Supplementary Material for IIEEE SMCA-22-12-3314

Likelihood Analysis of Imperfect Data

Jian-Bo Yang, Dong-Ling Xu, Xiaobin Xu, and Chao Fu

S1. Proofs of Theorems and Corollaries

S1.1 Proof of Theorem 1

If Equation (14) is used to acquire $p_{i,j}$, $p_{i,j}$ will be the probability that has the same evidential meaning as likelihood $c_{i,j}$ according to the principle of likelihood. Putting it into Equation (13) results in

$$p_{h_i,e_0\wedge e_j} = \frac{p_{i,0}p_{i,j}}{\sum_{n=1}^{N} p_{n,0}p_{n,j}} = \left(p_{i,0}\times\frac{c_{i,j}}{\sum_{t=1}^{N}c_{t,j}}\right)\Bigg/\left(\sum_{n=1}^{N}p_{n,0}\times\frac{c_{n,j}}{\sum_{t=1}^{N}c_{t,j}}\right) = \frac{p_{i,0}c_{i,j}}{\sum_{n=1}^{N}p_{n,0}c_{n,j}} = \frac{p\left(h_i|I_0\right)p\left(t_j|h_i,I_0\right)}{\sum_{n=1}^{N}p\left(h_n|I_0\right)p\left(t_j|h_n,I_0\right)} = p\left(h_i\middle|e_j,I_0\right) \tag{S1-1}$$

Equation (S1-1) asserts that the *ER* rule becomes equivalent to Bayes' rule given Equation (14) with $p_{h_i,e_0\wedge e_j}$ being the probability generated by the orthogonal sum of two independent probability distributions. Therefore, Equation (14) is a sufficient condition for the assertion of the theorem.

In the following, it is limited to discrete cases to prove that the condition given by Equation (14) is also necessary in order to ensure that the *ER* process is a probabilistic inference process, while a continuous case can be discretised as shown later in this paper. A discrete case can in general be illustrated in Table S1-1. In Table S1-1, $s_{i,j}$ stands for the number of observations of both state h_i and test result t_j, S_i for the number of all observations for state h_i, T_j for the number of all observations for test result t_j, and S for the number of all observations in the experiment.

Table S1-1 Sample Data

Frequency		Test result					Total observation
		t_1	$\cdots$	t_j	$\cdots$	t_L	
System state	h_1	$s_{1,1}$	$\cdots$	$s_{1,j}$	$\cdots$	$s_{1,L}$	S_1
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	h_i	$s_{i,1}$	$\cdots$	$s_{i,j}$	$\cdots$	$s_{i,L}$	S_i
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	h_N	$s_{N,1}$	$\cdots$	$s_{N,j}$	$\cdots$	$s_{N,L}$	S_N
Total test		T_1	$\cdots$	T_j	$\cdots$	T_L	S

If test result t_j is observed independently, combining evidence e_0 (prior distribution) and probabilistic evidence e_j (test result t_j) by the *ER* rule must result in posterior probability $s_{i,j}/T_j$ for any $i=1,\cdots,N$, which is the relative frequency or the conditional probability of state h_i given test result t_j in the experiment. That is, there must be

$$p_{h_i,e_0\wedge e_j} = \frac{p_{i,0}p_{i,j}}{\sum_{n=1}^{N}p_{n,0}p_{n,j}} = \frac{s_{i,j}}{T_j} \quad \forall h_i\in\Theta \tag{S1-2}$$

In the *prior* distribution e_0 generated from Table S1-1, there is $p_{n,0}=S_n/S$ for any $n=1,\cdots,N$. Putting this into Equation (S1-2) leads to

$$\frac{\left(S_i/S\right)p_{i,j}}{\sum_{n=1}^{N}\left(S_n/S\right)p_{n,j}} = \frac{s_{i,j}}{T_j} \quad \forall h_i\in\Theta \tag{S1-3}$$

In the above equation, note that for evidence e_j the three terms: $\sum_{n=1}^{N}\left(S_n/S\right)p_{n,j}$, S and T_j are all constant with respect to *i* for any $h_i\in\Theta$. Since the above equation is true for any $h_k\in\Theta$, dividing Equation (S1-3) for h_k by Equation (S1-3) for h_i leads to the following equation

$$p_{k,j} = \left(\frac{s_{k,j}}{S_k}\right)\left(\frac{S_i}{s_{i,j}}\right)p_{i,j} \quad \forall k\in\{1,\cdots,N\} \tag{S1-4}$$

On the other hand, as e_j is a probability distribution on the whole set of states, there is

$$\sum_{k=1}^{N}p_{k,j}=1 \tag{S1-5}$$

Solving Equations (S1-4) and (S1-5) leads to Equation (14) with $c_{i,j}=s_{i,j}/S_i$. This proves that Equation (14) is also necessary. □

S1.2 Proof of Corollary 1

Evidence e_j and total ignorance *prior* e_0 for this corollary can be profiled as the following basic probability distributions

$$e_j = \left\{\left(h_i,p_{i,j}\right),\forall h_i\in\Theta,\sum_{h_i\in\Theta}p_{i,j}=1;\left(\theta,\ 0\right),\forall\theta\subseteq\Theta,\theta\notin\Theta\right\}$$

$$e_0 = \{(\theta,\ 0),\ \forall \theta \subset \Theta;\ (\Theta,\ 1)\}$$

Applying Equation (10) to combine e_j and e_0 with $w_0 = w_1 = 1$ leads to

$$p_{\theta,e_0 \wedge e_1} = \begin{cases} 0 & \theta \subseteq \Theta, \theta \notin \Theta \\ \dfrac{p_{\theta,j} \times p_{\Theta,0}}{\sum_{D \in \Theta}\left(p_{D,j} \times p_{\Theta,0}\right)} = \dfrac{p_{\theta,j} \times 1}{\sum_{A \in \Theta}\left(p_{A,j} \times 1\right)} = p_{\theta,j} & \theta \in \Theta \end{cases}$$

Note that $p_{\theta,j} = p_{i,j}$ for any $\theta = h_i \in \Theta$ □

S1.3 Proof of Theorem 2

In Equation (10), the joint probability mass that both e_0 and e_j exactly support state θ is given by $m_{\theta,0}m_{\theta,j}/K$, where K is a normalisation factor common to all θ, with $m_{\theta,0}$ and $m_{\theta,j}$ calculated by $m_{\theta,0} = w_0 p_{\theta,0} = w_0 S_\theta / S$ and $m_{\theta,j} = w_j p_{\theta,j}$, respectively. If Equation (23) is used to acquire basic probability $p_{\theta,j}$, we will have

$$\begin{aligned} m_{\theta,0}m_{\theta,j}/K &= w_0 \frac{S_\theta}{S} w_j p_{\theta,j} \Big/ K \\ &= w_0 w_j S_\theta \left(c_{\theta,j} \Big/ \left(\sum_{A \subseteq \Theta} c_{A,j}\right)\right) \Big/ (SK) \\ &= w_0 w_j S_\theta \left(\frac{s_{\theta,j}}{S_\theta} \Big/ \left(\sum_{A \subseteq \Theta} \frac{s_{A,j}}{S_A}\right)\right) \Big/ (SK) \\ &= \frac{s_{\theta,j}}{T_j}\left(w_0 w_j T_j \Big/ \left(SK \sum_{A \subseteq \Theta} \frac{s_{A,j}}{S_A}\right)\right) = \frac{s_{\theta,j}}{T_j} K_{0,j} \quad \forall \theta \subseteq \Theta \end{aligned}$$

$$\text{with } K_{0,j} = w_0 w_j T_j \Big/ \left(SK \sum_{A \subseteq \Theta} \frac{s_{A,j}}{S_A}\right) \qquad \text{(S1-6)}$$

In Equation (S1-6), $s_{\theta,j}/T_j$ is the relative frequency of the $s_{\theta,j}$ observations of both state θ and test result t_j over all the T_j observations of t_j, or the posterior probability that state θ is true given that test result t_j is observed. Note that $K_{0,j}$ in Equation (S1-6) is constant for any θ. $m_{\theta,0}m_{\theta,j}/K$ is thus proportional to $s_{\theta,j}/T_j$ for any θ. According to the principle of likelihood [5], the former has the same evidential meaning as the latter. As such, Equation (23) is a sufficient condition for the assertion of this theorem.

