Propagation of Interval and Probabilistic Uncertainty in Cyberinfrastructure-Related Data Processing and Data Fusion

Christian Servin

Computational Science Program
University of Texas at El Paso
El Paso, Texas 79968, USA
christians@utep.edu

1. Need for Data Processing and Data Fusion

- For many quantities y, it is not easy (or even impossible) to measure them directly.
- Instead, we measure related quantities $x_1, \ldots, x_n$, and use the known relation $y = f(x_1, \ldots, x_n)$ to estimate y.
- Such *data processing* is especially important for cyberinfrastructure-related heterogenous data.
- Example of heterogenous data – geophysics:
  - first-arrival passive (from actual earthquakes) and active seismic data (from seismic experiments);
  - gravity data;
  - surface waves, etc.
- Before we start processing data, we need to first *fuse* data points corresponding to the same quantity.

2. Need to Take Uncertainty into Consideration

- The result $\widetilde{x}$ of a measurement is usually somewhat different from the actual (unknown) value x.
- Usually, the manufacturer of the measuring instrument (MI) gives us a bound Δ on the measurement error:

$$|\Delta x| \le \Delta, \text{ where } \Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$$

- Once we know the measurement result $\widetilde{x}$, we can conclude that the actual value x is in $[\widetilde{x} - \Delta, \widetilde{x} + \Delta]$.
- In some situations, we also know the probabilities of different values $\Delta x \in [-\Delta, \Delta]$.
- In this case, we can use statistical techniques.
- However, often, we do not know these probabilities; we only know that x is in the interval $\mathbf{x} \stackrel{\text{def}}{=} [\widetilde{x} - \Delta, \widetilde{x} + \Delta]$.
- In this case, we need to process this interval data.

3. Measurement Uncertainty: Traditional Approach

- Usually, a meas. error $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$ is subdivided into *random* and *systematic* components $\Delta x = \Delta x_s + \Delta x_r$:
  - the systematic error component Δx_s is usually defined as the expected value $\Delta x_s = E[\Delta x]$, while
  - the random error component is usually defined as the difference $\Delta x_r \stackrel{\text{def}}{=} \Delta x - \Delta x_s$.
- The random errors Δx_r corresponding to different measurements are usually assumed to be independent.
- For Δx_s, we only know the upper bound Δ_s s.t. $|\Delta x_s| \le \Delta_s$, i.e., that Δx_s is in the *interval* $[-\Delta_s, \Delta_s]$.
- Because of this fact, *interval computations* are used for processing the systematic errors.
- Δx_r is usually characterized by the corr. probability distribution (usually Gaussian, with known σ).

4. Expert Estimates and Fuzzy Data

- There is no guarantee of expert's accuracy.
- We can only provide bounds which are valid with some degree of certainty.
- This degree of certainty is usually described by a number from the interval $[0, 1]$.
- So, for each $\beta \in [0, 1]$, we have an interval $\mathbf{x}(\alpha)$ containing the actual value x with certainty $\alpha = 1 - \beta$.
- The larger certainty we want, the broader should the corresponding interval be.
- So, we get a nested family of intervals corresponding to different values α.
- *Alternative:* for each x, describe the largest α for which x is in $\mathbf{x}(\alpha)$; this α_{largest} is a membership function $\mu(x)$.

5. How to Propagate Uncertainty in Data Processing

- We know that $y = f(x_1, \ldots, x_n)$.
- We estimate y based on the approximate values $\widetilde{x}_i$ as $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.
- Since $\widetilde{x}_i \neq x_i$, we get $\widetilde{y} \neq y$; it is desirable to estimate the approximation error $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$.
- Usually, measurements are reasonably accurate, i.e., measurement errors $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$ are small.
- Thus, we can keep only linear terms in Taylor expansion: $\Delta y = \sum\limits_{i=1}^{n} C_i \cdot \Delta x_i$, where $C_i = \dfrac{\partial f}{\partial x_i}$.
- For systematic error, we get a bound $\sum\limits_{i=1}^{n} |C_i| \cdot \Delta_{si}$.
- For random error, we get $\sigma^2 = \sum\limits_{i=1}^{n} C_i^2 \cdot \sigma_i^2$.

