). This would increase drastically the dimension of the solution space, hence the complexity of the landscape in which to search for the maximum of the likelihood function. We thus discard that option in the present work and shall now review the basic quantitative geometrical properties of individual geological ruptures at different timescales.

5.3. Direct Observations of Earthquake and Damage Distribution Around a Fault

[49] Statistical studies of natural seismicity on and around a fault, and of the structure of damage within and around fault zones, are relatively scarce. Most of them focus on brittle deformation features perpendicular to the observed fault planes, i.e., in the direction that corresponds to the thickness of the clusters we intend to identify.

5.3.1. Distribution of Damage Across Faults

[50] *Vermilye and Scholz* [1998] studied the density of microcracks as a function of the distance from the fault for two strike-slip faults of different sizes in the same locality and rock type. They concluded that the crack density at the fault core is independent of the size of the fault and decays exponentially away from it. It seems that this claim comes from a misinterpretation of their Figures 6 and 12 (a confusion of the two axes), as they show unambiguously that the decay is in fact logarithmic, not exponential. Extrapolating this dependence implies that the density of microfractures would drop to zero or to a background level at a finite distance from the fault, defined as the typical process zone width. This typical process zone width was found to be linearly related to the length of the fault (using additional data of different types). More recent data from the Caleta Coloso fault in Chile published by *Faulkner et al.* [2006] provide a different result, according to which the density of microcracks is found to decay exponentially with distance from the fault core. Other data (compiled in Figure 3.14 of *Scholz* [2002]) suggest that the gouge thickness within a fault scales with its length.

[51] Those observations suggest that if the seismicity rate depends on the local density of damage, the kernel F should decay in the direction normal to the fault with tails that are thin enough so that all its statistical moments remain finite. This result allows us to choose a kernel which may be a function of the covariance matrix. The typical thickness of a cluster could be taken to depend linearly on its length, but the proportionality constant is not known.

5.3.2. Distribution of Earthquakes Across Faults

[52] *Liu et al.* [2003] showed that the spatial dispersion of aftershock epicenters normal to the direction of the Landers main rupture could not be explained by location errors alone, and that the event density decays exponentially with distance away from the fault [see *Liu et al.*, 2003, Figure 6]. In Liu et al.'s Figure 6, we can also notice that the decay on both sides of the fault is approximately symmetric. These data must be considered with caution as they concern aftershocks that may occur on distinct faults, and not on the main fault zone (yet the main fault and the distinct faults may constitute a fault zone). They also reveal the dynamics of the fault pattern on rather short timescales. In any case, this suggests that all statistical moments of the earthquake distribution in the direction normal to a fault are finite (see section 5.3.1).

[53] In a recent work, *Hauksson* [2010] performed a similar analysis of the seismicity occurring close (and in the direction normal) to major fault zones in southern California. This study shows a varied behavior depending on how data are selected. The first case deals with the full set of events in the catalog that occur close to a given fault. Earthquake density shows a peak that sometimes coincides with the core of the fault zone, or is sometimes clearly offset on one of its side (note that the earthquake location relative to the fault takes account of the dip of the fault). The average shape of the decay of the earthquake density with distance from the fault displays a zone with a constant density level within 2 km of the fault core, then decays as a power law at larger distances with an exponent $\simeq$1.5 to 2. When selecting only events that are aftershocks of a larger event (such as Landers), the exponent is significantly larger ($\simeq$3), thus suggesting a more localized kernel for short timescales. The density decays over distances larger than 10 km, thus defining a wide *seismic damage zone*. Such a large scale suggests that events are likely to occur on different individual faults, a conclusion supported by the fit of the Landers aftershock sequence by *Ouillon et al.* [2008], who showed that 16 planes are necessary to explain the data at the scale defined by average location errors. The data of *Hauksson* [2010] thus cannot be used to constrain F directly. His results suggest, however, that the statistical moment of second order is finite and may be reliably estimated for the distribution of events in the direction normal to the clusters. This feature will be used in section 6.1.

[54] A more extensive study is proposed by *Powers and Jordan* [2010], who study seismicity patterns across various vertical faults in California. Using stacks of data belonging to different fault classes, they are able to quantify average distribution properties at much finer scales than

does *Hauksson* [2010]. They find that a power law scaling holds up to a fault-normal distance of 3–6 km, with an exponent lower than 2 (except in Parkfield). This scaling holds only over an order of magnitude in distance, as they also evidence an inner characteristic scale d up to which the density of events is constant. The average value of d is 60 ± 20 m in northern California, 220 ± 40 m in southern California, and 120 ± 30 m in Parkfield and is argued to provide a measure of the size of a volumetric damage zone.

5.3.3. Distribution of Earthquakes and Damage Along Faults

[55] The structure of fault zones is very complex along their strike [*Martel et al.*, 1988; *Martel*, 1990; *Willemse et al.*, 1997]. This complexity mainly stems from the interactions among subfaults which grew and finally connected to compose a larger fault. Fault zones thus display bends, pull-aparts and step overs at every scale, which appear as a segmentation process. This heterogeneity of the fault zone thus ensures that earthquakes do not occur uniformly over a simple plane. We are unable to describe this heterogeneity so that we cannot put any constraint on F in that direction. Note that the disorder due to fault linking process certainly also holds in the vertical direction [*Nemser and Cowan*, 2009], hence contributing to the complexity of the distribution along the fault dip as well. Plotting histograms of the number of events as a function of depth for individual faults or for entire regions, *Scholz* [2002] showed that the organization of seismicity is quite complex and that the observed distribution sometimes displays multiple modes. Indeed, earthquake occurrence is governed by several mechanical factors that are influenced by the frictional and rupture properties of rocks, which are all depth dependent: lithostatic pressure, fluid pressure, as well as temperature (hence chemical processes such as stress corrosion) are known to increase with depth. As these parameters also influence the bulk rheology of rocks, deformation mechanisms depend on depth too. The existence of horizontal rheological boundaries as well as of the free surface also favor a symmetry breaking of the seismicity patterns along the vertical direction. Yet, there is no clear model that is able to predict the shape of the earthquake distribution as a function of depth (note that a maximum depth may anyway be defined). Our choice for F should thus not be too much constrained along the vertical direction. Note that the finite thickness of the seismogenic crust ensures a finite (if not well defined) variance of the set of associated events along the dip of the fault. We have no proof that a well-defined and finite variance also holds along the strike of the fault. For long faults, we may even need a kernel F with spatially variable parameters.

5.3.4. Implications for F

[56] The observation of damage around and along faults both on the long and short timescales shows that it is reasonable to assume that the distribution of seismic events within an individual cluster possesses finite second-order moments, i.e., a variance, in any direction. It thus makes sense to consider a kernel F which depends on the covariance matrix of the earthquakes' locations.

5.4. Dynamics of a Single Fault

[57] A few other field observations may help us to better define the shape of F. They mainly deal with the observed cumulative displacement profile along a fault (which results from the successive occurrence of earthquakes on that fault since its formation), as well as with the coseismic displacement over a rupture plane during a single seismic event. These observations are important as displacement gradients control the stress field, hence the probable location of other earthquakes. *Scholz* [2002] showed that cumulative displacement profiles along faults were self-similar, i.e., that they were almost identical when normalized by the fault length. Such a scaling law has also been documented for individual earthquakes, as coseismic displacement is observed to be proportional to the size of the rupture [see also *Scholz*, 2002]. As the overall shape of the displacement field does not depend much on the scale of observation, this suggests that the kernel F should be independent of the fault size. Thus, there is no need for different kernels to describe distinct faults. However, as the proportionality coefficient relating size and displacement differs between the single earthquake (where it is in the range 10^{-4} to 10^{-3}) and the finite fault case (where it is in the range 10^{-2} to 10^{-1}), the aspect ratios of F should not be considered to be constant from one fault to another.

[58] In a recent work, *Manighetti et al.* [2005] noted that cumulative (incremental) displacement profiles along faults (during an earthquake) display a strong spatial asymmetry that was not taken into account in the previously mentioned scaling laws. They also noted that this asymmetry is correlated with the presence of heterogeneities within the Earth's crust, and that displacement maxima occur close to those heterogeneities, then considered as potential nucleation locii of rupture. *Ben-Zion and Andrews* [1998] observed such an asymmetry when modeling the propagation of a rupture along an interface between materials with different mechanical properties (which could be the damage zone on one side and the more intact host rock on the other side in the case of natural faults). This asymmetry of displacement profiles suggests that the seismic activity triggered by faults should also be asymmetric (yet we have no prior idea of how the former asymmetry should translate into the latter), so that the mode and barycenter of F should not coincide. In section 6.1 we check whether it is possible to choose F in a rigorous manner so that it displays this property.

5.5. Geometrical Self-Similarity

[59] Many studies document the geometrical self-similarity of faults and earthquakes catalogs, which made concepts such as fractality or multifractality very popular in Earth Sciences and tectonics. Spatial self-similarity relies on two different kinds of statistics: the full size distribution of single objects (like fault or earthquake rupture length) and the scale dependence of the statistical moments of the spatial density distribution of objects (quantified by a set of generalized fractal dimensions or a multifractal spectrum).

[60] The study of faults and of earthquake ruptures systematically shows that they are not single-sized objects, but that their lengths are distributed over a very broad interval. Different types of size distributions have been proposed, but it seems that a majority of studies agree with a power law. This means that the rupture length probability density distribution scales as $P(L) \simeq L^{-a}$, with most observations suggesting that a is universally close to 2 for faults [*Sornette and Davy*, 1991] and to 3 for earthquakes. This implies that

the generic kernel F should be scalable so that it can fit with clusters of different sizes (note that the EM algorithm could perfectly be defined without those degrees of freedom, considering a kernel F that is independent on the covariance matrix, for example).

[61] The distribution of earthquakes and faults is not uniform, but features strong, intermittent bursts of spatial density [see, e.g., *Ouillon and Sornette*, 1996]. This complexity can be generally summarized in a well-defined multifractal spectrum $(\alpha, f(\alpha))$. This notation means that any box centered on an object (a fault or an earthquake) features a cumulated mass of similar objects that scales as a power of the box size, with exponent α. This exponent, coined *singularity strength*, may vary in space so that set of points with the same value of α is itself distributed in space with a well-defined fractal dimension $f(\alpha)$. Indeed, a multifractal distribution can be seen as a complex mixture of self-similar subsets of data points with similar singularity strengths. However, this mixture is much more subtle and difficult to model than the one defined in section 4, especially because it is fundamentally nonlinear. The multifractal formalism implies that the local exponent α varies in space in a very complex manner, so that the shape of the kernel F should also be able to display this kind of spatial variability.

[62] When dealing with faulting, multifractality quantifies the way small faults aggregate into clusters that behave as larger faults, themselves grouping into larger clusters defining some other faults, and so on [*Ouillon et al.*, 1995, 1996]. Such clusters exist at all scales, which makes more complex the identification of individual faults. The same holds for the clustering of earthquakes. Many studies suggested that within a fault network, the average α value (corresponding to the so-called correlation dimension) should be close to 1.7 on 2-D maps of fault traces. *Ouillon et al.* [1995, 1996], considering the network of fault traces observed in Saudi Arabia, and taking account of several biases that can alter the multifractal spectrum [*Ouillon and Sornette*, 1996], suggested an average value closer to 1.8, so that a fractal dimension of 2.8 may hold in 3-D,if one assumes isotropy (which is certainly a strong assumption). In the case of earthquakes, the correlation dimension seems to be close to 2.2 [*Kagan and Knopoff*, 1980].

[63] Despite the fact that the exponents measured on fault and earthquake catalogs may vary from region to region, it seems that their difference is significant, so that none of them can be reliably used to constrain the shape of the kernel F. Anyway, the above scaling laws would suggest that F should exhibit power law tails, with exponent values possibly informed by the multifractal spectrum properties mentioned above. The problem with this approach is that multifractality refers to a global property, in general isotropic, to which all events contribute even if they belong to different faults. The knowledge of the exponents quantifying the singularity strengths of the multifractal spectrum helps only if the underlying distribution is invariant under rotation, but tectonic deformations clearly display privileged directions. Thus, we should use a kernel behaving as a power law with an exponent that possibly varies with azimuth. Such a complex kernel clearly stands out of the scope of this study.

[64] In addition, power law kernels have the inconvenience of requiring large-scale or small-scale cutoffs (depending on the local value of α), to remove the pathological behavior associated with the mathematical singularity of the power law which impacts on the calculation of statistical moments such as the mean or the variance. These quantities are necessary for estimating the position and size of a cluster, and ultimately control their computed values. The values of those cutoffs, and their possible origin, are still largely unknown [*Sornette and Werner*, 2005a].

[65] Maybe the most important feature of the multifractal formalism is that it also implies that whatever the scale, each event is the center of a self-similar cluster, so that, strictly speaking, one should set $K = N$ in the EM algorithm. In a sense, this is the approach that is used when fitting earthquake sequences using cascade models of seismicity such as the ETAS model. In cascade models of seismicity, each event is the center of a component of the cascade in time as well as in space. Unfolding the cascade can be performed by using model-dependent approaches (such as that of *Zhuang et al.* [2002, 2004], who use the ETAS model) or model-independent approaches (such as that of *Marsan and Lengliné* [2008]). The former assumes an analytical dependence of part of the kernel parameters on the magnitude of the source event, so that it deals with a marked process. The latter does not assume any analytical form of the kernel, yet it has to be the same for all events, or to depend on their magnitude. In addition, these models assume linear superposition of the kernel contribution of each event to the triggering intensity. In effect, both methods separate all events in two broad families: a set of independent (or background) events, and a set of triggered events ("aftershocks" in the standard language; see *Helmstetter et al.* [2003] for a detailed discussion of how to understand the concept of aftershocks in this context). Empirical evidence shows that the share of independent events decreases as the magnitude threshold of the considered catalog decreases [*Sornette and Werner*, 2005b]. Thus, the set of independent events that are determined from such studies cannot be seriously considered as the seeds around which clusters are formed, i.e., as the potential centers of clusters. Indeed, as all events have to be considered as the center of a cluster, it does not help at all for a data clustering approach. ETAS-like models and their derived techniques search for probabilistic relationships of causality between events, which are not adapted for the goal of defining faults.

[66] These arguments thus suggest that it is not obvious how to use the quantitative knowledge concerning the self-similarity of earthquake distributions in order to constrain the kernel F that is needed to develop an efficient clustering algorithm. The problem is that the multifractal description does not naturally fit with the concept of data clustering with the goal of identifying intermediate-scale structures (faults), essentially because the multifractal formalism is based on only two relevant characteristic scales: the whole catalog and the individual event.