On the other hand, suppose it is required that the joint probability mass $m_{\theta,0}m_{\theta,j}/K$ must have the same evidential meaning as posterior probability, or the former be proportional to the latter:

$$m_{\theta,0}m_{\theta,j}/K = w_0 \frac{S_\theta}{S} w_j p_{\theta,j} \Big/ K = \frac{s_{\theta,j}}{T_j}\overline{K} \quad \forall \theta \subseteq \Theta \quad \text{(S1-7)}$$

where $\overline{K}$ is a positive constant so that $(s_{\theta,j}/T_j)\overline{K}$ is proportional to $(s_{\theta,j}/T_j)$. It can be shown that basic probability $p_{\theta,j}$ must be calculated by Equation (23).

In fact, note that in Equation (S1-7) the terms w_0, w_j, S, K, T_j and $\overline{K}$ are all constant with regard to any $\theta \subseteq \Theta$. As Equation (S1-7) is required for any $\theta \subseteq \Theta$, dividing Equation (S1-7) for A by Equation (S1-7) for θ leads to the following equation

$$p_{A,j} = \left(\frac{s_{A,j}}{S_A}\right)\left(\frac{S_\theta}{s_{\theta,j}}\right) p_{\theta,j} \quad \forall A \subseteq \Theta \qquad \text{(S1-8)}$$

Since $p_{A,j}$ is the basic probability that test result t_j points to state A, from Equation (8), the following equation holds

$$\sum_{A \subseteq \Theta} p_{A,j} = 1 \qquad \text{(S1-9)}$$

Then, solving Equations (S1-8) and (S1-9) for $p_{\theta,j}$ leads to

$$p_{\theta,j} = \frac{s_{\theta,j}}{S_\theta} \Big/ \left(\sum_{A \subseteq \Theta} \frac{s_{A,j}}{S_A}\right) = \frac{c_{\theta,j}}{\sum_{A \subseteq \Theta} c_{A,j}} \quad \forall A \subseteq \Theta \quad \text{(S1-10)}$$

Hence, Equation (23) is also a necessary condition for the assertion of the theorem. □

S1.4 Proof of Corollary 3

Evidence e_j and total ignorance prior e_0 for this corollary can be profiled as the following basic probability distributions

$$e_j = \left\{\left(\theta,\ p_{\theta,j}\right), \forall \theta \subseteq \Theta,\ \sum_{\theta \subseteq \Theta} p_{\theta,j} = 1\right\}$$

$$e_0 = \{(\theta,\ 0),\ \forall \theta \subset \Theta;\ (\Theta,\ 1)\}$$

Applying Equation (10) to combine e_j and e_0 with $w_0 = w_1 = 1$ leads to

$$p_{\theta,e_0 \wedge e_1} = \begin{cases} 0 & \theta = \varnothing \\ \dfrac{p_{\theta,j} \times p_{\Theta,0}}{\sum_{A \subseteq \Theta}\left(p_{A,j} \times p_{\Theta,0}\right)} = \dfrac{p_{\theta,j} \times 1}{\sum_{A \subseteq \Theta}\left(p_{A,j} \times 1\right)} = p_{\theta,j} & \theta \neq \varnothing \end{cases}$$

□

S2. Illustration of likelihood inference

Given Theorem 2 and Corollary 2, we are now in a position to analyse imbalanced data with ambiguity to support evidence-based decision making and come back to investigate Example 1. Let $c_{1,1}$, $c_{2,1}$ and $c_{\Theta,1}$ represent the likelihoods that an individual is expected to have a positive test result given that he is a steroid user, steroid free and in an unknown state of steroid use, respectively. From Table 2 and Equation (20), $c_{1,1}$, $c_{2,1}$ and $c_{\Theta,1}$ are calculated as follows:

$c_{1,1} = \frac{190}{203} = 0.936$, $c_{2,1} = \frac{270}{1830} = 0.1475$,
$c_{\Theta,1} = \frac{60}{187} = 0.3209$

Let $p_{1,1}$, $p_{2,1}$ and $p_{\Theta,1}$ represent the basic probabilities that a positive test result points to the three states: steroid user, steroid free and unknown, respectively. From Equation (23), $p_{1,1}$, $p_{2,1}$ and $p_{\Theta,1}$ are given as follows.

$p_{1,1} = \frac{c_{1,1}}{c_{1,1}+c_{2,1}+c_{\Theta,1}} = \frac{0.936}{1.4044} = 0.6664$,
$p_{2,1} = \frac{0.1475}{1.4044} = 0.1051$, $p_{\Theta,1} = \frac{0.3209}{1.4044} = 0.2285$

A positive test result using test Method I can then be profiled as the following *BPD*.

$$e_1 = \left\{\begin{matrix}(Steroid\ user, 0.6664), (Steroid\ free, 0.1051),\\ (Unknown,\ 0.2285)\end{matrix}\right\} \quad \text{(S2-1)}$$

The above likelihood distribution shows that a positive test result for an athlete means that the probabilities of the athlete being a steroid user, steroid free and unknown are 0.6664, 0.1051 and 0.2285, respectively. If the prior distribution of the athlete's state of steroid use is completely unknown, as given by the following vacuum basic probability distribution [29].

$$e_0 = \left\{(Steroid\ user, 0), (Steroid\ free, 0), (Unknown, 1)\right\} \quad \text{(S2-2)}$$

It can be shown that, by using the *ER* rule of Equation (10), the combination of e_1 given by Equation (S2-1) with e_0 given by Equation (S2-2) results in e_1 itself.

Note that uniform distribution is not the same as completely unknown. In fact, if Equation (10) is used to combine e_1 of Equation (S2-1) with the following uniform distribution on the two states: steroid user and steroid free

$$e_0^{'} = \left\{(Steroid\ user, 0.5), (Steroid\ free, 0.5), (Unknown, 0)\right\}$$

It will result in the following combined distribution.

$$\left(e_0^{'} \wedge e_1\right) = \left\{\begin{matrix}(Steroid\ user, 0.729), (Steroid\ free, 0.271),\\ (Unknown, 0)\end{matrix}\right\}$$

The above result shows the difference that a uniform prior and a completely unknown prior can make in inference when there is ambiguity in data, while there is no such difference when there is no ambiguity in data, as proved in Corollary 1 and Corollary 2. If a prior distribution is not known in advance, unknown should be explicitly modelled as vacuum evidence in inference, such as by Equation (S2-2), rather than assuming uniform prior. Otherwise, unintentional or biased inference results may be incurred. This is because a uniform distribution does not really mean unknown but is as informative as assuming that it is equally likely that the athlete could be a steroid user or steroid free with no ambiguity or unknown. As such, the inferred probability of 0.729 for the steroid user state in the above result will not be credible unless the uniform prior accurately represents the athlete's true prior condition.

Since the ambiguity in evidence e_1 is as high as 22.85%, it could be controversial to use e_1 alone to infer whether the athlete is a steroid user or steroid free. A more sensible and less controversial approach is to gather more evidence before a robust decision could be made beyond reasonable doubt.

Suppose two other different methods: Method II and Method III are available for detecting steroid use, and the effectiveness of Method II and Method III is investigated by two independent experiments as shown in Table S2-1 and Table S2-2. In Table S2-1, more steroid users are tested using Method II, while in Table S2-2 more steroid free individuals are tested using Method III, so the two datasets are imbalanced in different orientations. Nevertheless, both Method II and Method III are effective in detecting steroid use in the sense that among steroid users they generate positive result over 96% and 94% of the occasions, respectively.