6. How to Propagate Uncertainty in Data Fusion: Case of Probabilistic Uncertainty

- *Reminder:* we have several estimates $\widetilde{x}^{(1)}, \ldots, \widetilde{x}^{(n)}$ of the same quantity x.
- *Data fusion:* we combine these estimates into a single estimate $\widetilde{x}$.
- *Case:* each estimation error $\Delta x^{(i)} \stackrel{\text{def}}{=} \widetilde{x}^{(i)} - x$ is normally distributed with 0 mean and known st. dev. $\sigma^{(i)}$.
- *How to combine:* use Least Squares, i.e., find $\widetilde{x}$ that minimizes $\sum\limits_{i=1}^{n} \dfrac{(\widetilde{x}^{(i)} - \widetilde{x})^2}{2 \cdot (\sigma^{(i)})^2}$;
- *Solution:* $\widetilde{x} = \dfrac{\sum\limits_{i=1}^{n} \widetilde{x}^{(i)} \cdot (\sigma^{(i)})^{-2}}{\sum\limits_{i=1}^{n} (\sigma^{(i)})^{-2}}.$

7. Data Fusion: Case of Interval Uncertainty

- In some practical situations, the value x is known with interval uncertainty.
- This happens, e.g., when we only know the upper bound $\Delta^{(i)}$ on each estimation error $\Delta x^{(i)}$: $|\Delta x^{(i)}| \le \Delta_i$.
- In this case, we can conclude that $|x - \widetilde{x}^{(i)}| \le \Delta^{(i)}$, i.e., that $x \in \mathbf{x}^{(i)} \stackrel{\text{def}}{=} [\widetilde{x}^{(i)} - \Delta^{(i)}, \widetilde{x}^{(i)} + \Delta^{(i)}]$.
- Based on each estimate $\widetilde{x}^{(i)}$, we know that the actual value x belongs to the interval $\mathbf{x}^{(i)}$.
- Thus, we know that the (unknown) actual value x belongs to the intersection of these intervals:

$$\mathbf{x} \stackrel{\text{def}}{=} \bigcap_{i=1}^{n} \mathbf{x}^{(i)} = [\max(\widetilde{x}^{(i)} - \Delta^{(i)}), \min(\widetilde{x}^{(i)} + \Delta^{(i)})].$$

8. Propagation of Uncertainty: Challenges

- In the *ideal world:*
  - we should have an *accurate* description of data uncertainty;
  - based on this description, we should use *well-justified* and *efficient* algorithms to propagate uncertainty.
- In *practice*, we are often not yet in this ideal situation:
  - the description of uncertainty is often only *approximate*,
  - the algorithms for uncertainty propagation are often *heuristics*, i.e., not well-justified, and
  - the algorithms for uncertainty propagation are often *not* very computationally *efficient*.

9. What We Do in This Dissertation

- In Chapter 2, we show that the traditional idea of *random* and *systematic* components is an *approximation*:
  - we also need *periodic* components;
  - this is important in environmental studies.
- In Chapter 3, on the example of a fuzzy *heuristic*, we show how a heuristic can be *formally justified*.
- In Ch. 4, we show how to process more *efficiently*; e.g.:
  - first, we process data type-by-type;
  - then, we fuse the resulting models.
- All these results assume that we have a good description of the uncertainty of the original data.
- In practice, we often need to extract this info from the data; extraction methods are described in Ch. 5.

10. Chapter 2: Towards More Accurate Description of Uncertainty

- Often, the differences $r = \Delta x - s$ corr. to nearby times are strongly correlated.
- For example, meteorological sensors may have daytime or nighttime biases, or winter and summer biases.
- To capture this correlation, environmental scientists proposed a semi-heuristic 3-component model of Δx.
- In this model, the difference $\Delta x - \Delta x_s$ is represented as a combination of:
  - a "truly random" error Δx_r (which is independent from one measurement to another), and
  - a new "periodic" component Δx_p.
- We provide a theoretical explanation for this heuristic three-component model.

11. Error Components: Analysis

- We want to represent measurement error $\Delta x(t)$ as a linear combination of several components.
- We consider the most detailed level of granularity, w/each component determined by finitely many parameters c_i.
- Each component is thus described by a finite-dimensional linear space

$$L = \{c_1 \cdot x_1(t) + \ldots + c_n \cdot x_n(t) : c_1, \ldots, c_n \in \mathbb{R}\}.$$

- In most applications, signals are smooth and bounded, so we assume that $x_i(t)$ is smooth and bounded.
- Finally, for a long series of observations, we can choose a starting point arbitrarily: $t \to t + t_0$.
- It is reasonable to require that this change keeps us within the same component, i.e.,

$$x(t) \in L \Rightarrow x(t + t_0) \in L.$$

12. Error Components: Main Result

- A function $x(t)$ of one variable is called *bounded* if

$$\exists M\,\forall t\,(|x(t)| \le M).$$

- We say that a class F of functions of one variable is *shift-invariant* if

$$\forall x(t)\,(x(t) \in F \Rightarrow \forall t_0\,(x(t+t_0) \in F)).$$

- By an *error component* we mean a shift-invariant finite-dimensional linear space of functions