6. Gaussian Mixture Approach

6.1. Choice of the Kernel

[67] In section 5, we examine in detail how the theoretical and observed geometry of earthquake clusters that emerge over various time and space scales could inform us on the shape of the kernel F needed for the cluster analysis.

Unfortunately, this synthesis of the available knowledge suggests that a constructive approach to derive physically a plausible kernel may not yet be the most efficient one. In this section, we show that our lack of knowledge can actually be used to deduce unambiguously a reasonable analytical form of kernel F. The key idea is to quantify the impact of the lack of knowledge combined with what we know using the framework of information theory. As more and better data are incorporated in improved statistical and physical models, the proposed approach of combining these models with information theory provides a general framework to improve on the kernel that we propose now. Therefore, the choice of the kernel proposed here should be considered only as a first approximation step, which is open to significant improvement in the future.

[68] In order to identify earthquake clusters, we need to organize their description. Given a spatial distribution of events assumed to belong to the same cluster, it is natural to characterize it by the hierarchy of its moments. The first one is the first-order moment, which defines the mean position noted $\vec{\mu}$ as it is a 3-D vector. The mean position is a simpler quantity to compute than the mode of the distribution of event positions, especially if the distribution is multimodal within a cluster, a possibility that we cannot reject at present. The second quantity is the second-order centered moment of the distribution of event positions, otherwise known at the covariance matrix Σ of the locations of the events. The tensor Σ describes both the size and orientation of the cluster under study.

[69] Higher-order moments of the distribution constrain further the distribution of events within a cluster. For example, in 1-D, the third and fourth-order moments are related to the skewness and kurtosis of the distribution, which provide characteristic measures of the distribution asymmetry and distance to the Gaussian law. In the 3-D case, the corresponding quantities are the third- and fourth-order moments tensors, which are much more delicate to estimate reliably from empirical data. For the sake of simplicity and robustness, as measures of a given cluster, we will only use (1) its mean position $\vec{\mu}$ and (2) its covariance matrix Σ. This implies that the distribution of events within a cluster falls off sufficiently rapidly (faster than a power law with exponent 3) at large distances, such that the components of the covariance matrix are finite and well defined. This hypothesis is compatible with many of the various empirical laws of earthquakes and damage spatial density around faults mentioned in section 5.

[70] In information theory, the lack of knowledge about a stochastic process is quantified by the entropy of the underlying distribution. In the discrete case, if event i has a probability P_i to occur (or if its relative frequency weight is P_i), the entropy of the distribution P is defined as

$$H[P] = -\sum_i P_i ln P_i. \tag{12}$$

A very good knowledge about the process means that P is sharply peaked around a single value, and $H[P]$ will be close to 0. In contrast, zero knowledge about the underlying process (just like playing dice with equivalent probabilities for all possible outcomes) corresponds to the situation in which all events i have the same probability to occur. The entropy $H[P]$ then reaches its maximum value, which is $\ln N$ if the N different events i are equally possible. This formalism can be extended to the continuous case with the definition of the entropy of a spatial stochastic process with density P as

$$H[P] = -\int_{\text{space}} P(\vec{x}) \ln P(\vec{x}) d\vec{x}. \tag{13}$$

Our problem is to determine the kernel P that best models the shape of a cluster about which we know only (1) its mean position $\vec{\mu}$ and (2) its covariance matrix Σ (i.e., its size). This amounts to finding the function P which maximizes the entropy $H[P]$, in the presence of the constraints on the value of $\vec{\mu}$ and Σ. Indeed, maximizing $H[P]$ means that we look for the kernel P which assumes the minimum information beyond what we can determine from $\vec{\mu}$ and Σ. Maximizing $H[P]$ with fixed $\vec{\mu}$ and Σ is solved in a standard way using the method of Lagrange multipliers. The unique solution for P is the multivariate Gaussian distribution with mean $\vec{\mu}$ and covariance matrix Σ [see *Bishop*, 2007].

[71] This result is fundamental for the remaining of this paper and should thus be understood very clearly. We do not assume that all earthquake clusters are organized according to a multivariate Gaussian distribution. We just use the fact that the Gaussian distribution is the most general *prior* distribution (which assumes the least additional information) to describe a cluster, provided only its location, size and orientation are determined.

[72] For graphical purpose, we shall represent the determined multivariate Gaussian kernels using plane segments. To each Gaussian kernel mean $\vec{\mu}$ and covariance matrix Σ, we will show a plane centered at the position of the mean $\vec{\mu}$ of the Gaussian, while its orientation will be given by the principal axes of its covariance matrix Σ corresponding to the two largest eigenvalues. Those two largest eigenvalues will also determine the size of the plane, following the convention of *Ouillon et al.* [2008] (see section 7).

[73] Section 7 illustrates the Gaussian mixture approach on a synthetic example. It also elaborates a strategy to choose the number of Gaussian kernels used in the cluster decomposition, as it is one of the latent variables of the problem.

6.2. Possible Caveats

[74] Having chosen the multivariate Gaussian kernel for F, the data set studied below will be processed using the algorithm given in section 4.3. Recall that it uses a maximization of the likelihood as a criterion for the convergence toward a solution.

[75] There are some cases for which the validity of this approach can be questioned. Imagine a data set featuring N data points that we try to fit using K Gaussian kernels. Also imagine that within this data set, we can define a small cluster featuring three points that are spatially isolated from the others. Assuming that K is large, and depending on the initialization of the kernels, one may reach a situation where one of the kernels will account only for these three points and will thus literally collapse on them: this means that the covariance matrix of that kernel will depend only on the spatial locations of those points, all other data points having a vanishing influence on it. It thus follows that this

covariance matrix will be singular, as the smallest eigenvalue will be zero, so that the likelihood contribution of this kernel will be infinite. The total log likelihood will thus be infinite too, whatever the solution provided for the rest of the data set. In that case, maximizing the likelihood does not have any meaning. More generally, this problem occurs in a D-dimensional space when the D volume of a cluster is null. It is thus not appropriate for the identification of perfectly planar clusters in 3-D, a situation that is however unlikely in earthquake catalogs. Note that in the 3-D case, a kernel can also collapse on a set of two events, or on a single data point.

[76] Two solutions can be proposed to solve this problem. The first one consists in checking *a posteriori* that none of the kernels correspond to a singular covariance matrix. This can be easily done during the computations, and the corresponding kernel can then be removed from the set of kernels. The other (and best) solution consists in eliminating this possibility directly inside the data set. In that case, we eliminate events that define local geometrical structures that are likely to be characteristics of random catalogs instead of genuine geological features. This procedure is described in section 8.

7. Application of the EM Method With Gaussian Kernels to a Synthetic Example

7.1. Generating the Data Set

[77] The synthetic example we present here is similar to the one we considered by *Ouillon et al.* [2008] for the sake of comparison. Events are located over a set of three vertical faults. Here, λ stands for longitude, ϕ stands for latitude, and z stands for depth. The center of the first fault is located at $\lambda = -0.01°$ and $\phi = 0$. The center of the second fault is located at $\lambda = +0.01°$ and $\phi = 0$. Both faults have their centers at $z = 10$ km, while they are strictly parallel with a zero strike, vertical dip, length of 20 km and width of 10 km. The third fault is normal to the former ones, and its center is located at $\lambda = 0$, $\phi = 0$, $z = 10$ km. Its width is 10 km while its length is 40 km. We locate randomly (using a uniform distribution) some events over each plane. The first two faults carry 100 events each, whereas the third one carries 200 events as it is twice the length. There are thus 400 events in total. A random deviation to each event location is added in the normal direction to the plane it belongs to, mimicking location errors. This deviation is chosen uniformly within [−0.01 km; +0.01 km], so that it is comparable with relative location uncertainties in real relocated catalogs. This data set is shown in Figure 3a. We assume that the user measures the position of each event, which is determined within an uncertainty of ±0.01 km in each spatial direction. The number of faults and their geometrical properties are not known and the goal of the exercise is to demonstrate how good is the EM method with Gaussian kernels to recover these faults and their properties.

7.2. Fitting the Data

[78] Our algorithm uses the following variables and notations. The kernels will be noted G for Gaussian. A given kernel i of a set of K kernels is denoted $G_i^K(\vec{x}|\vec{\mu}_i^K, \Sigma_i^K)$. The index i of the kernels runs from 1 to K. For each kernel, the corresponding covariance matrix elements will be noted $\sigma_{i,mn}^K$, where indices m and n stand for x, y or z. Each covariance matrix can be diagonalized, and its eigenvalues are noted $\sigma_{i,j}^K$, where $j = 1$ for the largest eigenvalue, $j = 2$ for the intermediate eigenvalue and $j = 3$ for the smallest eigenvalue. The corresponding eigenvectors are noted $\vec{v}_{i,j}^K$. Following *Ouillon et al.* [2008], we assume that the size of the fault in any of its eigendirections is $\sqrt{12}$ times the standard deviation of its associated seismicity in that direction.

[79] In a first step, we fit this data set using a single Gaussian distribution; that is, we start with $K = 1$. The data set is first divided into two distinct subsets controlled by a probability p such that $0 < p < 1$. Each event in the whole catalog has a probability $1 - p$ to belong to a first set called the training set, while it has a probability p to belong to the validation set. For this example, we set $p = 0.1$ which means that, on average, about 40 events will belong to the validation set while about 360 events will belong to the training set. For each event, we thus draw a random number between 0 and 1, compare it to p and decide whether the data point belongs to the validation set or to the training set. After selecting those two subsets, we change the spatial location of each event by adding a random perturbation similar to the one which has been used when generating the catalog (in order to simulate location uncertainties; it is thus chosen randomly within [−0.01 km; +0.01 km]). Both subsets can then be viewed as independent realizations of the dynamics of the same underlying fault pattern. We then fit the training set using only one cluster, whose solution allows to compute the likelihood of the obtained configuration, using equation (11). The log likelihood is then normalized by the number of data points in the training set, so that we obtain a mean likelihood per data point (which we hereafter refer to as L_{pd}). Using the same optimal kernel parameters we obtained for the training set, we now compute the L_{pd} of the validation set conditioned to that kernel. In general, this likelihood will be smaller than the former. We are thus left with two L_{pd} conditioned to the same kernel: one for the training set, the other for the validation set. We perform this procedure 10 times, selecting different training and validation sets for a fixed value of p, perturbing randomly their location according to spatial uncertainties, so that we can provide an average value and error bars for the L_{pd} in the case $K = 1$.

[80] We then add a second Gaussian kernel, which amounts to fit the data with two kernels $G_1^2(\vec{x}|\vec{\mu}_1^2, \Sigma_1^2)$ and $G_2^2(\vec{x}|\vec{\mu}_2^2, \Sigma_2^2)$. We call this operation *kernel splitting* as the primary kernel is split into two secondary kernels (the primary kernel is simply the one obtained in the last of the 10 fits for $K = 1$). The initial positions and covariance matrices of the secondary kernels are not chosen at random but depend on the properties of the primary kernel. We choose them such that

$$\vec{\mu}_1^2 = \vec{\mu}_1^1 + \frac{\sqrt{3}}{2}\sqrt{\sigma_{1,1}^1}\,\vec{v}_1^1, \tag{14}$$

$$\vec{\mu}_2^2 = \vec{\mu}_1^1 - \frac{\sqrt{3}}{2}\sqrt{\sigma_{1,1}^1}\,\vec{v}_1^1, \tag{15}$$

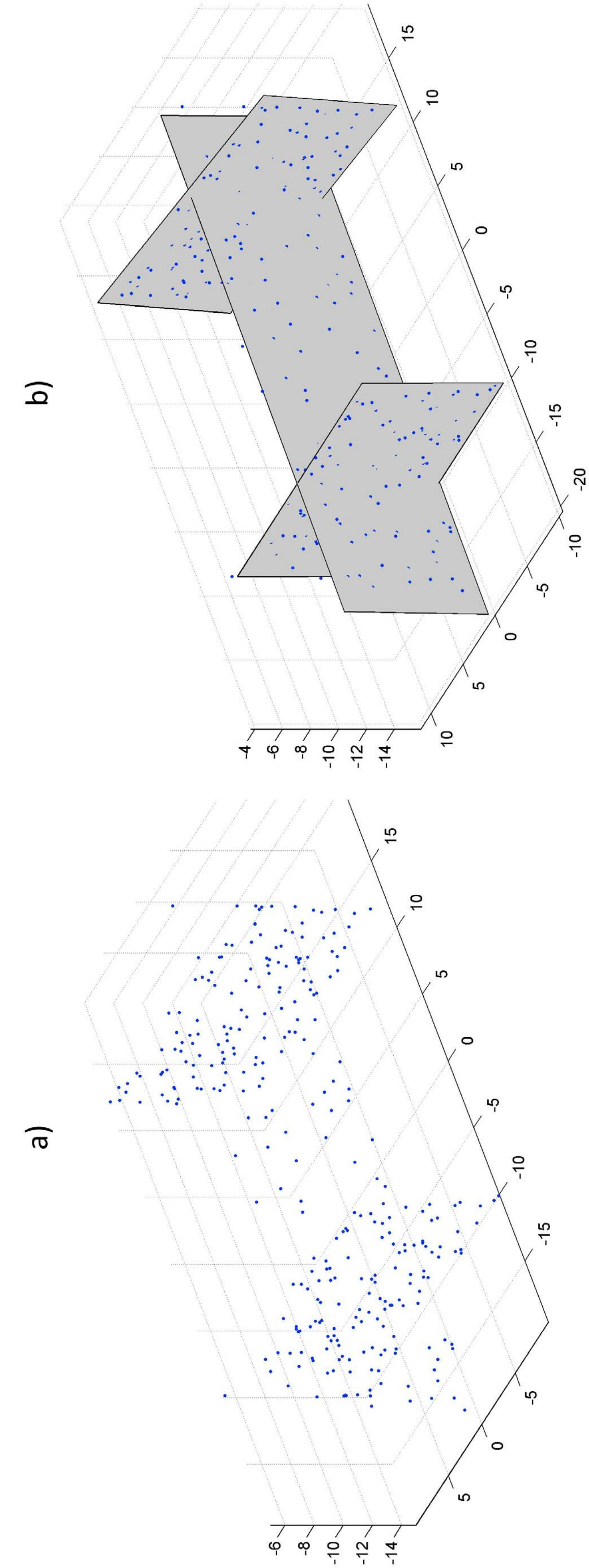

Figure 3. (a) The synthetic example analyzed in section 7. See Table 1 for its imposed geometrical parameters. (b) Data clustering of the synthetic set of events shown in Figure 3a, using three Gaussian kernels. See Table 1 for their properties.