Table S2-1 Sample Data Using Method II

Frequency		Test results		
		Positive	Negative	Inconclusive
System states	*Steroid user*	492	14	6
	Steroid free	25	115	3
	Unknown	12	66	8

Table S2-2 Sample Data Using Method III

Frequency		Test results		
		Positive	Negative	Inconclusive
System states	*Steroid user*	260	12	4
	Steroid free	400	2300	45
	Unknown	70	160	36

If the athlete is also tested using Method II and Method III and the results are both positive, what is the probability that the athlete is a steroid user? To answer the question, the two new pieces of evidence need to be combined with the first piece of evidence e_1 acquired by using Method I.

Suppose $p_{1,1,II}$, $p_{2,1,II}$ and $p_{\Theta,1,II}$ represent the basic probabilities that a positive test result generated by using Method II points to states h_1, h_2 and Θ, respectively. From Equation (23) and Table S2-1, we get $p_{1,1,II} = 0.754$, $p_{2,1,II} = 0.137$ and $p_{\Theta,1,II} = 0.109$.

Similarly, suppose $p_{1,1,III}$, $p_{2,1,III}$ and $p_{\Theta,1,III}$ represent the basic probabilities that a positive test result generated by using Method III points to states h_1, h_2 and Θ, respectively. From Equation (23) and Table S2-2, we get $p_{1,1,III} = 0.697$, $p_{2,1,III} = 0.108$ and $p_{\Theta,1,III} = 0.195$.

The positive test result generated by using Method II and Method III can then be profiled as the following basic probability distributions, represented by $e_{1,II}$ and $e_{1,III}$, respectively.

$$e_{1,II} = \left\{\begin{matrix}(Steroid\ user, 0.754), (Steroid\ free, 0.137),\\ (Unknown, 0.109)\end{matrix}\right\} \quad \text{(S2-3)}$$

$$e_{1,III} = \left\{ \begin{matrix} (Steroid\ user, 0.697), (Steroid\ free, 0.108), \\ (Unknown, 0.195) \end{matrix} \right\} \quad \text{(S2-4)}$$

Using Equation (10), the combination of the three pieces of independent evidence from the three positive test results, i.e. e_1 from Equation (S2-1), $e_{1,II}$ and $e_{1,III}$ above, is given by

$$\left(e_1 \wedge e_{1,II} \wedge e_{1,III}\right) = \left\{ \begin{matrix} (Steroid\ user, 0.96), \\ (Steroid\ free, 0.03), (Unknown, 0.01) \end{matrix} \right\}$$

The assertion that the athlete is a steroid user is highly supported by the three pieces of evidence in that the combined probability for this assertion is 0.96, that against it is only 0.03 and unknown about it is 0.01. While there is significant ambiguity in each of the three pieces of evidence (23%, 11% and 20%, respectively), the ambiguity in the combined evidence is reduced to 1%. This is one of the main features of *ER* that ambiguity diminishes with accumulation of more evidence. Based on such an inference result, one may assert that the athlete is a steroid user beyond reasonable doubt. Therefore, it should be less controversial to come up with a guilty verdict and take disciplinary actions against the athlete.

However, if the two new test results using Method II and Method III were both negative, the conclusion would be rather different. In such a case, the combined probability for the above assertion would be only 0.3189 but that against it would be 0.5859, although the ambiguity would still be as low as 9.52%. Given such mixed test results, however, it would be premature to take any action again the athlete. In this case, it should be sensible to collect more pieces of evidence for analysis before a final decision could be made.

In the example, if we estimate prior from sample data, we obtain the following three different prior distributions from Table 2, Table S2-1 and Table S2-2, respectively.

$$e_{0,I} = \left\{ \begin{matrix} (Steroid\ user, 0.0914), (Steroid\ free, 0.8243), \\ (Unknown, 0.0842) \end{matrix} \right\}$$

$$e_{0,II} = \left\{ \begin{matrix} (Steroid\ user, 0.6910), (Steroid\ free, 0.1930), \\ (Unknown, 0.1161) \end{matrix} \right\}$$

$$e_{0,III} = \left\{ \begin{matrix} (Steroid\ user, 0.0840), (Steroid\ free, 0.8351), \\ (Unknown, 0.0809) \end{matrix} \right\}$$

As they are very different, we have only conducted likelihood inference without considering any of the priors. If the prior distribution of the athlete's state of steroid use becomes available, it can and indeed should be treated as a piece of independent evidence in the *ER* framework and combined with other evidence to generate more robust and less ambiguous conclusions.

S3. Likelihood inference for fault diagnosis

In this section, a case study about fault diagnosis for rail track maintenance management is used to demonstrate how *ER* can be implemented for likelihood inference with ambiguous data collected from an engineering system. A more detailed description and analysis of the case can be found in Section S4 of the Supplementary Material.

The data sources of the case study are three sensors installed in different parts of a train, which are separated by springs. Therefore, these sensors are deemed to record train acceleration data independently, denoted by $f_1(t)$, $f_2(t)$, $f_3(t)$, in the sense that how one sensor generates data does not depend on how the other sensors work. These are continuous readings in nature and are each discretized into five equal intervals ($t_{i,j}$, $i = 1,\ldots,5$; $j = 1, 2, 3$, as shown in Table S3-1 to Table S3-3) for illustration purpose. The irregularity of rail track is measured by absolute vertical displacement, denoted by $Ir(t)$. $Ir(t)$ is recorded by running a special train with expensive high accuracy instruments. It is also a continuous variable and is discretized into three system states: *normal* (h_1), *transient* (h_2) and *faulty* (h_3) according to policy guidance for rail track maintenance management.

In this case study, there are 880 imperfect records out of 10309 collected in total. Table S3-1 shows all the cases where imperfect data were collected from different sensors mounted in different parts of a train, where "√" or "×" means that the reading $f_k(t)$ ($k = 1, 2$ or 3) and $Ir(t)$ were recorded or were not recorded at time t. The last row in Table S3-1 shows the total number of missing datasets for each case. For example, Case 1 means that all acceleration data were recorded from the three sensors, but the corresponding absolute vertical displacement $Ir(t)$ was not recorded.

Table S3-1 10 different cases of missing data

Case	1	2	3	4	5	6	7	8	9	10
$f_1(t)$	√	√	×	√	×	×	×	×	√	√
$f_2(t)$	√	×	√	√	√	√	√	×	√	×
$f_3(t)$	√	√	×	×	√	√	×	√	×	×
$Ir(t)$	×	×	√	√	√	×	×	×	×	×
Total	350	50	100	100	100	50	40	40	50	0

Note that such missing information means that there is a degree of ambiguity or unknown that rail track is in any of the three defined states: *normal* (h_1), *transient* (h_2) and *faulty* (h_3). So, missing information is referred to as unknown, or global ignorance, and measured by beliefs assigned to the system space $\Theta = \{h_1, h_2, h_3\}$, as shown in Table S3-2, Table S3-3 and Table S3-4 below. In an engineering system, there can be local ignorance as well, or beliefs assigned to subsets of states. For example, the state of a rail track may be judged to be not *normal*, that is, it could be in either *transient* state or *faulty* state. Such locally ambiguous information can be measured by a belief assigned to the subset of states $\{h_2, h_3\}$ as a whole without a need to assume that the belief has to be further assigned to h_2 or h_3 individually. This way, ambiguity in one data source is duly respected and explicitly measured without having to making unnecessary assumptions, and later can be reduced by combining evidence from another data source. In general, ambiguity can be reduced by accumulating evidence.