$$L = \{c_1 \cdot x_1(t) + \ldots + c_n \cdot x_n(t) : c_i \in \mathbb{R}\}.$$

- **Theorem:** *Every error component is a linear combination of the functions*

$$x(t) = \sin(\omega \cdot t) \text{ and } x(t) = \cos(\omega \cdot t).$$

13. Practical Conclusions

- Let f be the measurements frequency (how many measurements we perform per unit time).
- When $\omega \ll f$, the values $\cos(\omega \cdot t)$ and $\sin(\omega \cdot t)$ practically do not change with time.
- Indeed, the change period is much larger than the usual observation period.
- Thus, we can identify such low-frequency components with *systematic* error component.
- When $\omega \gg f$, the phases of the values $\cos(\omega \cdot t_i)$ and $\cos(\omega \cdot t_{i+1})$ differ a lot.
- For all practical purposes, the resulting values of cosine or sine functions are independent.
- Thus, high-frequency components can be identified with *random* error component.

14. Practical Conclusions (cont-d)

- *Result:* every error component is a linear combination of $\cos(\omega \cdot t)$ and $\sin(\omega \cdot t)$.
- *Notation:* let f be the measurements frequency (how many measurements we perform per unit time).
- *Reminder:*
  - we can identify low-frequency components ($\omega \ll f$) with *systematic* error component;
  - we can identify high-frequency ones ($\omega \gg f$) with *random* error component.
- *Easy to see:* all other error components $\cos(\omega \cdot t)$ and $\sin(\omega \cdot t)$ are periodic.
- *Conclusion:* we have indeed justified to the semi-empirical 3-component model of measurement error.

15. Chapter 3: Towards Justification of Heuristic Techniques for Processing Uncertainty

- As we have mentioned, some methods for processing uncertainty are *heuristic*.
- Such methods lack justification and are, therefore, less reliable.
- Usually, techniques for processing interval and probabilistic uncertainty are well-justified.
- However, many techniques for processing expert (fuzzy) data are still heuristic.
- In Chapter 3:
  - we consider a practically efficient heuristic fuzzy technique for decision making under uncertainty;
  - we show how this heuristic can be formally justified.

16. Traditional Approach to Decision Making: Reminder

- The quality of each possible alternative is characterized by the values of several quantities.
- For example, when we buy a car, we are interested in its cost, its energy efficiency, its power, size, etc.
- For each of these quantities, we usually have some desirable range of values.
- Often, there are several different alternatives all of which satisfy all these requirements.
- The traditional approach assumes that there is an objective function that describes the user's preferences.
- We then select an alternative with the largest possible value of this objective function.

17. Traditional Approach to Decision Making: Limitations

- The traditional approach to decision making assumes:
  - that the user knows exactly what he or she wants — i.e., knows the objective function – and
  - that the user also knows exactly what he or she will get as a result of each possible decision.
- In practice, the user is often uncertain:
  - the user is often uncertain about his or her own preferences, and
  - the user is often uncertain about possible consequences of different decisions.
- It is therefore desirable to take this uncertainty into account when we describe decision making.

18. Fuzzy Target Approach (Huynh-Nakamori)

- For each numerical characteristic of a possible decision, we form two fuzzy sets:
  - $\mu_i(x)$ describing the users' ideal value;
  - $\mu_a(x)$ describing the users' impression of the actual value.
- For example, a person wants a well done steak, and the steak comes out as medium well done.
- In this case, $\mu_i(x)$ corresponds to "well done", and $\mu_a(x)$ to "medium well done".
- The simplest "and"-operation (t-norm) is $\min(a, b)$; so, the degree to which x is both actual *and* desired is

$$\min(\mu_a(x), \mu_i(x)).$$

- The degree to which there exists x which is both possible and desired is $d = \max_x \min(\mu_a(x), \mu_i(x))$.

19. Fuzzy Target Approach: Successes and Remaining Problems

- The above approach works well in many applications.
- *Example:* it predicts how customers select a hand-crafted souvenir when their ideal ones is not available.
- *Problem:* this approach is heuristic, it is based on selecting:
  - the simplest possible membership function and
  - the simplest possible "and"- and "or"-operations.
- Interestingly, we get *better* predictions than with more complex membership functions and "and"-operations.
- In Chapter 3, we provide a justification for the above semi-heuristic target-based fuzzy decision procedure.

20. Chapter 4: Towards More Computationally Efficient Techniques for Processing Uncertainty

- *Fact:* computations often take a lot of time.
- *One of the main reasons:* we process a large amount of data.
- So, a natural way to *speed up* data processing is:
  - to divide the data into smaller parts,
  - to process each smaller part separately, and then
  - to combine the results of data processing.
- In particular, when we are processing huge amounts of heterogenous data, it makes sense:
  - to first process different types of data type-by-type,
  - and then to fuse the resulting models.
- This idea is explored in the first sections of Chapter 4.