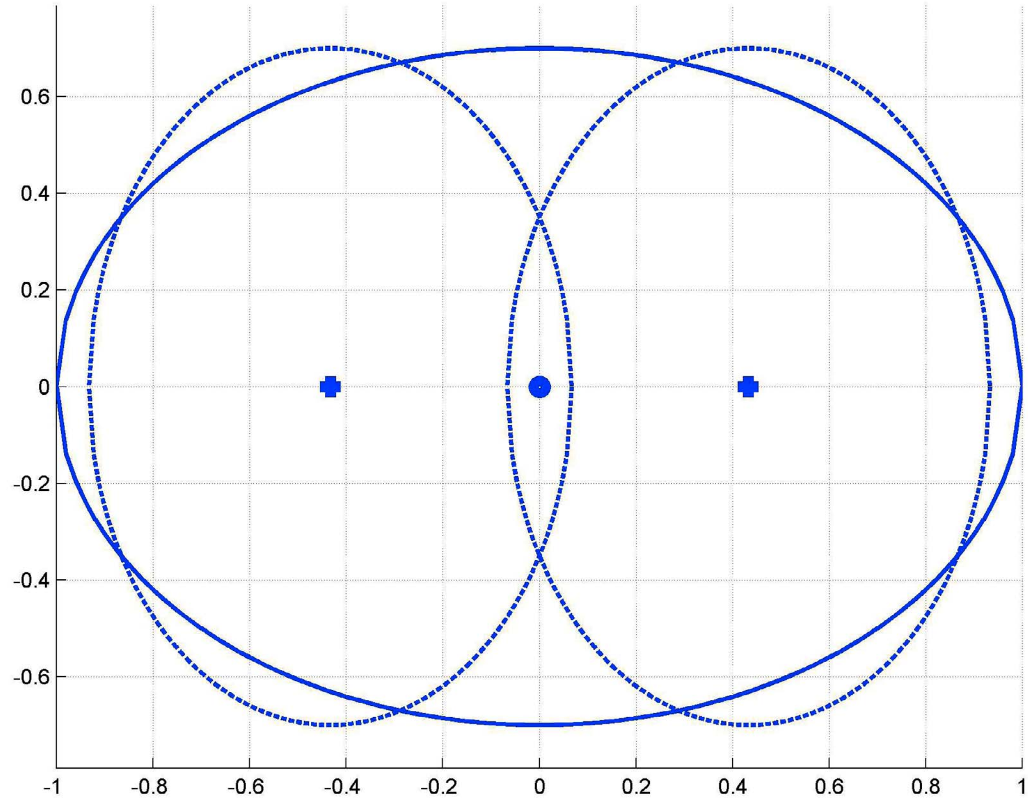

Figure 4. A 2-D multivariate Gaussian kernel, represented by its one standard deviation uncertainty ellipse using a solid line and centered on the solid circle at (0,0), is replaced by two smaller-scale kernels represented by dashed lines and centered on thick crosses. We call this operation *kernel splitting*. See section 7.2 for detailed explanations.

$$\sigma^2_{1,1} = \sigma^2_{2,1} = \frac{1}{4}\sigma^1_{1,1}, \quad (16)$$

$$\sigma^2_{1,2} = \sigma^2_{2,2} = \sigma^1_{1,2}, \quad (17)$$

$$\sigma^2_{1,3} = \sigma^2_{2,3} = \sigma^1_{1,3}. \quad (18)$$

[81] Figure 4 shows the geometrical meaning of this substitution in the 2-D case. In Figure 4, each multivariate Gaussian kernel is represented using its one standard deviation ellipse (which becomes an ellipsoid in 3-D). The primary Gaussian kernel is drawn using a solid line, and its center is represented with a solid circle located at (0;0). In this example, the standard deviation of the primary Gaussian kernel along the long axis is set to 1, while its standard deviation along the short axis is set to 0.7. This primary Gaussian kernel is then replaced by the two secondary kernels drawn using dashed lines, and centered on the thick crosses. For each of these secondary kernels, the standard deviation along the vertical axis of Figure 4 is the same as the one of the primary kernel in that direction, i.e., 0.7. In the horizontal direction of Figure 4, the standard deviation of each secondary kernel is set to half the standard deviation of the primary kernel in that direction, i.e., 0.5. Both secondary kernels are centered at ($\pm\sqrt{3}/2$; 0), respectively. This ensures that the global covariance matrix of the set of secondary kernels is identical to that of the primary kernel. We are thus left with a new configuration which has the same global covariance matrix as the primary one, except that it now features more degrees of freedom to fit the data, and the two secondary Gaussians can now be used as the starting configuration for the EM algorithm (while the primary kernel is removed from the set of kernels). In the 3-D case, our procedure is also to split the primary cluster into two secondary kernels along its largest axis, and to leave untouched the other eigendirections.

[82] We then consider the full synthetic data set, reassign each event either to the training set or to the validation set using the same probability p, and perturb their original spatial location with new random increments. Those two subsets are thus continuously changing while the algorithm proceeds, which allows a better exploration of the influence of all events, as well as of location uncertainties, in the fitting process. The training data set is then fitted using K = 2 kernels, and the L_{pd} is computed for the best solution. The L_{pd} of the validation set conditioned to the same set of kernels is also computed and stored. Once again, we generate 10 different training/validation data sets to provide averages and error bars.

[83] We then split the thickest kernel and start to fit the data with K = 3 kernels, and so on. The thickest kernel is

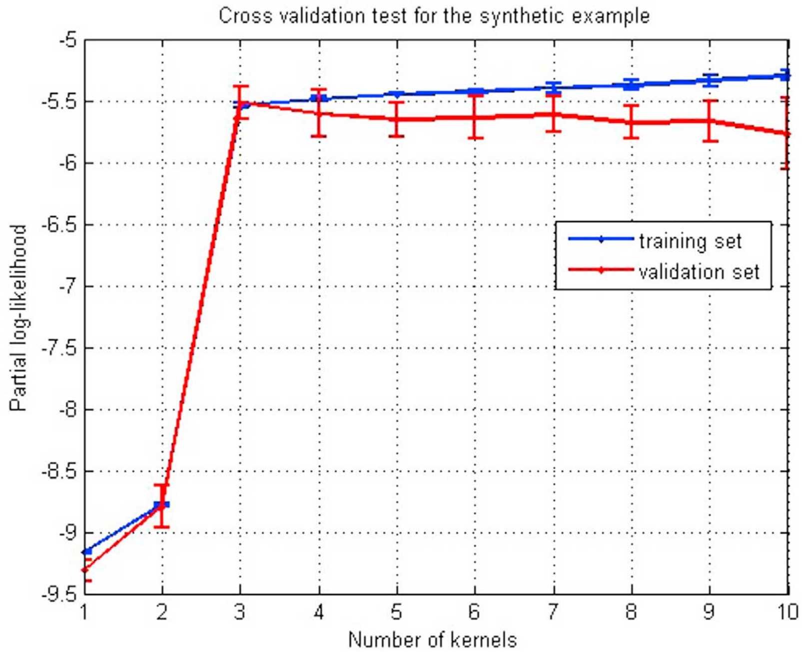

Figure 5. Average values of the likelihood per data point (L_{pd}) and one standard deviation error bars for the training sets (top curve) and the validation sets as a function of the number of kernels when fitting the data set shown on Figure 3a.

defined as the one which features the largest variance across its principal plane. For each value of K, we select a training data set and a validation data set, perturb the events spatial location, and compute the L_{pd} for both data sets using the same kernels configuration. Figure 5 shows the L_{pd} for the training and validation data sets versus the number of kernels used in the fit. One can check that the likelihood increases without any bound in the case of the training data set: the more kernels we use to fit the data, the larger the likelihood. However, the increase is much slower for $K > 3$, indicating that increasing K further may not be so efficient. The result is somewhat different for the validation set, which is an independent data set. The L_{pd} of the validation set first increases with K in a manner similar to the training data set: both data sets are similarly explained by the corresponding set of kernels. For values of K larger than 3, one can check that the L_{pd} of the validation set decreases slowly, which means that adding more clusters to fit the training set provides less and less consistency with an independent data set sampling the same underlying faults. This comes from the fact that increasing K allows to fit smaller-scale details of the training set whose geometry is more and more controlled by location uncertainties. This leads us to conclude that one should not use more than $K = 3$ clusters to fit the data set under study, which recovers the true number of faults in this synthetic example.

[84] Fitting the full data set with $K = 3$ kernels, the EM algorithm converges to the pattern shown in Figure 3b. We check that it fits almost perfectly the planes used to simulate the data set. Table 1 shows the parameters of the synthetic simulation as well as the cluster parameters retrieved by this modified EM algorithm.

[85] Of course, this methodology features a tuning parameter p which value is chosen arbitrary (but a value $p = 0.1$ is often assumed in cross-validation schemes); p must not be too large, as the training set must feature as many events as possible to provide a reliable fit of the fault pattern itself (in order not to waste too many data). In the other hand, p must not be too small, as the validation set may then feature too few events which would provide an unreliable value for the L_{pd} of the validation set (we checked that if p is too small, then this L_{pd} may feature very strong fluctuations

Table 1. Correspondence Between the Parameters of the Input Fault Planes Used to Generate the Synthetic Catalog Shown in Figure 3a and Those of the Fitting Kernels Found by the EM Algorithm[a]

Label	Longitude	Latitude	Depth (km)	Strike	Dip	Length (km)	Width (km)	σ_3 (km)	N
A	−0.100	0.000	10.000	180.00	90.00	20.000	10.000	0.006	100
	−0.100	−0.003	10.372	180.03	89.92	19.666	10.056	0.003	100.46
B	+0.100	0.000	10.000	180.00	90.00	20.000	10.000	0.006	100
	+0.100	0.006	9.791	180.05	89.96	21.241	10.014	0.003	99.58
C	0.000	0.000	10.000	270.00	90.00	40.000	10.000	0.006	200
	−0.017	0.000	10.486	270.00	89.85	38.756	10.249	0.003	199.96

[a]Longitudes and latitudes refer to those of the barycenters of the faults (or kernels). For each fault label, the first row corresponds to the input fault, while the second row corresponds to the fitted parameters that have been determined by the algorithm.

with K). The value $p = 0.1$ is a good compromise with a smooth behavior of the L_{pd} as a function of K, together with a nice convergence of the algorithm when fitting the training data set. We will thus use this value $p = 0.1$ in our implementation below. As the validation data set is smaller than the training data set, error bars are larger for the validation data set. Using $p = 0.05$ did not change the results (even for the natural catalog analyzed in section 8), whereas $p = 0.20$ increased the number of cases where some of the kernels collapsed on three or less data points.

8. Application of EM to the Mount Lewis Aftershock Sequence

8.1. The Mount Lewis Aftershock Sequence

[86] The $M_L = 5.7$ Mount Lewis earthquake occurred on 31 March 1986 on a right-lateral fault northeast of and oblique to the Calaveras fault in California [*Zhou et al.*, 1993]. Inversion of its centroid moment tensor suggests a right-lateral slip motion over a rupture plane striking N353° with a 79° dip, and a 9 km deep hypocenter. *Zhou et al.* [1993] studied the aftershock sequence in details, and deduced a main rupture size of typical radius of about 1.5–2.5 km. The seismicity pattern shown in Figure 6 is mainly due to the aftershocks following this event. Their focal mechanisms display strike-slip motion on steep faults, without obvious spatial relationships between them.

[87] The seismicity in that area has been relocated by *Kilb and Rubin* [2002], who used waveform cross-correlation techniques to determine the relative positions of 2747 events. While the relocated events were found to delineate a N-S trend at large scale, a more complex picture emerges at smaller scales. In particular, fine structures could be identified as forming series of E-W near-vertical faults of small dimensions ($\simeq$0.5–1 km long), with separations between them as small as $\simeq$200 m. *Kilb and Rubin* [2002] proposed that these structures result from the growth of a relatively young, right lateral N-S fault, whose displacement is accounted for by slips on secondary left-lateral E-W faults. Near the location of the main shock, the structure is much simpler and close to a single N-S plane. The occurrence of E-W secondary faults is more obvious to the north of the main shock, and the lengths of those faults seem to increase with distance to the main shock. The secondary faults located south of the main shock seem to have more variability in depth and strike, judging from the seismicity pattern.

[88] *Kilb and Rubin* [2002] argued that those E-W faults might reveal a preexisting faulting anisotropy in the area. They also noticed that two E-W clusters located to the north and south, respectively, of the main event were activated a few weeks later, suggesting a delayed growth process of the fault which generated the main shock.

[89] The catalog we used is the one of *Waldhauser and Schaff* [2008] which extends from January 1984 to May 2003. Yet, we only used events occurring after the Mount Lewis event, up to the end of the catalog. Note that this catalog features only events recorded by 6 or more stations. Figure 6 (left) shows that if most of events define clusters that could be easily identified with the naked eye, some of them define much sparser structures (if any structure at all) that preclude the identification of any reliable underlying fault. Before applying our modified EM algorithm in order to identify automatically all significant clusters, we thus need to address the problem of uncorrelated seismicity.

[90] This catalog features the 3-D (relative) locations of events, as well as information about their uncertainties. The epicentral uncertainty is given by the size (in km) of the axes of the 95% error ellipse, together with the azimuth of its large axis. The depth uncertainty is given by the 95% error interval (in km), so that we do not know the full covariance matrix of each event's location. In order to speed up computations, we determine the average of the three spatial standard deviations and assume that location uncertainty is isotropic. This simplified uncertainty sphere (of radius $\simeq$10 m for most events) is then used to build the reshuffled training and validation sets, following the procedure described in section 7, assuming a Gaussian distribution of earthquakes hypocentral location. This simplification has no real influence on the final results as we will check that the size of the clusters is significantly larger than location uncertainties.

8.2. Pretreatment to Sort Out Clustered and Unclustered Seismicity

[91] The synthetic example treated in section 7 is idealized in several respects, one of them being that the events are all assumed to be associated to some fault, possibly different from event to event, and that all existing faults are sampled by a sufficiently large number of events. In real data sets, which necessarily include only a limited history of seismic recording, several events can be isolated or, more generally, not associated with any cluster. This does not mean that they did not occur on a fault but that these particular faults associated with these isolated earthquakes carry too few events (perhaps just one) to reveal their existence via any clustering reconstruction algorithm. One can say that these faults are undersampled by the available seismic catalog.