The sample datasets discretized in this case study are shown from Table S3-2 to Table S3-4. A sample record f_{321}= [$f_1(t)$, $f_2(t)$, $f_3(t)$, $Ir(t)$]=[2.8992, 0.0088, 0.5371, 9.195] is chosen to

illustrate how the set of evidence acquired from the reading of the three sensors can be used to predict the state of the railway track, which is faulty because $Ir(t)$=9.195 > 8 (the threshold above which the track is faulty). From Table S3-2 to Table S3-4, this sample f_{321} is mapped to three pieces of evidence: $e_{3,1}$ $e_{2,2}$ and $e_{1,3}$ because $f_1(t)$ = 2.8992 falls into interval $t_{3,1}$ = [2.0405, 3.007] for sensor 1, $f_2(t)$ = 0.0088 falls into interval $t_{2,2}$ = [0.0051, 0.009] for sensor 2, and $f_3(t)$ = 0.5371 falls into interval $t_{1,3}$ = [0.4719, 0.5903] for sensor 3. Following the procedure as given in Section 3.1 and illustrated in Section 3.2, $e_{3,1}$, $e_{2,2}$ and $e_{1,3}$ are acquired from Table S3-2 to Table S3-4 as follows.

$$e_{3,1}=\{(h_1,0.0022),(h_2,0.0643),(h_3,0.8514),(\Theta,0.0822)\}$$

$$e_{2,2}=\{(h_1,0.3134),(h_2,0.3410),(h_3,0.1221),(\Theta,0.2235)\}$$

$$e_{1,3}=\{(h_1,0.2263),(h_2,0.2498),(h_3,0.1839),(\Theta,0.3399)\}$$

Table S3-2 Discretized sample datasets from sensor 1 ($f_1(t)$)

Frequency		$t_{1,1}$ [0.1075, 1.074)	$t_{2,1}$ [1.074, 2.0405)	$t_{3,1}$ [2.0405, 3.007)	$t_{4,1}$ [3.007, 3.9734)	$t_{5,1}$ [3.9734, 4.9399]	Missing readings
h_1	0<*Ir*≤5	9230	80	6	2	3	180
h_2	5<*Ir*≤8	167	22	4	0	0	19
h_3	8<*Ir*≤12	8	2	4	1	0	1
Θ	Unknown	404	26	14	6	0	130

Table S3-3 Discretized sample datasets from sensor 2 ($f_2(t)$)

Frequency		$t_{1,2}$ [0.0012, 0.0051)	$t_{2,2}$ [0.0051, 0.009)	$t_{3,2}$ [0.009, 0.0129)	$t_{4,2}$ [0.0129, 0.0168)	$t_{5,2}$ [0.0168, 0.0207	Missing readings
h_1	0<*Ir*≤5	3842	4571	933	149	6	0
h_2	5<*Ir*≤8	40	111	45	16	0	0
h_3	8<*Ir*≤12	1	3	4	1	7	0
Θ	Unknown	258	199	33	0	0	90

Table S3-4 Discretized sample datasets from sensor 3 ($f_3(t)$)

Frequency		$t_{1,3}$ [0.4719, 0.5903)	$t_{2,3}$ [0.5903, 0.7086)	$t_{3,3}$ [0.7086, 0.827)	$t_{4,3}$ [0.827, 0.9454)	$t_{5,3}$ [0.9454, 1.0638]	Missing readings
h_1	0<*Ir*≤5	3654	4464	1015	152	26	190
h_2	5<*Ir*≤8	90	79	23	7	4	9
h_3	8<*Ir*≤12	5	2	7	1	0	1
Θ	Unknown	335	118	26	11	0	90

Note that there are significant amounts of ambiguity in evidence $e_{2,2}$ and evidence $e_{1,3}$, measured by the probabilities of 0.2235 and 0.3399 assigned to system space Θ , respectively.

It is interesting to note that evidence $e_{3,1}$ acquired from sensor 1 to a large extent points to faulty state h_3, with a high probability of 0.8514, whilst evidence $e_{2,2}$ and evidence $e_{1,3}$, acquired from sensor 2 and sensor 3, respectively, point to non-faulty state h_1 or h_2 to larger degrees than those to h_3. Nevertheless, if these sensors are reliable, the three pieces of evidence should be combined to generate a more robust and less ambiguous diagnosis than any single sensor can provide. Using the *ER* rule or Equation (10) recursively and assuming that the weight of each piece of evidence is 1 for illustration purpose, we get the following combined result

$$(e_{3,1}\wedge e_{2,2}\wedge e_{1,3})=\left\{\begin{matrix}(h_1,0.0841),(h_2,0.1841),\\(h_3,0.7048),(\Theta,0.027)\end{matrix}\right\} \quad \text{(S3-1)}$$

Note that in engineering applications the weight of pieces of evidence generated from sensors should be estimated rather than assumed. In Section S4 of the supplementary material, a method for estimation of evidence weight is discussed in relation to this rail track maintenance management problem.

In the above result, the probability for faulty state h_3 is $p(h_3)$ = 0.7048 and the probability against h_3 is 0.2682 (0.0841+0.1841), with a very low probability of 0.027 left for unknown. This combined diagnosis result is much less ambiguous than any individual sensor can predict, providing a panoramic view for informative maintenance decision making.

To analyze the impact of ambiguity in data on probabilistic inference, let us use the common listwise deletion approach. to clean the ambiguous cases from the original data by deleting all the data of the last row and last column from each of Table1 S3-2 to Table S3-4. As shown in Section S4 of the supplementary material, the combined result generated by using the cleaned data with no ambiguity for the same sample f_{321} is given by

$$(e^*_{3,1}\wedge e^*_{2,2}\wedge e^*_{1,3})=\{(h_1,0.0055),(h_2,0.2076),(h_3,0.7869)\} \quad \text{(S3-2)}$$

where "*" means $e^*_{3,1}$, $e^*_{2,2}$ and $e^*_{1,3}$ are acquired from the cleaned data.

In Equation (S3-2), there is a high probability of 0.7869 for state h_3 and only a low probability of 0.2131 against it with no ambiguity. Although these results are not dramatically different from those of Equation (S3-1), they nevertheless provide an illusion about a higher degree in favor of h_3 with no ambiguity rather than what the data actually exhibits.

Note that the results of Equation (S3-1) were generated by means of recursive likelihood inference with prior distribution taken as total ignorance. This is appropriate for fault diagnosis in engineering systems where prior distributions about system states are difficult to estimate in general, a fault occurs with low probability, and prior distribution from sample data may not reflect the true prior condition of system states. In this rail track case study, not only can the prior distribution change with the change of the lengths of rail track sections where data was sampled, but the prior generated from the sample data of Table S3-2 (Table S3-3 or Table S3-4) does not represent the true prior condition of the rail track at the very location where the sample dataset f_{321} was taken. In fact, the data of Table S3-2 to Table S3-4 was recorded at a regular interval of every few meters along a section of the rail track when a train was moving on it, rather than from the same location where the sample record f_{321} was taken. While the data in Table S3-2 to Table S3-4 contain useful information about the relationship between the sensor readings and the irregularity of rail track, it does not

provide the information about the prior condition of the rail track at a particular location. To estimate the prior distribution of rail track irregularity at a specific location, many more datasets need to be generated from the very location, which may not be impossible but could be difficult, costly and impractical. Nevertheless, the impact of prior on fault diagnosis diminishes with accumulation of evidence acquired from more sensors.

S4. Fault diagnosis in railway track maintenance

S4.1 Evidence acquisition and combination

In this subsection, a case study on fault diagnosis for railway track is investigated to show that the *ER* process is a likelihood inference process and is capable of detecting system fault, even though the system fault is a low probability state in the sense that the sample data is severely imbalanced towards non-faulty states. This is significant because fault is often a small probability event in many engineering systems.

In railway systems, trains are guided on railway tracks by wheelsets (wheels and axles), which are connected to bogie frames via axle boxes and suspension. Due to such factors as heavy loads and uneven subgrade settlement, the geometric deformation of tracks can occur and is mainly expressed as vertical and lateral track irregularities. The abnormal vibration of train caused by track irregularities can lead to poor ride quality and even derailment. Therefore, track irregularity faults should be diagnosed and eliminated through maintenance to keep good ride quality and train safety. Rail track irregularities in conventional and high speed railways can be measured by track inspection vehicle such as GJ-4 and GJ-5. For example, a GJ-4 vehicle uses the inertial reference measurement method to calculate the vertical displacement (denoted as d_v) of irregularity by car-body-mounted accelerometers, displacement sensors, clinometers and gyroscopes. Although a GJ-4 vehicle can provide the precise estimation of vertical displacement, it needs very expensive clinometers and gyroscopes and requires specially designed structure for installation. This makes it difficult to use track inspection vehicles to conduct real-time monitoring for a large railway network.