21. Data Fusion under Interval Uncertainty: Reminder

- *Frequent practical situation:*
  - we are interested in a quantity u;
  - we have several measurements and/or expert estimates $u_1, \ldots, u_n$ of u.
- *Objective:* fuse these estimates into a single more accurate estimate.
- *Interval case:* each u_i is known with interval uncertainty.
- *Formal description:* for each i, we know the interval $\mathbf{u}_i = [u_i - \Delta_i, u_i + \Delta_i]$ containing u.
- *Solution:* u belongs to the intersection $\mathbf{u} \stackrel{\text{def}}{=} \bigcap\limits_{i=1}^{n} \mathbf{u}_i$ of these intervals.

22. Data Fusion under Probabilistic Uncertainty: Reminder

- *Probabilistic uncertainty:* each measurement error $\Delta u_i \stackrel{\text{def}}{=} u_i - u$ is normally distributed w/0 mean and known σ_i.
- *Technique:* the Least Squares Method (LSM)

$$\sum_{i=1}^{n} \frac{(u-u_i)^2}{2\sigma_i^2} \to \min .$$

- *Resulting estimate:* is

$$u = \frac{\sum_{i=1}^{n} u_i \cdot \sigma_i^{-2}}{\sum_{i=1}^{n} \sigma_i^{-2}}.$$

- *Standard deviation:*

$$\sigma^2 = \frac{1}{\sum_{i=1}^{n} \sigma_i^{-2}}, \qquad \text{with } \sigma^2 \ll \sigma_i^2.$$

23. New Problem: Different Resolution

- *Traditional data fusion:* fusing measurement results with different accuracy.
- *Additional problem:* different measurements also have different resolution.
- *Case study – geosciences:* estimating density $u_1, \ldots, u_n$ at different locations and depths.
- *Examples* of different geophysical estimates:
  - *Seismic data* leads to higher-resolution estimates $\widetilde{u}_1, \ldots, \widetilde{u}_n$ of the density values.
  - *Gravity data* leads to lower-resolution estimates, i.e., estimates $\widetilde{u}$ for the weighted average

$$u = \sum_{i=1}^{n} w_i \cdot u_i.$$

24. Why This Is Important

- *Reminder:* there are many sources of data for Earth models:
  - first-arrival passive seismic data (from the actual earthquakes),
  - first-arrival active seismic data (from the seismic experiments),
  - gravity data,
  - surface waves, etc.
- *At present:* each of these datasets is processed separately, resulting in several different Earth models.
- *Fact:* these models often provide complimentary geophysical information.
- *Idea:* all these models describe the properties of the same Earth, so it is desirable to combine them.

25. New Idea: Model Fusion

- *Objective:* to combine the information contained in multiple complementary datasets.
- *Ideal approach:* it is desirable to come up with techniques for joint inversion of these datasets.
- *Problem:* designing such joint inversion techniques is an important theoretical and practical challenge.
- *Status:* such joint inversion methods are being developed.
- *Practical question:* what to do while these methods are being developed?
- *Our practical solution:* fuse the Earth models coming from different datasets.

26. Model Fusion: Statistical Case

- *Objective:* find the values $u_1, \ldots, u_n$ of the desired quantity in different spatial cells.
- *Geophysical example:* u_i is the density at different

$$1 \text{ km } \times 1 \text{ km } \times 1 \text{ km cells.}$$

- *Input:* we have
  - high-resolution measurements, i.e., values $\widetilde{u}_i \approx u_i$ with st. dev. σ_i;
  - lower-resolution measurements, i.e., values $\widetilde{u}^{(k)}$ corresponding to blocks of neighboring cells:

$$\widetilde{u}^{(k)} \approx \sum_i w_i^{(k)} \cdot u_i, \text{ with st. dev. } \sigma^{(k)}.$$

- *Additional information:* a lower-resolution measurement result is representative of values within the block.

27. Model Fusion: Statistical Case (cont-d)

- *Formal description:* when $w_i^{(k)} \neq 0$, we have $\widetilde{u}^{(k)} \approx u_i$, with st. dev. $\delta^{(k)}$.
- *How to estimate $\delta^{(k)}$:* as the empirical st. dev. within the block.
- *High-resolution values (reminder):* $\widetilde{u}_i \approx u_i$ w/st. dev. σ_i.
- *Lower-resolution values (reminder):*

$$\widetilde{u}^{(k)} \approx \sum_i w_i^{(k)} \cdot u_i, \text{ with st. dev. } \sigma^{(k)}.$$

- *LSM Solution:* minimize the sum

$$\sum_i \frac{(u_i - \widetilde{u}_i)^2}{\sigma_i^2} + \sum_i \sum_k \frac{(u_i - \widetilde{u}^{(k)})^2}{(\delta^{(k)})^2} + \sum_k \frac{\left(\widetilde{u}^{(k)} - \sum_i w_i^{(k)} \cdot u_i\right)^2}{(\sigma^{(k)})^2}.$$

- *How:* differentiating w.r.t. u_i, we get a system of linear equations with unknowns $u_1, \ldots, u_n$.