[92] Given the complexity of faulting at all scales, the existence of these isolated events constitutes a problem in any clustering procedure, in order to keep small the number of reconstructed fault clusters. Not taking into account the existence of this diffuse seismicity would lead in any clustering procedure to a proliferation of spurious faults. Indeed, for a set of isolated events, the clustering algorithm tends to partition that data set into subsets of three events each, as this is the optimal solution for partitioning a more or less isotropic distribution with anisotropic clusters. Thus, in most cases, we would be confronted with the convergence problem raised in section 6.2, as some kernels will collapse on sets of three events or less, and the likelihood would diverge spuriously. The whole exercise would become meaningless and no insight on the relevant underlying fault network would be derived. We thus have first to identify the clustered part of seismicity, remove the uncorrelated seismicity and apply our EM procedure only to the clustered subset in order to identify and characterize individual clusters.

[93] A caveat is that the definition of uncorrelated seismicity is time and space scale dependent. For example, at large spatial scales, all events along the boundary between the Pacific and North American plates appear to be located within narrow and well-identified quasi-linear clusters. No part of the seismicity would be considered uncorrelated. But at smaller scales, as in Figure 6, many events appear quite

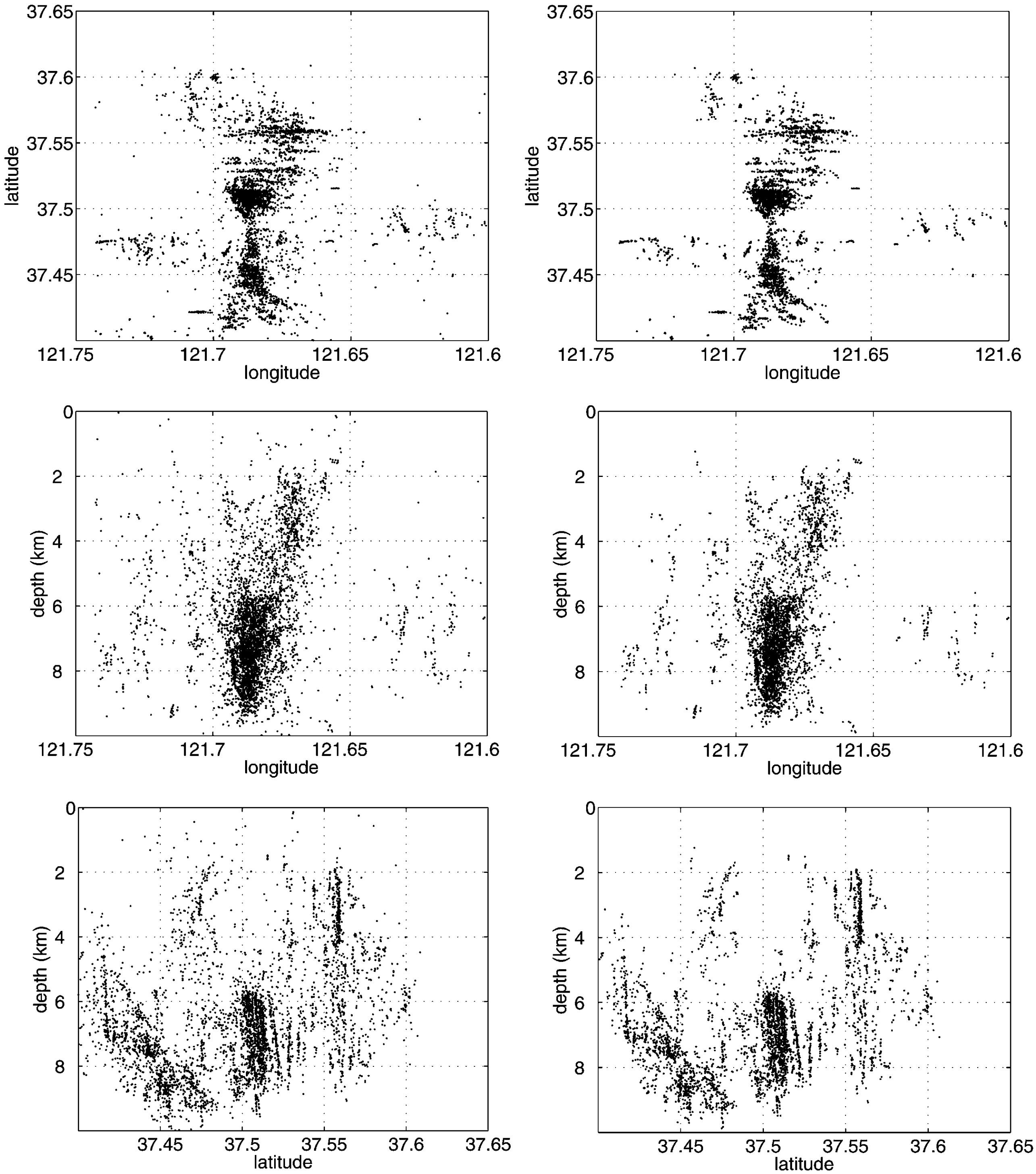

Figure 6. Spatial distribution of the seismic activity that followed the Mount Lewis event, shown along different projections. (left) The whole data set; (right) the clustered part of seismicity (see text for details).

distant from their nearest neighbors and do not define any specific pattern with them. Here, we avoid this issue of the scale dependence by focusing on the regional scale associated with the 1986 Mount Lewis earthquake.

[94] In contrast to the hypothesis underlying many previous studies and models, one should not assume that the uncorrelated seismicity is homogeneous. Following the concept that uncorrelated events belong to the subset of seismic activity occurring on undersampled faults, its spatial organization should correspond to the undersampling of a complex fault network. As an illustration of the distortions that may result from undersampling, *Eneva* [1994] showed that an undersampled monofractal distribution could be described by a spurious nontrivial multifractal spectrum.

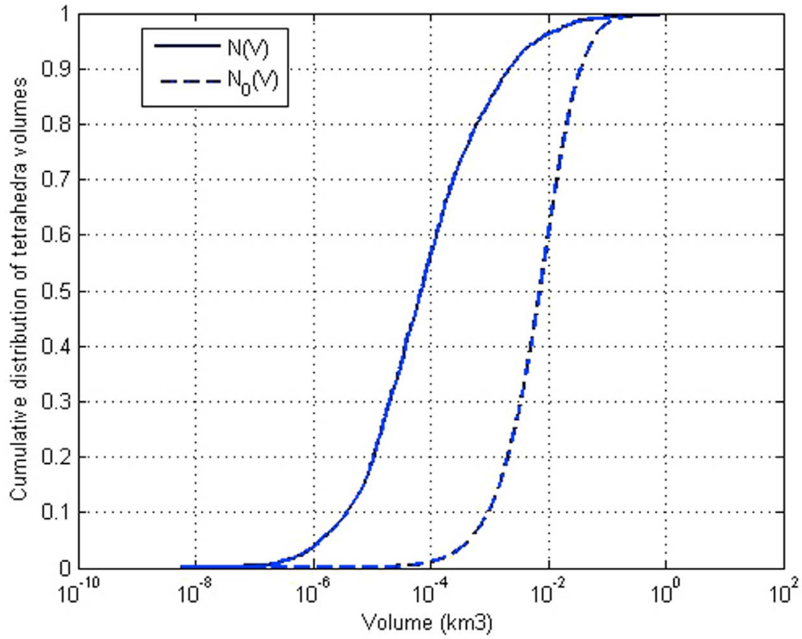

Figure 7. Cumulative distribution $N(V)$ of the volume V of tetrahedra formed with four nearest-neighbor events (see main text for details) for the full catalog (solid line) and the corresponding distribution N_0 (V) for randomized catalogs (dashed line).

[95] We now proceed to identify the uncorrelated seismicity by comparing the distribution of natural events to a reference distribution of uncorrelated seismicity. Our method consists in comparing the cumulative distribution $N(V)$ of the volumes V of tetrahedra formed with quadruplets of nearest neighbor events with distinct locations obtained from the natural catalog (Figure 7, solid curve) with the distribution $N_0(V)$ of the volumes of tetrahedra constructed with events belonging to a randomized catalog, where the horizontal (but not the vertical) dependence between events has been destroyed (Figure 7, dashed curve). We are guided by the logic that, in general, tetrahedra defined by clustered events should have smaller volumes than tetrahedra defined by unclustered seismicity events, because the former are supposed to be associated with local quasi-planar fault structures. The set of uncorrelated events is constructed by throwing at random (in the horizontal plane) the same number of events as in the real catalog, in a spatial domain whose size is identical to the natural one. Each event in the randomized catalog is located at the depth of a randomly chosen event in the real catalog.

[96] Let $V_0(0.05)$ be the 5% quantile of the distribution $N_0(V)$ of tetrahedra volumes formed with events of the randomized catalog, i.e., $N_0(V_0(0.05)) = 0.05$. We define the correlated seismicity as made of those events contributing to tetrahedra in the natural catalog with volumes smaller than $V_0(0.05)$. Quantitatively, we find $V_0(0.05) = 4.511 \times 10^{-4}$ km^3, which means that by definition of the quantile at probability level 5%, only 5% of tetrahedra volumes of the randomized catalog are smaller than $V_0(0.05)$. In contrast, we find $N(V_0(0.05)) = 0.77$, which means that about 77% of the tetrahedra constructed with events belonging to the natural catalog have a volume smaller than $V_0(0.05)$), thus reflecting a much stronger clustering on more planar structures. The difference between the natural and the randomized catalogs is also striking for the largest tetrahedra volume: 2 km^3 for the former compared with 31 km^3 for the latter.

[97] We thus select in the real catalog all events which belong to a tetrahedron of volume $V \leq V_0(0.05)$. By construction, these events have only a 0.05 probability of forming only by chance a tetrahedron with a small volume. All other events are removed and not considered in our analysis, as they are interpreted as belonging to undersampled faults. This set of rejected events is shown in Figure 8. Figure 6 (right) shows three projections of the set of clustered seismicity. These three projections can be compared with those on the left column showing the full initial catalog, sum of both clustered and uncorrelated seismicity. In all three projections, the clustered part of seismicity appears to be much more strongly anisotropic than the full catalog. It is also very clear that the unclustered seismicity shown in Figure 8 does not exhibit clear patterns, in contrast with the clustered part shown in Figure 6 (right). A local criterion using just quadruplets is therefore sufficient to sort the seismicity of the aftershocks of 1986 Mount Lewis earthquake into two classes, which show very distinctive differences in clustering, already to the naked eye.

[98] The tetrahedron is the simplest polyhedron, that allows us to define a volume, taken as a diagnostic of a nonplanar pattern. This pretreatment to sort out clustered and unclustered seismicity in terms of tetrahedra can be generalized to pentahedra and higher-order polyhedra, at the cost of more involved calculations. Furthermore, our choice of the tetrahedron amounts to the minimal selection procedure, in the sense that the constraint is the most local. Using higher-order polyhedra would select higher-order statistical properties. This is not judicious for the goal of defining large-scale fault clusters, which by themselves should

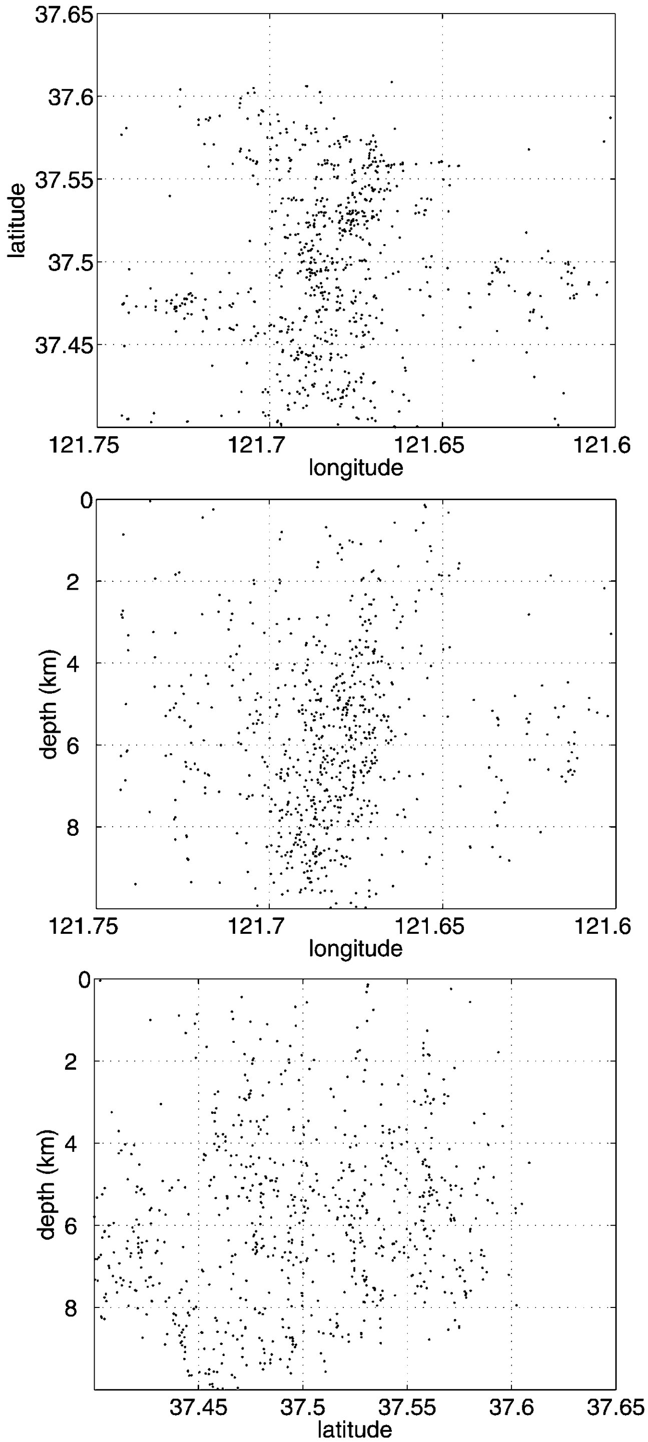

Figure 8. Unclustered events detected in the catalog shown in Figure 6 using a local criterion using quadruplets of events forming tetrahedra compared between the initial and randomized catalogs (see text for details).

embody the information of the important large-scale coherent structures (the faults) of seismicity.