As an alternative solution to this problem, irregularity fault identification and estimation can be done by using vibration acceleration data measured from the axle-boxes, bogies and car-bodies of in-service trains. The alternative solution can be applied on in-service trains using cheap accelerometers, so that real-time monitoring can be realized and the cost of measurement can be significantly reduced. As the distinctive signals of irregularity are hidden in the natural frequency of vehicle vibration, signal processing methods need to be used to extract frequency-domain features and identify irregularity faults.

According to Chinese railway line maintenance policy [1], track irregularity levels can be used as a specific standard for diagnosis and track maintenance. For example, when vehicle speed is limited in the interval [160, 200]km/h, dynamic management levels under the 42m waveband is given in Table S4-1, where Ir is the absolute value of d_v in millimeter (mm). Level I means that track is in good condition and only routine maintenance is required. For 5mm<Ir≤8mm (Level II), car-body vibration can discomfort passengers but is still tolerable from the maintenance point of view. If geometric deformation deteriorates further to Level III (8mm<Ir≤12mm), alarm must be generated and maintenance engineers have to do on-site repair as soon as possible. If Ir>12mm, serious defect occurs and poses a threat to train safety. In this case, a speed limit must be set immediately. Therefore, Level II is a transitional level from normal (Level I) condition (state) to abnormal (Level III) condition (state) or fault state as named in this paper.

Table S4-1 dynamic levels of vertical irregularity of track

(160 km/h~ 200 km/h)	Acceptance	Discomfort	Temporary repair	Speed limit
Level	I	II	III	IV
Standard(mm)	0≤Ir≤5	5<Ir≤8	8<Ir≤12	12<Ir

For railway safety, it is necessary to diagnose track irregularity to support the dynamic management and maintenance of rail tracks. In this subsection, we demonstrate how the *ER* inference process investigated in this paper can be used to diagnose track irregularity levels by using the data generated from the accelerometers mounted in the axle-box, bogie and car-body of a train, as shown in Figure S4-1, in comparison with Bayesian inference. In the next subsection, we will show that inaccuracy and ambiguity are inherent in such data and need to be treated with respect, instead of deleting inaccurate and missing datasets or imputing data under unrealistic assumptions.

Figure S4-1 shows the vertical vibration readings in time-domain recorded from a section of an operational railway line of about 2.357 kilometers. The data are measured by the accelerometers mounted in the axle-box, car-body and bogie of a GJ-4 vehicle and denoted by vr_1, vr_2 and vr_3, respectively. The vertical displacement d_v of the rail track is calculated by using the inertial reference measurement method. vr_1, vr_2 and vr_3 are sampled per 0.25m, so time step is t=1,…,T, with T being the number of total samples and T=(2.357/0.25)×10^3 = 9429.

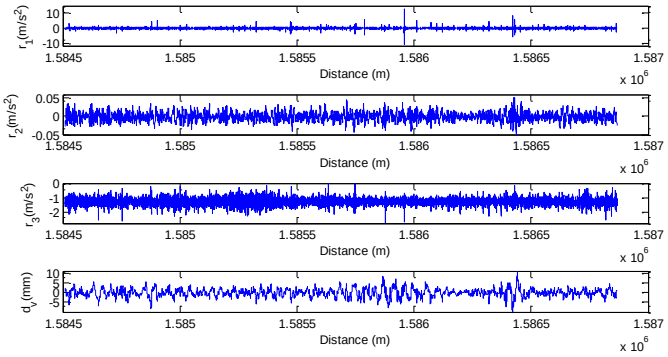

Figure S4-1 Vertical vibration readings and displacement in time-domain

At each step, the short-time Fourier transform is used to obtain the frequency amplitudes of acceleration with a window size of 5.25m. The mean values of the absolute amplitudes are denoted by $f_1(t)$, $f_2(t)$ and $f_3(t)$, respectively, and the absolute value of d_v by $Ir(t)$, as shown in Figure S4-2. Sample datasets are defined as S_1={$f_1(t)|t$=1, 2 ,…, T}, S_2={$f_2(t)|t$=1, 2,…, T},

$S_3=\{f_3(t)|t=1, 2,\ldots, T\}$, and $S_{Ir}=\{Ir(t)|t=1, 2,\ldots, T\}$, with $f_1(t)\in SI_1=[0.1705, 4.9399]$, $f_2(t)\in SI_2=[0.0012, 0.0207]$ and $f_3(t)\in SI_3=[0.4719,1.638]$. In this paper, the data sources f_1, f_2 and f_3 are assumed to be independent of each other because these accelerometers are installed on the different parts of a train separated by suspension and springs so that each accelerometer works on its own without affecting others.

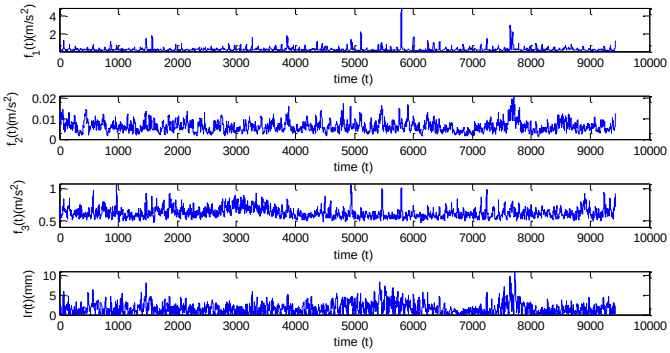

Figure S4-2 The mean values $f_1(t)$, $f_2(t)$, $f_3(t)$ and the absolute value $Ir(t)$

Since there is no sample such that $Ir(t)$>12mm (Level IV), in this subsection, the whole set of states is given by $\Theta=\{h_1(\text{I}), h_2(\text{II}), h_3(\text{III})\}$. For simplicity of discussion, SI_k (k=1, 2, 3) is uniformly divided into 5 subintervals or bins denoted by $e_{j,k}$ (j=1, 2, …, 5), and then the number of samples cast in $e_{j,k}$ is counted as shown in Table S4-2, Table S4-3 and Table S4-4. Note that while this casting approach is simple to generate frequency tables, it does not make full use of the information of a continuous variable and can cause approximation error in the process of generating frequency tables, leading to concern on accuracy in machine learning (ML). In ML, variables can be discretized continuously using the reference point approach [2].

Table S4-2 The casting result of sample $f_1(t)$

Frequency		$e_{1,1}$ [0.1075,1.074)	$e_{2,1}$ [1.074,2.0405)	$e_{3,1}$ [2.0405,3.007)	$e_{4,1}$ [3.007,3.9734)	$e_{5,1}$ [3.9734,4.9399]	Total observation
h_1	I:0<Ir≤5	9130	80	6	2	3	9221
h_2	II:5<Ir≤8	167	22	4	0	0	193
h_3	III:8≤Ir≤12	8	2	4	1	0	15
Total cast		9305	104	14	3	3	9429

Table S4-3 The casting result of sample $f_2(t)$

Frequency		$e_{1,2}$ [0.0012,0.0051)	$e_{2,1}$ [0.0051,0.009)	$e_{3,2}$ [0.009,0.0129)	$e_{4,2}$ [0.0129,0.0168)	$e_{5,2}$ [0.0168,0.0207]	Total observation
h_1	I:0<Ir≤5	3596	4537	933	149	6	9221
h_2	II:5<Ir≤8	37	98	42	16	0	193
h_3	III:8≤Ir≤12	0	3	4	1	7	15
Total cast		3633	4368	979	166	13	9429

Table S4-4 The casting result of sample $f_3(t)$

Frequency		$e_{1,3}$ [0.4719,0.5903)	$e_{2,3}$ [0.5903,0.7086)	$e_{3,3}$ [0.7086,0.827)	$e_{4,3}$ [0.827,0.9454)	$e_{5,3}$ [0.9454,1.0638]	Total observation
h_1	I:0<Ir≤5	3603	4428	1012	152	26	9221
h_2	II:5<Ir≤8	86	73	23	7	4	193
h_3	III:8≤Ir≤12	5	2	7	1	0	15
Total cast		3694	4503	1042	160	30	9429

According to the casting results, likelihood $c_{i,j,k}$ can be calculated as shown in Table S4-5, Table S4-6 and Table S4-7, to which the j^{th} piece of evidence $e_{j,k}$ of the data source S_k is expected to occur given that the i^{th} (i=1, 2, 3) state (h_i) is true. Then, Equation (14) in the paper is used to acquire the corresponding basic probability $p_{i,j,k}$ that evidence $e_{j,k}$ points to the i^{th} state h_i, as listed in Table S4-8, Table S4-9 and Table S4-10. Since the missing data is deleted in this subsection, all data are cast into h_1(I), h_2(II) or h_3(III) in this casting process, without any data cast to Θ or any of its other subsets.