28. Model Fusion: Interval Case

- *Quantities of interest:* values $u_1, \ldots, u_n$ of the desired quantity in different spatial cells.
- *Objective:* find the ranges $\mathbf{u}_1, \ldots, \mathbf{u}_n$ of possible values of $u_1, \ldots, u_n$.
- *High-resolution measurements:* values $\widetilde{u}_i \approx u_i$ with bound Δ_i:
$$\widetilde{u}_i - \Delta_i \le u_i \le \widetilde{u}_i + \Delta_i.$$
- *Lower-resolution measurements:* values $\widetilde{u}^{(k)}$ corresponding to blocks of neighboring cells:
$$\widetilde{u}^{(k)} \approx \sum_i w_i^{(k)} \cdot u_i, \text{ with bound } \Delta^{(k)}.$$
- *Resulting constraint:*
$$\widetilde{u}^{(k)} - \Delta^{(k)} \le \sum_i w_i^{(k)} \cdot u_i \le \widetilde{u}^{(k)} + \Delta^{(k)}.$$

29. Model Fusion: Interval Case (cont-d)

- *Additional information:* a priori bounds on u_i:

$$\underline{u}_i \le u_i \le \overline{u}_i.$$

- *Additional information:* a priori bounds on the changes between neighboring cells:

$$-\delta_{ij} \le u_i - u_j \le \delta_{ij}.$$

- *High-resolution measurements (reminder):*

$$\widetilde{u}_i - \Delta_i \le u_i \le \widetilde{u}_i + \Delta_i.$$

- *Lower-resolution measurements (reminder):*

$$\widetilde{u}^{(k)} - \Delta^{(k)} \le \sum_i w_i^{(k)} \cdot u_i \le \widetilde{u}^{(k)} + \Delta^{(k)}.$$

- *Objective:* minimize and maximize each u_i under these constraints.

30. Model Fusion: Interval Solution

- *Problem.* Minimize (Maximize) u_i under the following constraints:
  - $\underline{u}_i \le u_i \le \overline{u}_i$.
  - $-\delta_{ij} \le u_i - u_j \le \delta_{ij}$.
  - $\widetilde{u}_i - \Delta_i \le u_i \le \widetilde{u}_i + \Delta_i$.
  - $\widetilde{u}^{(k)} - \Delta^{(k)} \le \sum_i w_i^{(k)} \cdot u_i \le \widetilde{u}^{(k)} + \Delta^{(k)}$.
- *Current solution method:* linear programming.
- *Objective:* provide more efficient algorithms for specific geophysical cases.
- *Preliminary results:* some such algorithms have been developed.

Need for Data . . .
Need to Take . . .
Expert Estimates and . . .
How to Propagate . . .
How to Propagate . . .
Chapter 2: Towards . . .
Chapter 3: Towards . . .
Chapter 4: Towards . . .
Chapter 5: Towards . . .
Home Page
Title Page
Page 31 of 91
Go Back
Full Screen
Close
Quit

31. Numerical Experiments

- *What we have done:* proof-of-concept experiments.
- *Simplifications:*
  - equal weights w_i;
  - simplified datasets.
- *Conclusion:* the fused model improves accuracy and resolution of different Earth models.

32. Model Fusion: Input Data, $\widetilde{X}_1 \neq \frac{1}{4} \cdot \sum_{i=1}^{4} \widetilde{x}_i$

$\widetilde{x}_1 = 2.0$	$\widetilde{x}_2 = 3.0$
$\widetilde{x}_3 = 5.0$	$\widetilde{x}_4 = 6.0$

$$\widetilde{X}_1 = 3.7$$

33. Result of Model Fusion: $\widetilde{X}_1 = \frac{1}{4} \cdot \sum_{i=1}^{4} x_i$

$x_1 = 1.7$	$x_2 = 2.7$
$x_3 = 4.7$	$x_4 = 5.7$

34. Example Where Fuzzy Information Helps

- In machine learning:
  - we know how to classify several known objects, and
  - we want to learn how to classify new objects.
- For example, in a biomedical application:
  - we have microarray data corresponding to healthy cells and
  - we have microarray data corresponding to different types of tumors.
- Based on these samples, we would like to be able, given a microarray data, to decide
  - whether we are dealing with a healthy tissue or with a tumor, and
  - if it is a tumor, what type of cancer does the patient have.