8.3. Spatial Data Clustering of the Mount Lewis Sequence

[99] We now apply our data clustering method to the clustered part of the Mount Lewis sequence shown in Figure 6 (right), as defined in the previous section. We shall first present the results of the cross-validation scheme, which allows to determine the number of kernels $K_{\max}$ that should be used to fit the data, in a way similar to section 7. Then, we shall look at the obtained fault patterns using different values of K, before studying in more details the statistical properties of the clusters obtained for $K = K_{\max}$, i.e., of the optimal set of fault planes proposed to explain the earthquake catalog.

8.3.1. Cross Validation Results

[100] Figure 9 shows the dependence on K of the L_{pd} of the training and validation data sets (using $p = 0.1$). Figure 9 is thus analog to Figure 5. As observed for the synthetic data set in section 7, error bars are much larger for the validation data set (bottom curve) than for the training data set (top curve). The middle dashed curve shows, for comparison, the dependence of L_{pd} when we fit the full data set (so that $p = 0$; there is thus no validation data set in this case). The L_{pd} of the training data set in the case $p = 0.1$ is larger than the L_{pd} when using the full data set because the former features less events, which makes the fit easier. Figure 9 shows that the L_{pd} increase very quickly with K when less than 10 kernels are used for the fit. Then, the likelihoods increase more slowly without exhibiting any bound for the training sets, as increasing K provides better fits. In the case of the validation set, the increase is even slower, and the L_{pd} saturate at a constant value when $K > 100$, suggesting that increasing K further does not bring any significant information about the structure of the data set. However, unlike the synthetic case, the L_{pd} of the validation set does not decrease when K increases beyond a critical value, which makes an accurate measure of $K_{\max}$ much more difficult. Figure 9 displays results obtained for $K = 1$ to 130, suggesting that $K_{\max} \simeq 100$. Fitting the training data set with K values larger than 130 proved very unstable as we found an increasing number of cases where Gaussian kernels collapse on clusters featuring only three or fewer events, preventing the convergence of the algorithm. Note that in the case where such a collapse is found, the corresponding kernel is removed and we restart a fit with $K - 1$ kernels. Computations using $p = 0.05$ or $p = 0.2$ did not change significantly the results due to fluctuations of L_{pd} for the validation data set.

[101] We now switch to the results obtained when fitting the whole data set with a varying number of clusters. For example, Figure 10 shows the sets of fitting planes obtained when increasing K from 1 to 130 (we only show the cases $K = 1$, 3, 5, 10, 30,50, 90, 110, 130). Each plot shows a 2-D epicentral projection of the data set as well as of the corresponding set of planes. The darkest plane is the one that will be split in the next step by going from K to $K + 1$ to form the initial conditions of the EM procedure. Figure 11 (top) shows the solution obtained with $K = 100$ planes, while Figure 11 (bottom) shows a stereographic projection of the corresponding set of planes (Figure 11, bottom left) and of the associated set of poles (Figure 11, bottom right). Figure 11 (bottom left) shows, as seen from above, the intersection of the planes with the lower hemisphere of a unit sphere, assuming that all planes have been translated in space so that their centers coincide with that of the sphere. It thus provides a direct visualization of the anisotropy of the set of fault traces. Figure 11 (bottom right) shows the intersection with the same hemisphere of the set of vectors normal to each

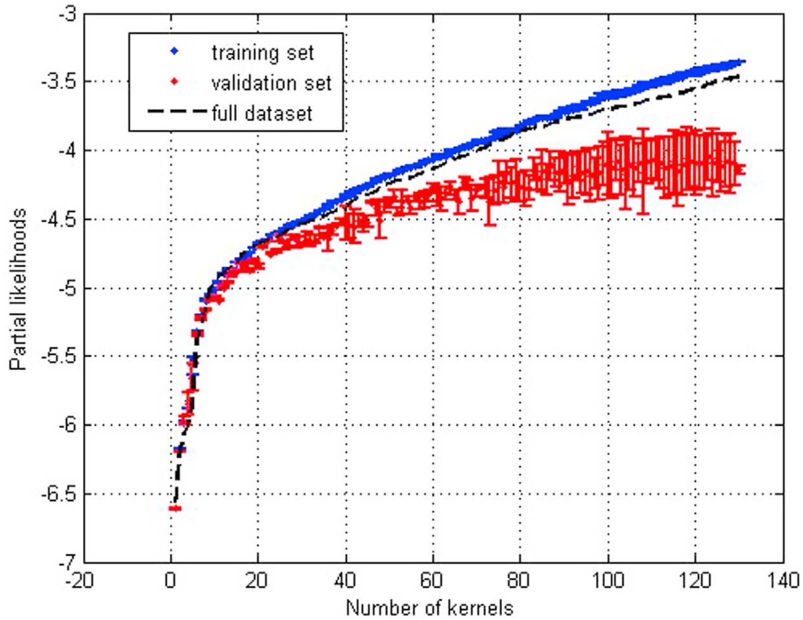

Figure 9. Likelihoods per data point (L_{pd}) of the training data set (top curve) and of the validation data set (bottom curve) for the Mount Lewis catalogue with p = 0.1. The middle dashed curve shows the L_{pd} of the fit of the full catalogue (p = 0).

plane (one extremity of the vector being attached to the center of the unit sphere). Figure 11 (bottom right) is particularly useful when many faults are plotted. Figures 11 (bottom left) and 11 (bottom right) are indeed equivalent and help to visualize quickly the main anisotropy directions if any. For example, in the case of poles, vertical planes correspond to points located on the edge (equator) of Figure 11 (bottom), while horizontal planes correspond to points located at the center of Figure 11 (bottom). For K = 100, both Figures 11 (bottom left) and 11 (bottom right) indicate that a large majority of clusters define nearly vertical structures, as most poles concentrate near the equator, yet their azimuths are highly variable. It also seems that a majority of clusters dip to the west, whereas the focal mechanism of the main rupture indicated a dip to the east (note that the main shock belongs to the set of unclustered events and is thus not part of the fitted data set). Looking at Figure 10, we also notice that most of the newly added planes are then generated in the central part the catalog (within [0; 3] km along the vertical direction), suggesting that this small-scale area features smaller structures. If this is the case, the associated set of events would reveal the fuzzy structure of a fault zone (as depicted on Figure 2b).

[102] The solution obtained for K = 130 displays a very large anomalous cluster (which is about to be split in the next iteration), with a thickness of several kilometers. This kind of solution does not show up systematically for very large values of K (e.g., it is observed for K = 120 but not for K = 123) but seems to be more frequent when K is significantly larger than 100. Our interpretation is that clusters are first used to fit small-scale features (which correspond to the largest fluctuations of the density of local events) to maximize the likelihood; then, when all small-scale features are correctly fitted, supplementary kernels (i.e., for $K > 100$) are possibly used to fit larger-scale residuals of the distribution, hence the computation of such anomalous large-scale clusters.

[103] Figure 9 shows that the error bars preclude the determination of a very precise value of $K_{\max}$. Once a value of $K_{\max}$ has been chosen, we can fit the whole data set using that optimal number of kernels. In the following, we shall thus discuss and compare in more details the statistical properties of individual clusters as a function of K, focusing on their orientation and size. We shall see that the results also suggest $K_{\max} \simeq 100$.

8.3.2. Anisotropy Properties

[104] The standard way for studying the anisotropy properties of a given fault network is the use of stereographical plots. These plots require the knowledge of the strike λ and the dip δ of the clusters. We now introduce another parameter β which we simply define as $\beta = \lambda \bmod \pi$, which is thus a measure of the azimuth of a fault, independent of its dip.

[105] The distribution of dips does not show a strong dependence on K when $K > 100$. When binning the dip axis by steps of 10°, we find that the entropy of the histogram of fault dips reaches a local approximate minimum for $K \simeq 100$. Performing the same analysis on the histogram of β values, a minimum entropy is also observed when $K \simeq 100$. This suggests that for $K < 100$, using too few kernels to fit the data forces some of them to gather small-scale uncorrelated clusters into larger ones with a more or less random orientation. When $K > 100$, we try to explain the data with too many kernels, so that we fit smaller-scale features that are likely to possess a geometry depending more heavily on location errors: they are thus oriented more randomly too. This result once again suggests that the optimal number of clusters that one should use to fit the catalog should be set to $K \simeq 100$. When considering the joint distribution of (λ, δ) (by binning both axes by steps of 10°), the corresponding

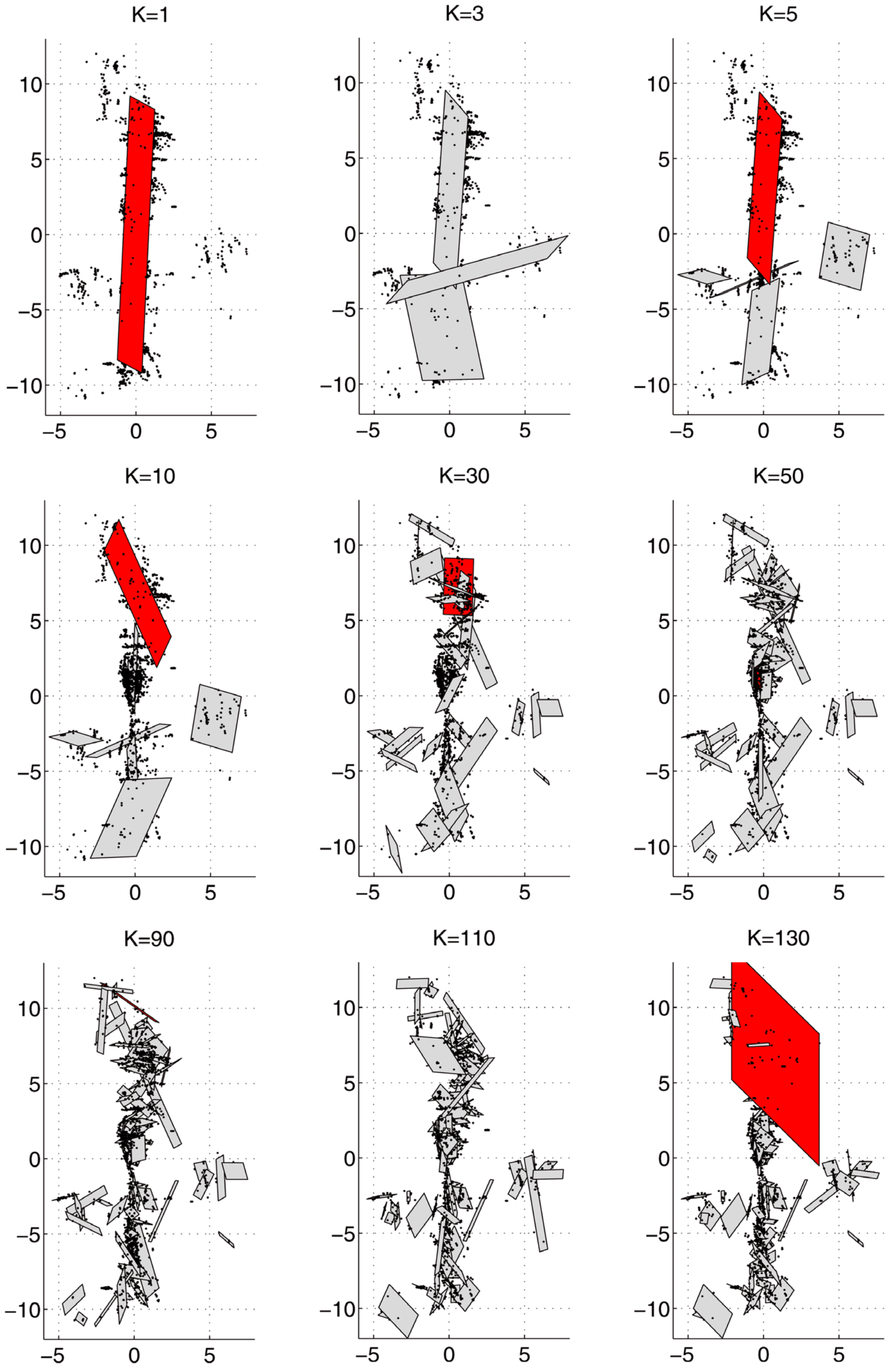

Figure 10. Fit of the catalog of clustered events shown in Figure 6 with various number of planes. All scales in km. The darkest plane (red) is the one that will be split in the next step by going from K to K+1 to form the initial conditions of the EM procedure used in the text.

entropy monotonically increases with K, which tempers the previous conclusion. The minimum observed dip for the fits obtained for $K = 100$ to $K = 130$ varies between 13° and 17°, but the solution for $K = 100$ features the largest proportion of planes with a dip larger than 30° (99%), as well as the largest proportion (84%) of planes with a dip larger than 60°. The cumulative distribution of dips for $K = 100$ is shown on Figure 12.

[106] Figure 13 shows a rose histogram of the observed β values in the case $K = 100$. One can clearly notice three main directions of faulting: the main one strikes about N-S, while the other ones strike N75° and N110°–115°, respectively (and thus seem to be conjugate directions of faulting). No correlation was found between dip and strike or β. We shall now look at the distribution of the size of clusters.