Table S4-5 The likelihood $c_{i,j,1}$ from data source S_1

Likelihood	$e_{1,1}$	$e_{2,1}$	$e_{3,1}$	$e_{4,1}$	$e_{5,1}$
h_1	$c_{1,1,1}$ = 0.9901	$c_{1,2,1}$ = 0.0087	$c_{1,3,1}$ = 0.0007	$c_{1,4,1}$ = 0.0002	$c_{1,5,1}$ = 0.0003
h_2	$c_{2,1,1}$ = 0.8653	$c_{2,2,1}$ = 0.1140	$c_{2,3,1}$ = 0.0207	$c_{2,4,1}$ = 0.0000	$c_{2,5,1}$ = 0.0000
h_3	$c_{3,1,1}$ = 0.5333	$c_{3,2,1}$ = 0.1333	$c_{3,3,1}$ = 0.2667	$c_{3,4,1}$ = 0.0667	$c_{3,5,1}$ = 0.0000

Table S4-6 The likelihood $c_{i,j,2}$ from data source S_2

Likelihood	$e_{1,2}$	$e_{2,2}$	$e_{3,2}$	$e_{4,2}$	$e_{5,2}$
h_1	$c_{1,1,2}$ = 0.3900	$c_{1,2,2}$ = 0.4920	$c_{1,3,2}$ = 0.1012	$c_{1,4,2}$ = 0.0162	$c_{1,5,2}$ = 0.0007
h_2	$c_{2,1,2}$ = 0.1917	$c_{2,2,2}$ = 0.5078	$c_{2,3,2}$ = 0.2176	$c_{2,4,2}$ = 0.0829	$c_{2,5,2}$ = 0.0000
h_3	$c_{3,1,2}$ = 0.0000	$c_{3,2,2}$ = 0.2000	$c_{3,3,2}$ = 0.2667	$c_{3,4,2}$ = 0.0667	$c_{3,5,2}$ = 0.4667

Table S4-7 The likelihood $c_{i,j,3}$ from data source S_3

Likelihood	$e_{1,3}$	$e_{2,3}$	$e_{3,3}$	$e_{4,3}$	$e_{5,3}$
h_1	$c_{1,1,3}$ = 0.3907	$c_{1,2,3}$ = 0.4802	$c_{1,3,3}$ = 0.1097	$c_{1,4,3}$ = 0.0165	$c_{1,5,3}$ = 0.0028
h_2	$c_{2,1,3}$ = 0.4456	$c_{2,2,3}$ = 0.3782	$c_{2,3,3}$ = 0.1192	$c_{2,4,3}$ = 0.0363	$c_{2,5,3}$ = 0.0207
h_3	$c_{3,1,3}$ = 0.3333	$c_{3,2,3}$ = 0.1333	$c_{3,3,3}$ = 0.4667	$c_{3,4,3}$ = 0.0667	$c_{3,5,3}$ = 0.0000

In order to demonstrate the equivalence and difference between Bayesian inference and the *ER* process, data record $f_{321}=[f_1(t), f_2(t), f_3(t), Ir(t)]=[2.8992, 0.0088, 0.5371, 9.195]$ is chosen from the datasets as a test record, and analyzed using the following three alternative inference methods based on both the *ER* process and Bayesian inference. From the sample data (I_0) shown in Table S4-2 to Table S4-4, it is clear that test record f_{321} activates three pieces of evidence: $e_{3,1}$, $e_{2,2}$ and $e_{1,3}$ as the values of $f_1(t)$, $f_2(t)$ and $f_3(t)$ are in those three bins, respectively. It is also noted that this sample is taken at fault state h_3 as $Ir(t)$ = 9.195. The corresponding likelihoods can then be acquired

from Table S4-5 to Table S4-7. In addition, the prior probabilities can be generated directly from the last column of Table S4-2 by $[p_{10}, p_{20}, p_{30}]^T$=[0.9779, 0.0205, 0.0016] ^{T}.

Table S4-8 Basic probability $p_{i,j,1}$ from data source S_1

Probability	$e_{1,1}$	$e_{2,1}$	$e_{3,1}$	$e_{4,1}$	$e_{5,1}$
h_1	$p_{1,1,1}$ = 0.4145	$p_{1,2,1}$ = 0.0339	$p_{1,3,1}$ = 0.0022	$p_{1,4,1}$ = 0.0032	$p_{1,5,1}$ = 1
h_2	$p_{2,1,1}$ = 0.3622	$p_{2,2,1}$ = 0.4453	$p_{2,3,1}$ = 0.0720	$p_{2,4,1}$ = 0.0000	$p_{2,5,1}$ = 0
h_3	$p_{3,1,1}$ = 0.2233	$p_{3,2,1}$ = 0.5208	$p_{3,3,1}$ = 0.9258	$p_{3,4,1}$ = 0.9968	$p_{3,5,1}$ = 0

Table S4-9 Basic probability $p_{i,j,2}$ from data source S_2

Probability	$e_{1,2}$	$e_{2,2}$	$e_{3,2}$	$e_{4,2}$	$e_{5,2}$
h_1	$p_{1,1,2}$ = 0.6704	$p_{1,2,2}$ = 0.4101	$p_{1,3,2}$ = 0.1728	$p_{1,4,2}$ = 0.0975	$p_{1,5,2}$ = 0.0014
h_2	$p_{2,1,2}$ = 0.3296	$p_{2,2,2}$ = 0.4232	$p_{2,3,2}$ = 0.3717	$p_{2,4,2}$ = 0.5002	$p_{2,5,2}$ = 0.0000
h_3	$p_{3,1,2}$ = 0.0000	$p_{3,2,2}$ = 0.1667	$p_{3,3,2}$ = 0.4555	$p_{3,4,2}$ = 0.4023	$p_{3,5,2}$ = 0.9986

Table S4-10 Basic probability $p_{i,j,3}$ from data source S_3

Probability	$e_{1,3}$	$e_{2,3}$	$e_{3,3}$	$e_{4,3}$	$e_{5,3}$
h_1	$p_{1,1,3}$ = 0.334	$p_{1,2,3}$ = 0.4842	$p_{1,3,3}$ = 0.1578	$p_{1,4,3}$ = 0.138	$p_{1,5,3}$ = 0.1198
h_2	$p_{2,1,3}$ = 0.381	$p_{2,2,3}$ = 0.3814	$p_{2,3,3}$ = 0.1713	$p_{2,4,3}$ = 0.3037	$p_{2,5,3}$ = 0.8802
h_3	$p_{3,1,3}$ = 0.285	$p_{3,2,3}$ = 0.1344	$p_{3,3,3}$ = 0.6709	$p_{3,4,3}$ = 0.5583	$p_{3,5,3}$ = 0.0000

Method 1: First of all, all the three pieces of evidence $e_{3,1}$, $e_{2,2}$ and $e_{1,3}$ are combined with the prior distribution e_0 by applying the *ER* rule with w_i=1 (i=0, 1, 2, 3) assumed. Equation (13) is used to calculate the combined result as follows, where $p_{h_k}(e_0 \wedge e_{3,1} \wedge e_{2,2} \wedge e_{1,3})$ is the probability that $e_{3,1}$, $e_{2,2}$, $e_{1,3}$ and e_0 jointly support the k^{th} state h_k (k=1, 2, 3).