35. Chapter 5: Towards Better Ways of Extracting Information About Uncertainty from Data

- Previous methods assume that we have a good description of the uncertainty.
- In practice, often, we do not have this information.
- We need to extract uncertainty information from the data.
- In Chapter 5, we describe how this uncertainty information can be extracted from the data.

36. Extracting Uncertainty from Data: What Is Known

- *Traditional approach:* use a "standard" (more accurate) measuring instrument SMI.
- *Idea:* values $\widetilde{x}_S$ measured by SMI are accurate: $\widetilde{x}_S \approx x$, so $\widetilde{x} - \widetilde{x}_S \approx \Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$.
- *Limitation:* for cutting-egde measurements, we do not have more accurate instruments, these are the best.
- *Example:* the Eddy convariance tower provides the best estimates for Carbon flux.
- *Idea:* if we have two similar measuring instruments, we can estimate $\Delta x^{(1)} - \Delta x^{(2)}$ as $\widetilde{x}^{(1)} - \widetilde{x}^{(2)}$.
- If both error are normally distributed with st. dev. σ, then $\Delta x^{(1)} - \Delta x^{(2)}$ is also normal, with variance $2\sigma^2$.
- So, we can determine σ from observations.

37. Normally Distributed Measurement Errors with Mean 0 and Unknown Variance V

- If we have two similar MIs, then $\widetilde{x}^{(1)} - \widetilde{x}^{(2)} = \Delta x^{(1)} - \Delta x^{(2)}$ is normally distributed w/variance $V' = 2V$.
- From the sample of differences, we estimate V' and estimate V as $V'/2$.
- *Example:* two nearby Eddy Covariance towers.
- In geosciences, we usually have only one seismic map, only one gravity map, etc.
- In general, we have several measurement results $\widetilde{x}^{(i)}$ with variances V_i.
- Here, the variance e_{ij} of the difference $\widetilde{x}^{(i)} - \widetilde{x}^{(j)}$ is equal to $e_{ij} = V_i + V_j$.
- We have 3 equations $e_{12} = \ldots$, $e_{23} = \ldots$, $e_{13} = \ldots$ for 3 unknown variances, so, e.g., $V_1 = \dfrac{e_{12} + e_{13} - e_{23}}{2}$.

38. Need to Go Beyond Normal Distributions, and Resulting Problem

- The distribution of measurement errors is sometimes *not normal* (e.g., in measuring fluxes).
- *In such cases*, in addition to variance V, we need to know skewness and other characteristics.
- *In general*, reconstruction of an asymmetric distribution is not unique.
- *Proof:* Δx and $\Delta y = -\Delta x$ lead to the same distribution for differences $\widetilde{x}^{(1)} - \widetilde{x}^{(2)} = \Delta x^{(1)} - \Delta x^{(2)}$.
- *Natural questions:*
  - which characteristics of the distribution Δx *can* we reconstruct?
  - what are efficient *algorithms* for this reconstruction?

39. Extracting Uncertainty from Data: Main Results

- *Problem:* if distribution of $\Delta x^{(i)}$ is skewed, we cannot distinguish between two distributions:
  - the distribution of $\Delta x^{(i)}$, and
  - its mirror image, the distribution of $-\Delta x^{(i)}$.
- *Our results:*
  - this is the only non-uniqueness, and
  - modulo this non-uniqueness, we can effectively reconstruct the distribution of $\Delta x^{(i)}$.