8.3.3. Size and Shape of Clusters

[107] Figure 14 shows the mean size of clusters (length L, width W and thickness T) as a function of K. The length and width of clusters have already been defined in section 7. The thickness is defined as $4\sigma_3$, where σ_3 is the standard

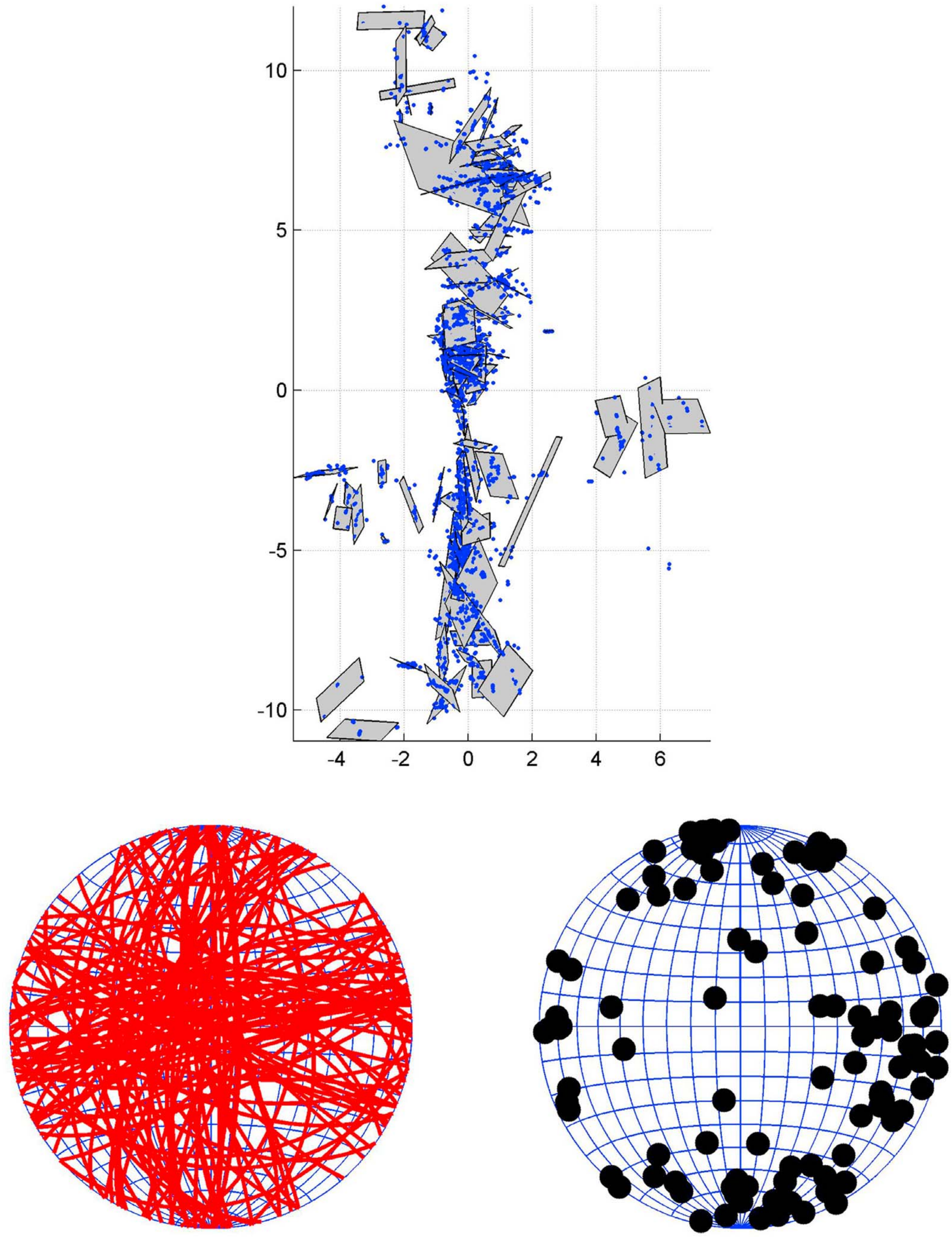

Figure 11. (top) Fit of the catalog of clustered events shown in Figure 6 with 100 planes and (bottom) their associated stereographic projections. All scales in km.

deviation of events in a direction normal to the fault plane (the factor 4 ensures that the thickness corresponds to the spatial domain covered by 96% of the associated events when the cluster obeys a genuine Gaussian distribution). The volume of a cluster is simply defined as the product of the three former quantities.

[108] Figure 14 reveals several features: first, all quantities decrease with K up to $K = 100$, and seem to reach a constant value beyond this threshold. The decay stems from the ability to fit smaller features by adding more kernels, but values of K larger than 100 seem of no use to decipher the small-scale structure of the catalog. This is thus a supplementary argument to propose $K \simeq 100$ as the maximum number of kernels to use to fit this data set. The linearity of the log-log plot in Figure 14 for large values of K (from 20 up to 100, which is less than an order of magnitude) reveals underlying power laws: $L \sim K^{-0.4}$, $W \sim K^{-0.55}$ and $T \sim K^{-0.85}$. These power laws can be restated as

$$K \sim L^{-2.5}, \quad W \sim L^{1.37}, \quad T \sim L^{2.1}. \tag{19}$$

The first scaling law $K \sim L^{-2.5}$ may correspond to a measure of the capacity fractal dimension $D_f = 2.5$ of the reconstructed fault network. The two other scaling laws

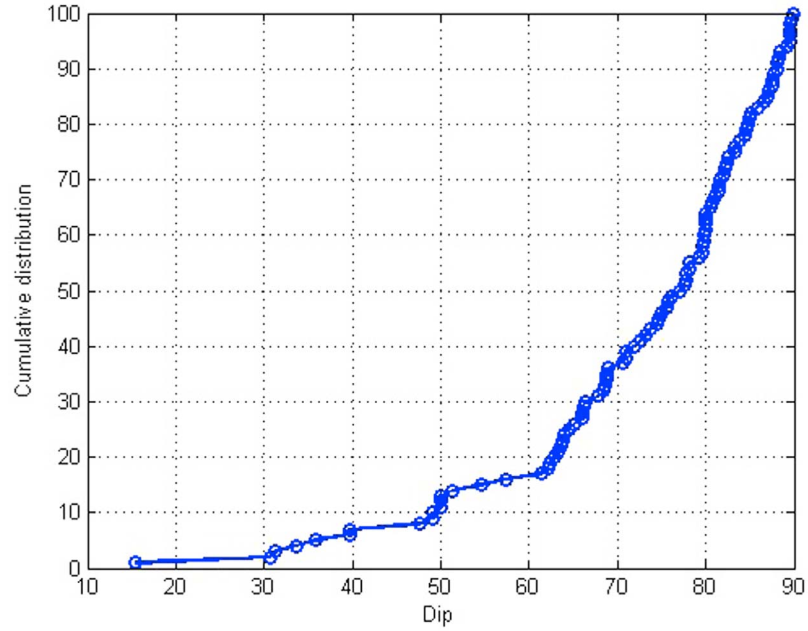

Figure 12. Cumulative distribution of fault dips for the fit of the Mount Lewis sequence with K = 100 kernels.

express the self-affinity properties of the faults in the network. The fact that the exponents 0.55/0.4 = 1.37 and 0.85/0.4 = 2.1 are larger than 1 implies that the scaling laws hold in the small size asymptotic (so-called "ultraviolet" regime): they suggest that small fault segments tend to become more and more one-dimensional with very thin width, the smaller they are. This is in agreement with the mechanics of shear localization, in which the smallest structures are localized at the scale of grain sizes while large structures are processed zones with thicker cores.

[109] Relating the number K of clusters to the mean volume of the clusters (defined as the product of their length, width and thickness) provides an estimate of the average properties of the fault network with respect to a unique size measure. Considering a characteristic length ε equal to the cubic root of the mean volume, we plot K as a function of ε in Figure 15. The two anomalous data points on the left of Figure 15 correspond to K = 120 and K = 130, and we already discussed their origin. Discarding these two points, the log-log plot appears to be linear for small values of ε, showing that $K \simeq \varepsilon^{-D}$ with $D \simeq 1.85$. This exponent compares well with the fractal dimension D_0 = 1.8–2.0 of hypocenters in southern California proposed by *Robertson et al.* [1995] using a high-quality relocated catalog. Yet we must acknowledge that the scaling we observe holds over a very limited range of scale, due to the small spatial extent of the data set.

[110] Figure 16 shows the probability density function of the length, width and thickness of clusters for K = 100. All distributions span a very limited scale interval, so that inferring the theoretical underlying distribution is not obvious. However, for length and width, this semilog plot shows that both are compatible with exponential distributions. The exponential decay suggests characteristic sizes of $\simeq$2.4 km for length and $\simeq$0.9 km for width. Both values are close from the true averages of cluster length and width for K = 100 (2.27 km and 0.81 km, respectively). The thickness of clusters spans a much smaller-scale interval and does not seem to obey any simple analytical form. It is almost flat in between its extreme values. It is worth noticing that when K = 100, the average value of σ_3 is about 72 m, while the average uncertainty of relative locations is about 8 m (one standard deviation). This suggests that the observed thickness of clusters does not stem from location uncertainties, but is rather a genuine geometrical feature of fault zones.

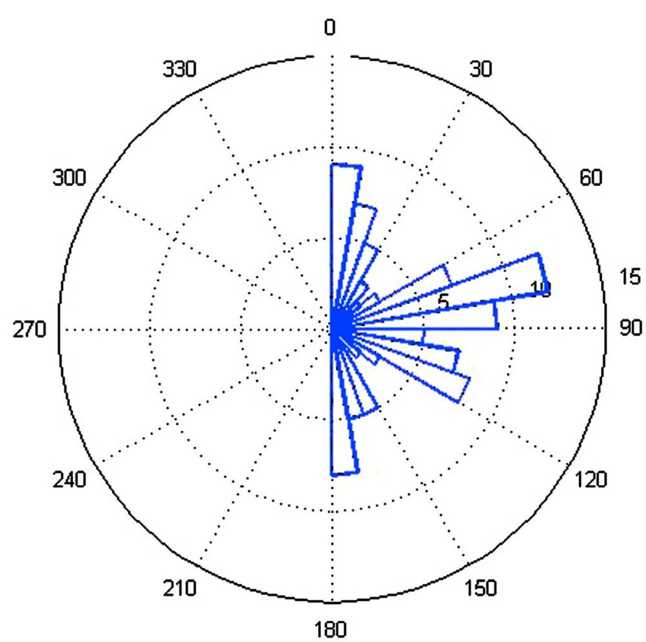

Figure 13. Histogram of β (azimuth of fault trace) for K = 100.

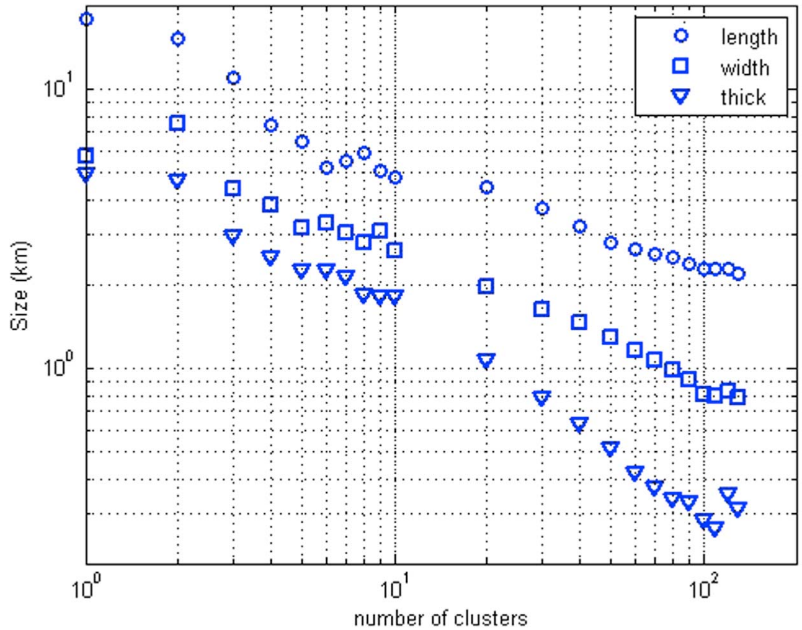

Figure 14. Mean length, width, and thickness of clusters as a function of K.

[111] We were unable to identify any correlation between the size parameters of clusters (length, width, or thickness) for given values of K. This may be due to the very limited interval covered by those sizes (about 1 order of magnitude at most). We thus computed the average value of the ratios L/W (hereafter ρ_{LW}) and L/T (hereafter ρ_{LT}) over all clusters for fixed values of K. Figure 17 shows that the ratio ρ_{LW} is approximately constant with K, suggesting that at any scale, the length of clusters is about 3 times their width (for K = 100 we get ρ_{LW} = 3.37). The other shape ratio, ρ_{LT}, shows much wilder fluctuations with K, and is about 10.7 for K = 100. Clusters thus appear as quite elongated structures with significant anisotropy in all spatial directions. Note that those results differ a little from what can be deduced from equation (19), as the latter allows to define ratios of the means and not means of the ratios.

8.4. Future and Necessary Improvements of the Method

[112] One may argue that the results presented above may be influenced by arbitrary choices of parameters governing the clustering algorithm, or may not be geologically plausible. We shall thus now briefly discuss those potential

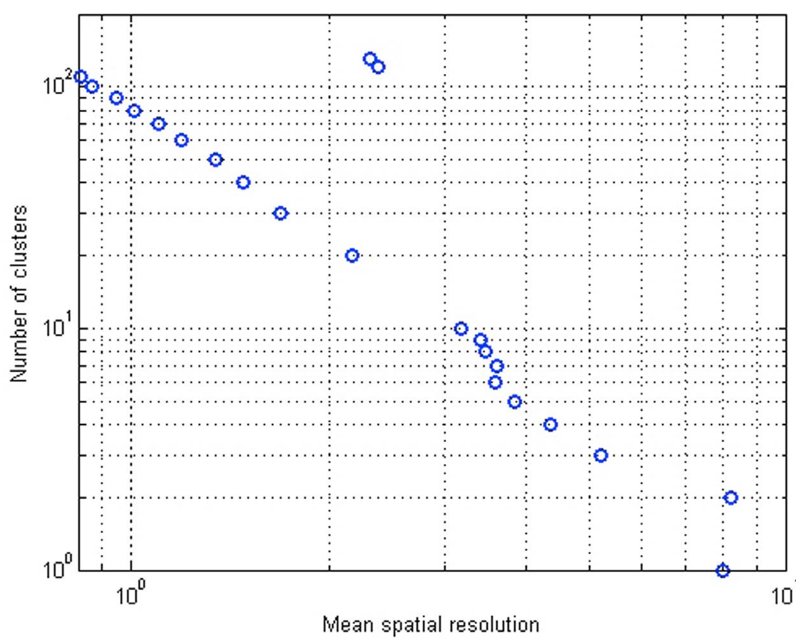

Figure 15. Relationship between the average spatial resolution ε and the number K of clusters.