$$[p_{h_1}(e_0 \wedge e_{3,1} \wedge e_{2,2} \wedge e_{1,3}),\ p_{h_2}(e_0 \wedge e_{3,1} \wedge e_{2,2} \wedge e_{1,3}),\ p_{h_3}(e_0 \wedge e_{3,1} \wedge e_{2,2} \wedge e_{1,3})]^T$$
$$=[0.4961,\ 0.3892,\ 0.1147]^T$$

Method 2: According to the conventional Bayesian inference as discussed in Section 2.1, the combined result of evidence e_0 and evidence $e_{3,1}$ can be generated by

$$p(h_i \mid e_{3,1}, I_0) = \frac{p(e_{3,1} \mid h_i, I_0)\, p(h_i \mid I_0)}{\sum_{i=1}^{3} p(e_{3,1} \mid h_i, I_0)\, p(h_i \mid I_0)} = \frac{c_{i,3,1} \times p_{i,0}}{\sum_{i=1}^{3} c_{i,3,1} \times p_{i,0}}$$

Similarly, take the above combined result as new *prior* probability, and continue to combine it with evidence $e_{2,2}$ and evidence $e_{1,3}$ recursively, leading to the following result.

$$\begin{bmatrix} p(h_1 \mid I_0, e_{3,1}, e_{2,2}, e_{1,3}),\ p(h_2 \mid I_0, e_{3,1}, e_{2,2}, e_{1,3}), \\ p(h_3 \mid I_0, e_{3,1}, e_{2,2}, e_{1,3}) \end{bmatrix}^T$$
$$=[0.4961,\ 0.3892,\ 0.1147]^T$$

Method 3: The *ER* rule is employed to combine the three pieces of evidence but take the prior as total ignorance. The combined result is given by

$$\begin{bmatrix} p_{h_1}(e_{3,1} \wedge e_{2,2} \wedge e_{1,3}),\ p_{h_2}(e_{3,1} \wedge e_{2,2} \wedge e_{1,3}), \\ p_{h_3}(e_{3,1} \wedge e_{2,2} \wedge e_{1,3}) \end{bmatrix}^T$$
$$=[0.0055,\ 0.2076,\ 0.7869]^T$$

The above analyses show that Method 1 and Method 2 lead to the same combined result, just as expected by Theorem 1; however, both methods fail to identify faulty state h_3 for this sample as the posterior probability for h_3 is as low as 0.1147. On the other hand, the combined probability generated using Method 3 with the *prior* taken as total ignorance leads to a credible result in the sense that the probability for h_3 is as high as 0.7869. This is a satisfactory result and reflects the true track irregularity level of this test sample to a great extent. This is significant because there are only 15 samples pointing to h_3 in all the sample datasets. Compared with the other states, h_3 is a very small probability state as revealed by the sample data. If the *prior* distribution e_0 generated from the imbalanced sample data is taken into account, which points to the normal state (h_1) to a great extent, it will have significant impact on the combined result for such a small probability event, leading to the missed identification of faulty state, just as shown by this example. It is a common situation in system fault diagnosis that a faulty state for a seemingly normal system is often a small probability event. As such, it makes sense to apply likelihood inference methods for fault diagnosis.

However, from the above results it would be inappropriate to conclude that no prior distribution should be used in inference for fault diagnosis. Rather, one should check whether a prior distribution can reflect true prior system conditions or not. In the above analysis, for example, the prior generated from Table S4-9 does not represent the true prior condition of rail track at the location where the test dataset f_{321} was recorded. In fact, the data of Table S4-9 are imbalanced because they were recorded from different locations when the inspection train was moving along a section of rail track, rather than from the same location where f_{321} was taken. While the dataset of Table S4-9 is believed to hold useful information about the relationship between the sensor readings and the irregularity of rail track, it may not provide adequate information about the prior condition of rail track at a particular location. To estimate the prior distribution of rail track irregularity at a specific location, many more datasets need to be generated from the very location, which may not be impossible but could be difficult, costly and impractical.

S4.2 Evaluating the weight of evidence

In the above analysis, the evidence acquired from each of the three sensors is assumed to be fully reliable. However, this is not the case as the quality of data collected from the three sensors is different, as shown in Figure S4-2. There is therefore a need to estimate the weight of evidence acquired from these data sources. In principle, the weight of evidence acquired from a data source or a sensor represents the importance of the role that the data source plays to provide correct identification of rail track irregularities. In this context, we need to consider two factors to estimate the weight of evidence in this case study. One is the weight of the information source f_k itself, and the other is the precision of the partitioned evidence intervals. It is suggested that the more important an information source is, the more exactly it follows the vertical irregularities. That is, the higher the vertical displacement, the larger f_k, and vice versa. Hence, we firstly define the relative changes of the readings $f_k(t)$ and $Ir(t)$ as follows.

$$Cf_k(t)=\frac{f_k(t)}{\max_t(f_k(t))} \tag{S4-1}$$

$$CIr(t)=\frac{Ir(t)}{\max_t(Ir(t))} \tag{S4-2}$$

The ability of f_k following Ir can then be described by

$$af_k=\sum_{t=1}^{T}\left|CIr(t)\text{-}Cf_k(t)\right| \tag{S4-3}$$

Obviously, the smaller af_k is, the more accurately f_k reflects the variation trends of Ir. As such, the weight of information source f_k can be defined as follows.

$$Rf_k=\frac{\min_i(af_i)}{af_k} \tag{S4-4}$$

Equation (S4-4) implies that f_k is the most important, with $Rf_k=1$, if $\min_i(af_i)=af_k$. The weight of the other sources is measured by comparison with the most important one.

On the other hand, evidence set $E^k=\{e_j^k \mid j=1,2,...,5\}$ for each data source was constructed by uniformly partitioning the interval $SI_k=[lb_k, ub_k]$, where $lb_k=\min_t(f_k(t))$ and $ub_k=\max_t(f_k(t))$ are the lower and upper bounds of $f_k(t)$. In practice, $f_k(t)$ is an accurate reading but has $\pm\delta\%$ observation error. To take into account the error, we add $\delta\%$ or $-\delta\%$ perturbation to each reading in dataset S_k and then calculate the number of those noisy readings falling outside of SI_k, denoted as T_k. If those noisy readings fall outside of the discussed domain SI_k, data source f_k will not be regarded as fully reliable. Therefore, the weight of information source f_k, i.e. evidence set E^k derived from observation error, can be defined as follows.

$$Rn_k=\left(T\text{-}T_k\right)/T \tag{S4-5}$$

Finally, the overall weight of evidence set E^k is synthesized as follows.

$$w_k=Rf_k\times Rn_k \tag{S4-6}$$

S4.3 Evidence acquisition and combination with missing data taken into account

In the case study shown in Section *S4.1*, incomplete data were not taken into account in the comparison study of Bayesian inference with the equivalent likelihood inference based on the *ER* Rule. In fault diagnosis for railway tracks, it is common to face missing data, because acceleration data is gathered from sensors installed in fast moving trains and such sensors can fail to record data from time to time.

In this case study, there are a total number of 880 incomplete sample datasets, which were not considered in the analysis of Section *S.1*. Table S3-1 shows all the cases where incomplete data were collected from different accelerometers mounted in different parts of the train.

After taking these incomplete sample datasets into account, we can construct new data casting results, as shown in Table S4-11 to Table S4-16, from which a set of new evidence can be acquired. In this situation, we also choose the data record f_{321} =[$f_1(t)$, $f_2(t)$, $f_3(t)$, $Ir(t)$] = [2.8992, 0.0088, 0.5371, 9.195] as the test record, and use the *ER* rule to combine the three pieces of the observed evidence $e_{3,1}$, $e_{2,2}$ and $e_{1,3}$. The weights of evidence related to data inaccuracy are calculated by Rn_1=0.9243, Rn_2=0.9437, Rn_3=0.9243 using Equation (S4-5) for $\delta\%$ =5%. The synthesized weights are given by Equation (S4-6) as w_1=0.9237， w_2=0.5311 and w_2=0.2068. We then get the final combined result as follows.

$$\begin{bmatrix} p_{h_1}(e_{3,1}\wedge e_{2,2}\wedge e_{1,3}),\ p_{h_2}(e_{3,1}\wedge e_{2,2}\wedge e_{1,3}), \\ p_{h_3}(e_{3,1}\wedge e_{2,2}\wedge e_{1,3}),\ p_{\Theta}(e_{3,1}\wedge e_{2,2}\wedge e_{1,3}) \end{bmatrix}^T$$
$$=\left[0.0458,\ 0.1148,\ 0.7611,\ 0.0783\right]^T \tag{S4-7}$$

Note that the above result is generated with the prior taken as total ignorance. The weight of the above final combined result is given by $w=0.9565$.