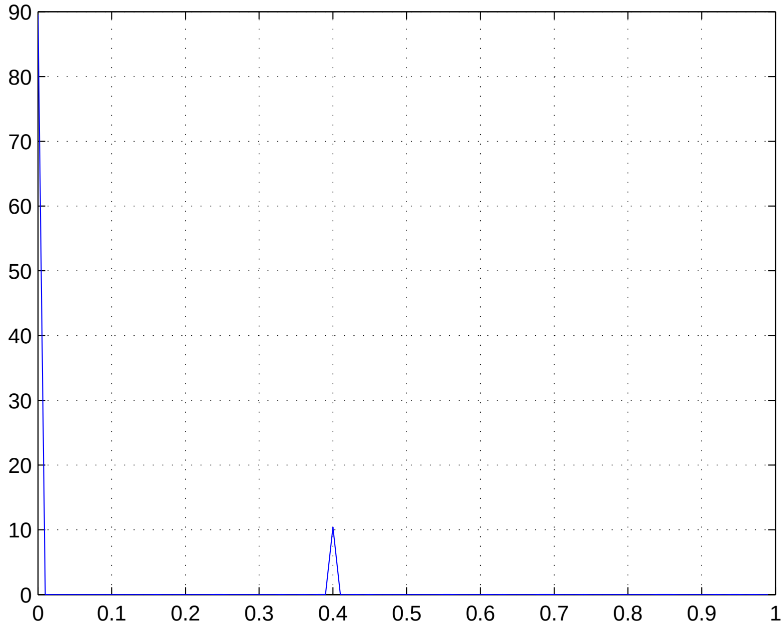
40. How to Gauge the Accuracy

- *Problem:* find a function $\rho(z)$ which satisfies the following two conditions:
  - $\rho(z) \geq 0$ for all z, and
  - $|F(\omega)|^2 = D(\omega)$ for given $D(\omega)$ (Fourier transform of the distribution of $\widetilde{x}^{(1)} - \widetilde{x}^{(2)}$).
- *Method:* of successive projections.
- We start with an arbitrary function $\rho^{(0)}(z)$.
- On the k-th iteration, starting with the result $\rho^{(k-1)}(z)$ of the previous iteration, we:
  - find the closest function $\rho'(x)$ to $\rho^{(k-1)}(z)$ which satisfies the 1st condition;
  - then, find the closest function $\rho^{(k)}(x)$ to $\rho'(z)$ which satisfies the 2nd condition.
- We continue this process until it converges.

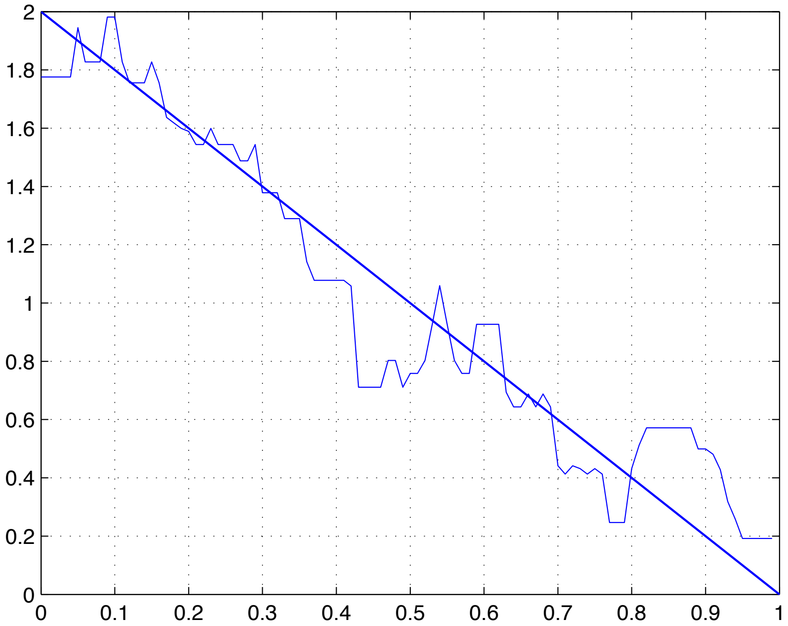
41. Resulting Algorithm

- We start with an arbitrary function $\rho^{(0)}(z)$.
- On the k-th iteration, we start with the function $\rho^{(k-1)}(z)$ obtained on the previous iteration, and we:
  - first, we compute $\rho'(z) = \max(0, \rho^{(k-1)}(z))$;
  - then, we apply Fourier transform to $\rho'(z)$ and get $F'(\omega)$;
  - after that, we compute $F^{(k)}(\omega) = \dfrac{\sqrt{|D(\omega)|}}{|F'(\omega)|} \cdot F'(\omega)$;
  - as the next approx. $\rho^{(k)}(z)$, we take the result of applying the inverse Fourier transform to $F^{(k)}(\omega)$.
- We continue this process until it converges; this enables us to recover many $\rho(x)$.

42. Example: Reconstructing 2-Peak Distribution



43. Example: Reconstructing Asymmetric Triangular Distribution



44. Summary: Main Problem

- In the *ideal world:*
  - we should have an *accurate* description of data uncertainty;
  - based on this description, we should use *well-justified* and *efficient* algorithms to propagate uncertainty.
- In *practice*, we are often not yet in this ideal situation:
  - the description of uncertainty is often only *approximate*,
  - the algorithms for uncertainty propagation are often *heuristics*, i.e., not well-justified, and
  - the algorithms for uncertainty propagation are often *not* very computationally *efficient*.

45. Summary: Conclusions

- In Ch. 2, we showed that the traditional idea of *random* and *systematic* components is an *approximation*:
  - we also need *periodic* components;
  - this is important in environmental studies.
- In Chapter 3, on the example of a fuzzy *heuristic*, we showed how a heuristic can be *formally justified*.
- In Ch. 4, we showed how to be more *efficient*; e.g.:
  - first, we process data type-by-type;
  - then, we fuse the resulting models.
- All these results assume that we have a good description of the uncertainty of the original data.
- In practice, we often need to extract this information from the data; this is described in Ch. 5.