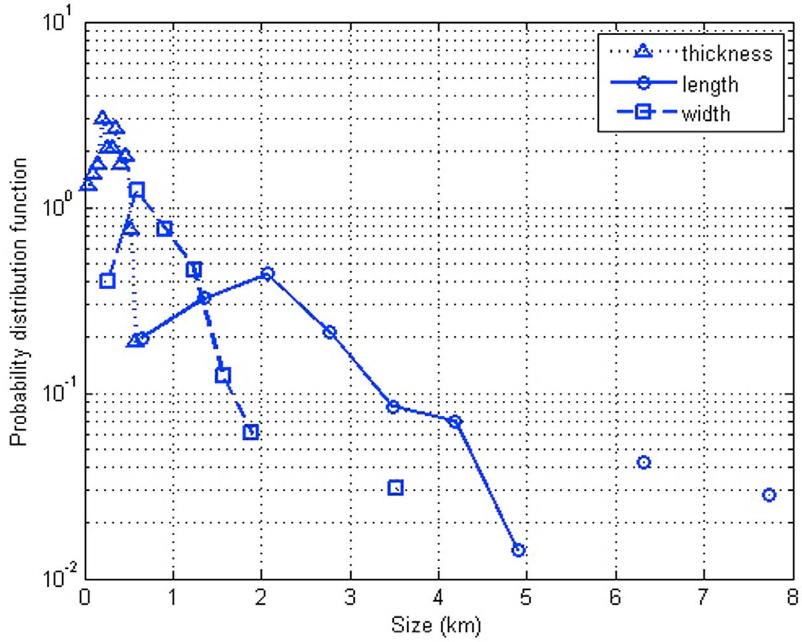

Figure 16. Probability distribution of length, width, and thickness of clusters for a number of clusters equal to $K = 100$.

criticisms and provide a rough picture of the next improvements of the method we plan to implement.

[113] Whatever the number of segments used to fit the Lewis catalog, our clustering algorithm leads to cases where fault planes intersect each other. While it is a documented fact that many natural settings present active branching faults, there is no evidence that the faults cross each other. Indeed, the existence of fault intersections seem to be incompatible with the mechanics of fault growth, due to screening effects. Moreover, one of the faults would be split into two segments separated by an offset which would lead to invert 3 fault segments in all. However, as the reconstructed planes intersect rather rarely and on their periphery in Figure 10 or 11, we expect that allowing faults to cross does not influence significantly the quantitative analyses of segment sizes and orientations presented in Figures 12–17. Having said that one should use the fraction of fault intersections as an additional metric for the quality

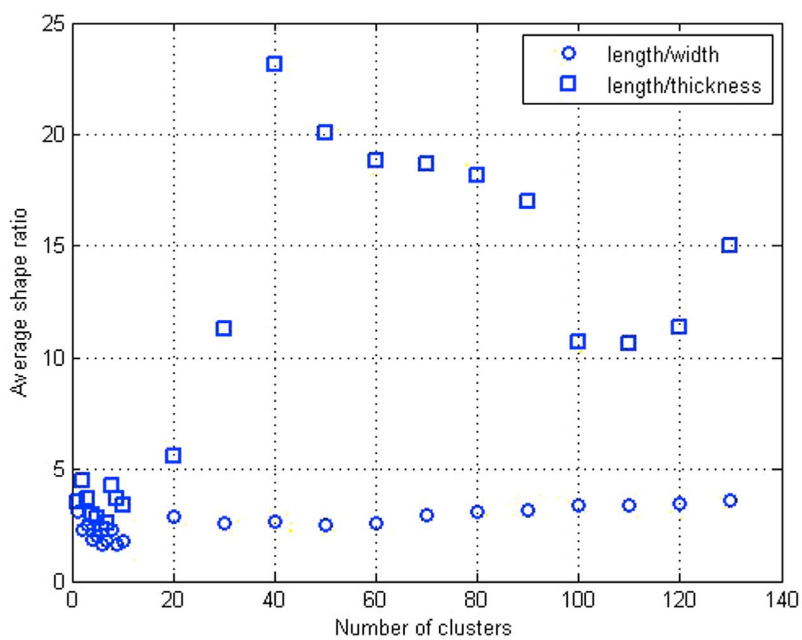

Figure 17. Dependence of mean shape ratios L/W and L/T as a function of number K of clusters.

of the fit of the set of earthquake hypocenters by the reconstructed fault network. One can also envision an upgrade of our algorithm that would consist in splitting a given fault when it is found to intersect another one. In any case, one must notice that the size of a segment in each direction is given by the eigenvalues of the associated cluster's correlation matrix, assuming that events are uniformly distributed over the segment [*Ouillon et al.*, 2008]. This simplification allows one to compute an approximative length and width, but more complex criteria should be used to check if two fault segments really intersect or not.

[114] The method we propose in this paper seems to be applicable to planar faults only. However, natural faults may exhibit some curvature, such as for listric faults. Using our algorithm, a curved fault featuring a uniform distribution of events would be approximated by a set of planar segments whose number would depend on the curvature of the fault and on the location uncertainties of the events. An improvement would consist in defining kernels that take into account a nonzero curvature. But, it is possible that many natural faults, which are interpreted as curved at a macroscopic resolution level, are actually composed of many individual approximately planar segments, organized to form a curved envelop. We also notice that the fit of the Lewis earthquake sequence has selected segments of kilometric size, for which curvature effects are not significant.

[115] More generally, one would like to define scenarios of geologically reasonable fault networks which would be difficult to reconstruct due to their geometry and/or hypocenter sampling bias. Many such scenarios can be imagined, yet it is difficult to prove that they would be geologically reasonable. This idea is too broad to help define a reliable strategy for testing the consistency of the obtained fault networks. Another strategy is to include as much independent geological information as possible during the fitting process, such as the "known" geometry of surface fault traces (see *Ouillon et al.* [1996] for methods and results showing their scale-dependent nature), earthquake focal mechanisms or the size of events. This is a promising direction for future research, if one remains critical about possible arbitrary constraints that could backfire with respect to the philosophy of "letting the data speak by itself and decide about the optimal fault network."

[116] The algorithm defined in the present paper still features some human "arbitrary" decisions. For example, the number of events in the training and validation data sets is controlled by the parameter p. We already mentioned that the choice of p does not significantly influence the results, which is due to the fact that our data set features a rather large amount of events. We thus expect p to play an even smaller role for larger data set sizes. A more critical parameter maybe the quantile used when sorting clustered and unclustered events as in section 8.2. We arbitrarily chose the q% quantile with $q = 5$ as it is a standard choice for significance tests in statistics. Changing q would increase or decrease the number of background events in the fitted data set. In the future, we propose to include events in the fitted data set according to their probability to belong to the set of clustered events. This will naturally remove the need for fixing a specific value for q. The last arbitrary choice is the way we split clusters to add a supplementary kernel in the fit. We chose to split the thickest cluster into two equal parts along its length. There is a large number of other criteria that could be used instead. Other reasonable choices that should be explored in the future include splitting the cluster that is the farthest away from a 3-D Gaussian shape, and/or splitting it in the direction where it is the farthest away from a 1-D Gaussian shape. Rather than splitting it just in its middle, we may also choose another location according to the shape of the cluster.

[117] The choice of a Gaussian cluster may be questioned as well, as we list in section 5, many observations suggesting a kernel with fatter tails. This can be achieved by replacing the multivariate Gaussian kernel by a multivariate Student's t kernel. As the number of degrees of freedom will vary from cluster to cluster, we expect to fit the data with kernels featuring tails with different thicknesses. We then hope to be able to classify clusters according to their tail behavior. This could lead to a significant improvement in the description of the fault network.

[118] Finally, we also plan to implement a fully Bayesian approach, which has the advantage of providing a probability distribution for the outcome solutions, including the optimal number of kernels to use. The spirit of the algorithm is similar to those of *Corduneanu and Bishop* [2001] and *Bishop* [2007, Chapter 10], who present two different basic algorithmic versions. The former allows determination of the optimal number of kernels that maximizes the probability of the model, while the latter derives the full distribution and thus allows averaging over the space of models. This approach can thus allow one to compute the probability of the presence of a fault at any location, which may be a very interesting information in areas where data are scarce. We plan to modify both algorithms in order to introduce uncertainties in data spatial location, which is in the spirit of Bayesian reasoning.

9. Discussion and Conclusion

[119] A fundamental and still unresolved question in brittle tectonics deals with the mechanics of fault growth. For example, strike-slip faults generally display discrete linear fault segments offset by compressive or tensile fault steps. It is believed that as time increases, those steps (or any other nonplanar fluctuation) are progressively damaged and destroyed to allow segments linkage into larger-scale segments, themselves offset by larger-scale steps [*de Joussineau and Aydin*, 2009; *Ben-Zion and Sammis*, 2003]. It seems that this process is self-affine (and not self-similar as claimed by *de Joussineau and Aydin* [2009]) so that the fault appears smoother and smoother as cumulative slip increases. In the end, combined with the mechanics of constrained fragmentation in the fault gouge, the active part of the fault reduces to a very well defined principal fracture surface which is nearly an Euclidean surface [*Ben-Zion and Sammis*, 2003]. In that picture, the damage zone around the principal fracture surface that is observed in the field is just a record of the past fault history and does not host any significant seismic activity. If surface field observations seem to support that model, it is still a matter of debate for faulting at seismogenic depth, notwithstanding the fact that this view may hold for mature faults but might be only a rough approximation for younger faults. On other grounds, fault segmentation is an important feature of faults that

ultimately controls the initiation, size and dynamics of coseismic ruptures, while the damaged zone may play a significant role in crustal fluid flow or act as a guide for surface waves.

[120] Fault segmentation was up to now largely inferred from the statistical analysis of fault maps. A few geophysical techniques may also be used for imaging tectonic structures at depth, but their rather poor resolution precludes segregating active from inactive components of the fault structures. In this work, we have proposed a new approach which takes account of seismicity itself to infer the three dimensional dynamics of faulting at depth. We are aware that earthquake catalogs do not sample properly and uniformly the deformation processes at depth (e.g, we miss aseismic creep events and slow earthquakes), but they reveal at least the minimum spatial extent of active processes. Earthquakes are never distributed in a uniform manner but display strong and complex clustering properties, which may even change with the considered timescales. The spatial shape of those clusters may nevertheless bear similarities with the structure of faults. We thus used a data clustering approach (based on the expectation maximization algorithm) to provide a probabilistic reconstruction of spatial clusters within an earthquake catalog, without taking account of their causal relationships. The main idea we developed is to approximate the local density of events by a linear superimposition (a mixture) of Gaussian kernels. This particular shape of kernels ensures that we do not assume anything about the clusters, except that they can be characterized by a position, an orientation and a finite spatial extent. This mixture approach, when combined with a cross validation scheme, allows to determine the number of clusters that compose the analyzed catalog; that is, it provides a maximum bound for the number of kernels needed to approximate coherent structures in the catalog. In the case of the Mount Lewis sequence that we have analyzed, we found that $\simeq$100 clusters are needed to explain the sequence of events that followed the Mount Lewis 1986 event.

[121] The properties of those clusters have been analyzed using standard statistical techniques. We could observe that the orientation of clusters is not random but reveals the existence of three distinct directions of faulting. The main one (nearly N0) corresponds to the current boundary conditions in northern California and is thus correlated with younger structures. It is, for example, fully compatible with the direction of the fault plane given by the focal mechanism of the main event. The two other directions (N70° and N110°) are most likely inherited from previous tectonic history. Our results show that those older structures still play a significant role in the seismogenesis in that area, and thus should not be discarded in seismic hazard analyses.

[122] The spatial extent of clusters provides an important insight about the scale of fault segmentation, which is thought to be a function of the cumulative slip over a fault [*de Joussineau and Aydin*, 2009]. We showed that both the length and width of clusters are exponentially distributed, i.e., display characteristic values. The typical length of clusters is found to be about 2–2.5 km, while their typical width is about 0.8–0.9 km. Such high resolution can only be achieved when using high-quality catalogs built with modern techniques of earthquake location. Assuming that the clusters we identified are genuine fault segments, then Figure 3 of de Joussineau and Aydin predicts that the maximum slip on the underlying faults should lie within [200 m; 2 km]. It is worth noticing that the size of the clusters is comparable to the assumed size of the main shock rupture, so that the latter may have broken a single fault segment. It is also similar to the typical fault segment length empirically observed by *Kilb and Rubin* [2002] in the same area.

[123] Finally, the typical thickness of clusters is found to be 290 m (which is the average value). A natural debate that arises is whether this quantity stems from location uncertainties and errors alone or reveals a genuine structural property of the fault. It is much larger than values that could be expected from location errors alone (which are about 30–40 m in the studied catalog). Our conclusion is thus that a fault cannot be considered as an extremely thin surface, but rather as a very elongated volume with finite thickness. One may think that crossing faults may artificially inflate the thickness of clusters to provide better fits at the crossings. This explanation can be ruled out, as shown by our study of the synthetic example studied in section 7, whose correct parameters for the three fault planes were correctly retrieved by our algorithm. It is worth noting that the thickness mentioned above is compatible with the thickness of the damage zone which is commonly observed around faults [*Ben-Zion and Sammis*, 2003] and with the one deduced from other studies of seismicity around faults [*Powers and Jordan*, 2010]. The latter suggests a damage zone thickness $2d$ within 100–600 m. The various faults in the Mount Lewis area may thus not be mature enough to have developed a significant fault core able to localize earthquake ruptures. This last conclusion is compatible with the idea that the Mount Lewis fault is a young structure which is still in its early growth process. Future work will focus on the active faults analyzed by *de Joussineau and Aydin* [2009] in order to check how the segmentation properties at the surface correlates with the size of earthquake clusters at depth.

[124] Fitting the data with an increasing number of kernels allows one to inspect spatial features at smaller and smaller scales within the catalog. It is thus akin to a multiscale analysis. In the latter, the resolution is fixed and one can compute some relevant parameters, such as the number of clusters as a function of scale, for instance. The present approach has an inverted logic: choosing the number of clusters, we could infer their position, size and orientation as well as scaling properties (see expression (19)).

[125] More work is needed to provide a stronger theoretical formalism of data clustering of multifractal sets of events. Some may see an incompatibility between the scale invariance of earthquake catalogs and the exponential distribution of the size of clusters, as one would certainly expect a power law distribution. It is then worth reminding a very popular and deterministic fractal segmentation model, i.e., the Cantor set. In its most classical construction process, surviving segments are divided into three equal parts, one of which being removed (and replaced by a hole) while the two other ones play the role of survivors for the next iteration. At each scale, all surviving segments possess the same size, but the whole set geometry (including the holes) is self-similar. There is thus no contradiction between the existence of characteristic fault or segment sizes (nonscaling) and self-

similarity of the fault network. *de Joussineau and Aydin* [2009] analyzed the distribution of fault segments length in various settings and found it to be a lognormal distribution, i.e., this distribution possesses a characteristic size while the fault network seems to display scale-invariant characteristics. *Sornette* [1991] has also provided examples to debunk the misconception that power law distributions of fault length are in any way related to the fractal geometrical organization of their network.