Table S4-11 Casting result of sample $f_1(t)$ with incomplete data

Frequency		$e_{1,1}$ [0.1075, 1.074)	$e_{2,1}$ [1.074, 2.0405)	$e_{3,1}$ [2.0405, 3.007)	$e_{4,1}$ [3.007, 3.9734)	$e_{5,1}$ [3.9734, 4.9399]	$e_{\Theta,1}$ Unknown	Total observation
h_1	I:0<Ir≤5	9230	80	6	2	3	180	9501
h_2	II:5<Ir≤8	167	22	4	0	0	19	212
h_3	III:8≤Ir≤12	8	2	4	1	0	1	16
Θ	Unknown	404	26	14	6	0	130	580
Total cast		9809	130	28	9	3	330	10309

Table S4-12 Casting result of sample $f_2(t)$ with incomplete data

Frequency		$e_{1,2}$ [0.0012, 0.0051)	$e_{2,2}$ [0.0051, 0.009)	$e_{3,2}$ [0.009,0.0129)	$e_{4,2}$ [0.0129, 0.0168)	$e_{5,2}$ [0.0168, 0.0207]	$e_{\Theta,2}$ Unknown	Total observation
h_1	I:0<Ir≤5	3842	4571	933	149	6	0	9501
h_2	II:5<Ir≤8	40	111	45	16	0	0	212
h_3	III:8≤Ir≤12	1	3	4	1	7	0	16
Θ	Unknown	258	199	33	0	0	90	580
Total cast		4141	4884	1015	166	13	90	10309

Table S4-13 Casting result of sample $f_3(t)$ with incomplete data

Frequency		$e_{1,3}$ [0.4719, 0.5903)	$e_{2,3}$ [0.5903, 0.7086)	$e_{3,3}$ [0.7086, 0.827)	$e_{4,3}$ [0.827,0.9454)	$e_{5,3}$ [0.9454, 1.0638]	$e_{\Theta,3}$ Unknown	Total observation
h_1	I:0<Ir≤5	3654	4464	1015	152	26	190	9501
h_2	II:5<Ir≤8	90	79	23	7	4	9	212
h_3	III:8≤Ir≤12	5	2	7	1	0	1	16
Θ	Unknown	335	118	26	11	0	90	580
Total cast		4084	4663	1071	171	30	290	10309

Table S4-14 Probability $p_{i,j,1}$ of incomplete data source S_1

Probability	$e_{1,1}$	$e_{2,1}$	$e_{3,1}$	$e_{4,1}$	$e_{5,1}$	$e_{\Theta,1}$
h_1	$p_{1,1,1}$ = 0.3287	$p_{1,2,1}$ = 0.0299	$p_{1,3,1}$ = 0.0022	$p_{1,4,1}$ = 0.0029	$p_{1,5,1}$ = 1	$p_{1,\Theta,1}$ = 0.0479
h_2	$p_{2,1,1}$ = 0.2665	$p_{2,2,1}$ = 0.3680	$p_{2,3,1}$ = 0.0643	$p_{2,4,1}$ = 0.0000	$p_{2,5,1}$ = 0	$p_{2,\Theta,1}$ = 0.2268
h_3	$p_{3,1,1}$ = 0.1692	$p_{3,2,1}$ = 0.4432	$p_{3,3,1}$ = 0.8514	$p_{3,4,1}$ = 0.8555	$p_{3,5,1}$ = 0	$p_{3,\Theta,1}$ = 0.1581
Θ	$p_{\Theta,1,1}$ = 0.2357	$p_{\Theta,2,1}$ = 0.1590	$p_{\Theta,3,1}$ = 0.0822	$p_{\Theta,4,1}$ = 0.1416	$p_{\Theta,5,1}$ = 0	$p_{\Theta,\Theta,1}$ = 0.5671

Table S4-15 Probability $p_{i,j,2}$ of incomplete data source S_2

Probability	$e_{1,2}$	$e_{2,2}$	$e_{3,2}$	$e_{4,2}$	$e_{5,2}$	$e_{\Theta,2}$
h_1	$p_{1,1,2}$ = 0.3675	$p_{1,2,2}$ = 0.3134	$p_{1,3,2}$ = 0.1591	$p_{1,4,2}$ = 0.1021	$p_{1,5,2}$ = 0.001	$p_{1,\Theta,2}$ = 0
h_2	$p_{2,1,2}$ = 0.1715	$p_{2,2,2}$ = 0.3410	$p_{2,3,2}$ = 0.3438	$p_{2,4,2}$ = 0.4912	$p_{2,5,2}$ = 0.000	$p_{2,\Theta,2}$ = 0
h_3	$p_{3,1,2}$ = 0.0568	$p_{3,2,2}$ = 0.1221	$p_{3,3,2}$ = 0.4049	$p_{3,4,2}$ = 0.4068	$p_{3,5,2}$ = 0.999	$p_{3,\Theta,2}$ = 0
Θ	$p_{\Theta,1,2}$ = 0.4042	$p_{\Theta,2,2}$ = 0.2235	$p_{\Theta,3,2}$ = 0.0922	$p_{\Theta,4,2}$ = 0.0000	$p_{\Theta,5,2}$ = 0.000	$p_{\Theta,\Theta,2}$ = 1

Table S4-16 Probability $p_{i,j,3}$ of incomplete data source S_3

Probability	$e_{1,3}$	$e_{2,3}$	$e_{3,3}$	$e_{4,3}$	$e_{5,3}$	$e_{\Theta,3}$
h_1	$p_{1,1,3}$ = 0.2263	$p_{1,2,3}$ = 0.4013	$p_{1,3,3}$ = 0.1531	$p_{1,4,3}$ = 0.1226	$p_{1,5,3}$ = 0.13	$p_{1,\Theta,3}$ = 0.0714
h_2	$p_{2,1,3}$ = 0.2498	$p_{2,2,3}$ = 0.3182	$p_{2,3,3}$ = 0.1555	$p_{2,4,3}$ = 0.2531	$p_{2,5,3}$ = 0.87	$p_{2,\Theta,3}$ = 0.1516
h_3	$p_{3,1,3}$ = 0.1839	$p_{3,2,3}$ = 0.1068	$p_{3,3,3}$ = 0.6271	$p_{3,4,3}$ = 0.4790	$p_{3,5,3}$ = 0.00	$p_{3,\Theta,3}$ = 0.2231
Θ	$p_{\Theta,1,3}$ = 0.3399	$p_{\Theta,2,3}$ = 0.1737	$p_{\Theta,3,3}$ = 0.0643	$p_{\Theta,4,3}$ = 0.1453	$p_{\Theta,5,3}$ = 0.00	$p_{\Theta,\Theta,3}$ = 0.5539

From the above result, it can be concluded that the probability for the fault state h_3 is given by $p(h_3) = 0.7611$, that against h_3 by $p^c(h_3) = 0.0458 + 0.1148 = 0.1606$, and the probability of unknown about h_3 by $p^r(h_3) = 0.0783$. In fault diagnosis, this is a satisfactory conclusion for the correct prediction of the fault state h_3 for the sample in question. Compared with the results generated in Section 2.1, this is a more realistic and credible conclusion because in this case the incomplete data was duly taken into account rather than being neglected as in the previous analyses.

References

[1] Ministry of Railways of the People's Republic of China: Railway Line Repair Rules. Chinese Railway Press, Beijing (2006).

[2] X. B. Xu, J. Zheng, J. B. Yang, D. L. Xu and Y. W. Chen, "Data classification using evidence reasoning rule", Knowledge-Based Systems, Vol.116, 2017, pp.144–151.