References

Allégre, C. J., J.-L. Le Mouel, and A. Provost (1982), Scaling rules in rock fracture and possible implications for earthquake prediction, *Nature*, *297*, 47–49, doi:10.1038/297047a0.

Ben-Zion, Y., and D. J. Andrews (1998), Properties and implications of dynamic rupture along a material interface, *Bull. Seismol. Soc. Am.*, *88*, 1085–1094.

Ben-Zion, Y., and C. G. Sammis (2003), Characterization of fault zones, *Pure Appl. Geophys.*, *160*, 677–715, doi:10.1007/PL00012554.

Bishop, C. M. (2007), *Pattern Recognition and Machine Learning*, 2nd ed., Springer, New York.

Bowman, D. D., G. Ouillon, C. G. Sammis, A. Sornette, and D. Sornette (1998), An observational test of the critical earthquake concept, *J. Geophys. Res.*, *103*(B10), 24,359–24,372, doi:10.1029/98JB00792.

Corduneanu, A., and C. M. Bishop (2001), Variational Bayesian model selection for mixture distributions, in *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*, edited by T. Richardson and T. Jaakkola, pp. 27–34, Morgan Kaufmann, Boston, Mass.

de Joussineau, G., and A. Aydin (2009), Segmentation along strike-slip faults revisited, *Pure Appl. Geophys.*, *166*, 1575–1594, doi:10.1007/s00024-009-0511-4.

Duda, R. O., P. E. Hart, and D. G. Stork (2001), *Pattern Classification*, 2nd ed., Wiley-Interscience, New York.

Eneva, M. (1994), Monofractal or multifractal: A case study of spatial distribution of mining induced seismic activity, *Nonlinear Processes Geophys.*, *1*, 182–190, doi:10.5194/npg-1-182-1994.

Faulkner, D. R., T. M. Mitchell, D. Healy, and M. J. Heap (2006), Slip on 'weak' faults by the rotation of regional stress in the fracture damage zone, *Nature*, *444*, 922–925, doi:10.1038/nature05353.

Gardner, J. K., and L. Knopoff (1974), Is the sequence of earthquakes in southern California, with aftershocks removed, Poissonian?, *Bull. Seismol. Soc. Am.*, *64*(5), 1363–1367.

Hauksson, E. (2010), Spatial separation of large earthquakes, aftershocks, and background seismicity: Analysis of interseismic and coseismic seismicity patterns in southern California, *Pure Appl. Geophys.*, *167*, 979–997, doi:10.1007/s00024-010-0083-3.

Hauksson, E., J. Woessner, and P. Shearer (2006), Associating seismicity to late Quaternary faults in southern California, *Eos Trans. AGU*, *87*(52), Fall Meet. Suppl., Abstract S11B-04.

Helmstetter, A., and D. Sornette (2002), Subcritical and supercritical regimes in epidemic models of earthquake aftershocks, *J. Geophys. Res.*, *107*(B10), 2237, doi:10.1029/2001JB001580.

Helmstetter, A., and D. Sornette (2003a), Foreshocks explained by cascades of triggered seismicity, *J. Geophys. Res.*, *108*(B10), 2457, doi:10.1029/2003JB002409.

Helmstetter, A., and D. Sornette (2003b), Importance of direct and indirect triggered seismicity in the ETAS model of seismicity, *Geophys. Res. Lett.*, *30*(11), 1576, doi:10.1029/2003GL017670.

Helmstetter, A., D. Sornette, and J.-R. Grasso (2003), Mainshocks are aftershocks of conditional foreshocks: How do foreshock statistical properties emerge from aftershock laws, *J. Geophys. Res.*, *108*(B1), 2046, doi:10.1029/2002JB001991.

Kagan, Y. Y. (1981a), Spatial distribution of earthquakes: the three-point moment function, *Geophys. J. R. Astron. Soc.*, *67*, 697–717.

Kagan, Y. Y. (1981b), Spatial distribution of earthquakes: the four-point moment function, *Geophys. J. R. Astron. Soc.*, *67*, 719–733.

Kagan, Y. Y. (1992), Correlations of earthquake focal mechanisms, *Geophys. J. Int.*, *110*, 305–320, doi:10.1111/j.1365-246X.1992.tb00876.x.

Kagan, Y. Y. (1994), Observational evidence for eathquakes as a nonlinear dynamic process, *Physica D*, *77*, 160–192, doi:10.1016/0167-2789(94)90132-5.

Kagan, Y. Y., and L. Knopoff (1978), Statistical study of the occurrence of shallow earthquakes, *Geophys. J. R. Astron. Soc.*, *55*, 67–86.

Kagan, Y. Y., and L. Knopoff (1980), Spatial distribution of earthquakes: The two-point correlation function, *Geophys. J. R. Astron. Soc.*, *62*, 303–320.

Kilb, D., and A. M. Rubin (2002), Implications of diverse fault orientations imaged in relocated aftershocks of the Mount Lewis, M_L 5.7, California, earthquake, *J. Geophys. Res.*, *107*(B11), 2294, doi:10.1029/2001JB000149.

Knopoff, L., Y. Y. Kagan, and R. Knopoff (1982), b-values for foreshocks and aftershocks in real and simulated earthquake sequences, *Bull. Seismol. Soc. Am.*, *72*(5), 1663–1675.

Liu, J., K. Sieh, and E. Hauksson (2003), A structural interpretation of the aftershock "cloud" of the 1992 M_w 7.3 Landers earthquake, *Bull. Seismol. Soc. Am.*, *93*(3), 1333–1344, doi:10.1785/0120020060.

Lyakhovsky, V., Y. Ben-Zion, and A. Agnon (2001), Earthquake cycle, fault zones, and seismicity patterns in a rheologically layered lithosphere, *J. Geophys. Res.*, *106*(B3), 4103–4120, doi:10.1029/2000JB900218.

MacQueen, J. (1967), Some methods for classification and analysis of multivariate observations, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, edited by L. M. Le Cam and J. Neyman, pp. 281–297, Univ. of Calif. Press, Berkeley.

Manighetti, I., M. Campillo, C. Sammis, P. M. Mai, and G. King (2005), Evidence for self-similar, triangular slip distributions on earthquakes: Implications for earthquake and fault mechanics, *J. Geophys. Res.*, *110*, B05302, doi:10.1029/2004JB003174.

Marsan, D., and O. Lengliné (2008), Extending earthquakes' reach through cascading, *Science*, *319*(5866), 1076–1079, doi:10.1126/science.1148783.

Martel, S. (1990), Formation of compound strike-slip fault zones, Mount Abbot Quadrangle, California, *J. Struct. Geol.*, *11*, 799–814.

Martel, S., D. D. Pollard, and P. Segall (1988), Development of simple fault zones in granitic rock, Mount Abbot Quadrangle, Sierra Nevada, California, *Geol. Soc. Am. Bull.*, *100*, 1451–1465, doi:10.1130/0016-7606(1988)100<1451:DOSSSF>2.3.CO;2.

Miltenberger, P., D. Sornette, and C. Vanneste (1993), Fault self-organization as optimal random paths selected by critical spatio-temporal dynamics of earthquakes, *Phys. Rev. Lett.*, *71*, 3604–3607, doi:10.1103/PhysRevLett.71.3604.

Nemser, E. S., and D. S. Cowan (2009), Downdip segmentation of strike-slip fault zones in the brittle crust, *Geology*, *37*(5), 419–422, doi:10.1130/G25619A.1.

Ouillon, G., and D. Sornette (1996), Unbiased multifractal analysis: Application to fault patterns, *Geophys. Res. Lett.*, *23*(23), 3409–3412, doi:10.1029/96GL02379.

Ouillon, G., D. Sornette, and C. Castaing (1995), Organization of joints and faults from 1 cm to 100 km scales revealed by optimized anisotropic wavelet coefficient method and multifractal analysis, *Nonlinear Processes Geophys.*, *2*, 158–177, doi:10.5194/npg-2-158-1995.

Ouillon, G., C. Castaing, and D. Sornette (1996), Hierarchical scaling of faulting, *J. Geophys. Res.*, *101*(B3), 5477–5487, doi:10.1029/95JB02242.

Ouillon, G., C. Ducorbier, and D. Sornette (2008), Automatic reconstruction of fault networks from seismicity catalogs: Three-dimensional optimal anisotropic dynamic clustering, *J. Geophys. Res.*, *113*, B01306, doi:10.1029/2007JB005032.

Pisarenko, V. F., A. Sornette, D. Sornette, and M. V. Rodkin (2008), New approach to the characterization of Mmax and of the tail of the distribution of earthquake magnitudes, *Pure Appl. Geophys.*, *165*, 847–888, doi:10.1007/s00024-008-0341-9.

Plesch, A., et al. (2007), Community Fault Model (CFM) for southern California, *Bull. Seismol. Soc. Am.*, *97*(6), 1793–1802, doi:10.1785/0120050211.

Powers, P. M., and T. H. Jordan (2010), Distribution of seismicity across strike-slip faults in California, *J. Geophys. Res.*, *115*, B05305, doi:10.1029/2008JB006234.

Reasenberg, P. A. (1985), Second-order moment of central California seismicity, 1969–1982, *J. Geophys. Res.*, *90*, 5479–5495, doi:10.1029/JB090iB07p05479.

Robertson, M., C. G. Sammis, M. Sahimi, and A. Martin, (1995), The 3-D spatial distribution of earthquakes in southern California with a percolation theory interpretation, *J. Geophys. Res.*, *100*, 609–620, doi:10.1029/94JB02463.

Scholz, C. H. (2002), *The Mechanics of Earthquakes and Faulting*, 2nd ed., Cambridge Univ. Press, Cambridge, U. K.

Scholz, C. H., N. H. Dawers, J.-Z. Yu, M. H. Anders, and P. A. Cowie (1993), Fault growth and fault scaling laws: Preliminary results, *J. Geophys. Res.*, *98*, 21,951–21,961, doi:10.1029/93JB01008.

Sornette, D. (1991), Self-organized criticality in plate tectonics, in *Spontaneous Formation of Space-Time Structures and Criticality*, *NATO ASI Ser., Ser. C*, vol. 349, edited by T. Riste and D. Sherrington, pp. 57–106, Kluwer Acad., Dordrecht, Netherlands.

Sornette, D. (2006), *Critical Phenomena in Natural Sciences: Chaos, Fractals, Self-Organization, and Disorder: Concepts and Tools*, 2nd ed., Springer, Heidelberg, Germany.

Sornette, D., and P. Davy (1991), Fault growth model and the universal fault length distribution, *Geophys. Res. Lett.*, *18*, 1079–1081, doi:10.1029/91GL01054.

Sornette, D., and M. J. Werner (2005a), Constraints on the size of the smallest triggering earthquake from the ETAS model, Baath's law, and observed aftershock sequences, *J. Geophys. Res.*, *110*, B08304, doi:10.1029/2004JB003535.

Sornette, D., and M. J. Werner (2005b), Apparent clustering and apparent background earthquakes biased by undetected seismicity, *J. Geophys. Res.*, *110*, B09303, doi:10.1029/2005JB003621.

Sornette, D., P. Davy, and A. Sornette (1990), Structuration of the lithosphere in plate tectonics as a self-organized critical phenomenon, *J. Geophys. Res.*, *95*, 17,353–17,361, doi:10.1029/JB095iB11p17353.

Sornette, D., P. Miltenberger, and C. Vanneste (1994), Statistical physics of fault patterns self-organized by repeated earthquakes, *Pure Appl. Geophys.*, *142*(3–4), 491–527, doi:10.1007/BF00876052.

Sornette, D., P. Miltenberger, and C. Vanneste (1995), Statistical physics of fault patterns self-organized by repeated earthquakes: Synchronization versus self-organized criticality, in *Recent Progresses in Statistical Mechanics and Quantum Field Theory*, edited by P. Bouwknegt et al., pp. 313–332, World Sci., Singapore.

Vermilye, J. M., and C. H. Scholz (1998), The process zone: A microstructural view of fault growth, *J. Geophys. Res.*, *103*, 12,223–12,237, doi:10.1029/98JB00957.

Waldhauser, F., and D. P. Schaff (2008), Large-scale relocation of two decades of northern California seismicity using cross-correlation and double-difference methods, *J. Geophys. Res.*, *113*, B08311, doi:10.1029/2007JB005479.

Wang, Y. M., J. Woessner, S. Husen, D. Sornette, and G. Ouillon (2009), *Reconstructing Fault Networks From Seismicity Catalog*, Adv. Sch. on Non-linear Dyn. and Earthquake Predict., Trieste, Italy.

Weatherley, D., P. Mora, and M. Xia (2002), Long-range automaton models of earthquakes: Power law accelerations, correlation evolution and mode switching, *Pure Appl. Geophys.*, *159*, 2469–2490, doi:10.1007/s00024-002-8743-6.

Willemse, E. J. M., D. C. P. Peacock, and A. Aydin (1997), Nucleation and growth of strike-slip faults in limestones from Somerset, U.K., *J. Struct. Geol.*, *19*, 1461–1477, doi:10.1016/S0191-8141(97)00056-4.

Wilson, K. G. (1975), The renormalization group: Critical phenomena and the Kondo problem, *Rev. Mod. Phys.*, *47*(4), 773–840, doi:10.1103/RevModPhys.47.773.

Zhou, Y., K. C. McNally, and T. Lay (1993), Analysis of the 1986 Mt. Lewis, California, earthquake: Preshock sequence-mainshock-aftershock sequence, *Phys. Earth Planet. Inter.*, *75*(4), 267–288, doi:10.1016/0031-9201(93)90005-T.

Zhuang, J., Y. Ogata, and D. Vere-Jones (2002), Stochastic declustering of space-time earthquake occurrences, *J. Am. Stat. Assoc.*, *97*, 369–380, doi:10.1198/016214502760046925.

Zhuang, J., Y. Ogata, and D. Vere-Jones (2004), Analyzing earthquake clustering features by using stochastic reconstruction, *J. Geophys. Res.*, *109*, B05301, doi:10.1029/2003JB002879.

G. Ouillon, Lithophyse, 4 rue de l'Ancien Sénat, 06300 Nice, France. (lithophyse@neuf.fr)

D. Sornette, D-MTEC, ETH Zürich, Kreuzplatz 5, CH-8032 Zürich, Switzerland. (dsornette@ethz.ch)