46. Acknowledgments

I would like to express my deep-felt gratitude:

- to my mentor Dr. Vladik Kreinovich;
- to members of my committee, Dr. Aaron Velasco, Dr. Scott Starks, and Dr. Luc Longpré;
- to Dr. Benjamin C. Flores, and to the Alliance for Minority Participation Bridge to the Doctorate program;
- to Dr. Craig Tweedie for his suggestions, comments, and guidance in this work;
- to Dr. Aaron Velasco, Dr. Vanessa Lougheed, Dr. William Robertson, and to all the GK-12 program staff; and
- last but not the least, to all the faculty and staff of the Computational Science Program.

47. Appendix to Chapter 2: Proof of the Theorem

- Shift-invariance means that, for some $c_i(t_0)$, we have

$$x_i(t+t_0) = c_{i1}(t_0)\cdot x_1(t) + \ldots + c_{in}(t_0)\cdot x_n(t).$$

- For n different values $t = t_1, \ldots, t = t_n$, we get a system of n linear equations with n unknowns $c_{ij}(t_0)$.
- The Cramer's rule solution to linear equations is a smooth function of all the coeff. & right-hand sides.
- Since all the right-hand sides $x_i(t_j+t_0)$ and coefficients $x_i(t_j)$ are smooth, $c_{ij}(t_0)$ are also smooth.
- Differentiating w.r.t. t_0 and taking $t_0 = 0$, for $c_{ij} \stackrel{\text{def}}{=} \dot{c}_{ij}(0)$, we get

$$\dot{x}_i(t) = c_{i1}\cdot x_1(t) + \ldots + c_{in}\cdot x_n(t).$$

48. Proof (cont-d)

- *Reminder:* $\dot{x}_i(t) = c_{i1} \cdot x_1(t) + \ldots + c_{in} \cdot x_n(t)$.
- A general solution of such system of equations is a linear combination of functions

$$t^k \cdot \exp(\lambda \cdot t), \text{ w}/k \in \mathrm{N}, k \geq 0, \lambda = a + \mathrm{i} \cdot \omega \in \mathrm{C}.$$

- Here,

$$\exp(\lambda \cdot t) = \exp(a \cdot t) \cdot \cos(\omega \cdot t) + \mathrm{i} \cdot \exp(a \cdot t) \cdot \sin(\omega \cdot t).$$

- When $a \neq 0$, we get unbounded functions for $t \to \infty$ or $t \to -\infty$.
- So, $a = 0$.
- For $k > 0$, we get unbounded t^k; so, $k = 0$.
- Thus, we indeed have a linear combination of sinusoids.

49. Appendix to Chapter 2: How to Propagate Uncertainty in the Three-Component Model

- We are interested in the quantity

$$y = f(x_1(t_{11}), x_1(t_{12}), \ldots, x_2(t_{21}), x_2(t_{22}), \ldots, x_n(t_{n1}), x_n(t_{n2}), \ldots).$$

- Instead of the actual values $x_i(t_{ij})$, we only know the measurement results $\widetilde{x}_i(t_{ij}) = x_i(t_{ij}) + \Delta x_i(t_{ij})$.
- Measurement errors are usually small, so terms quadratic (and higher) in $\Delta x_i(t_{ij})$ can be safely ignored.
- For example, if the measurement error is 10%, its square is 1% which is much much smaller than 10%.
- If the measurement error is 1%, its square is 0.01% which is much much smaller than 1%.
- Thus, we can safely linearize the dependence of Δy on $\Delta x_i(t_{ij})$.

50. How to Propagate Uncertainty (cont-d)

- *Reminder:* we can safely linearize the dependence of Δy on $\Delta x_i(t_{ij})$, so

$$\Delta y = \sum_i \sum_j C_{ij} \cdot \Delta x_i(t_{ij}), \text{ with } C_{ij} \stackrel{\text{def}}{=} \frac{\partial y}{\partial x_i(t_{ij})}.$$

- In general, $\Delta x_i(t_{ij}) = s_i + r_{ij} + \sum_\ell A_{\ell i} \cdot \cos(\omega_\ell \cdot t_{ij} + \varphi_{\ell i})$.
- Due to linearity, we have $\Delta y = \Delta y_s + \Delta y_r + \sum_\ell \Delta y_{p\ell}$, where

$$\Delta y_s = \sum_i \sum_j C_{ij} \cdot s_i; \quad \Delta y_r = \sum_i \sum_j C_{ij} \cdot r_{ij};$$

$$\Delta y_{p\ell} = \sum_i \sum_j C_{ij} \cdot A_{\ell i} \cdot \cos(\omega_\ell \cdot t_{ij} + \varphi_{\ell i}).$$

- *We know:* how to compute Δy_s and Δy_r.
- *What is needed:* propagation of the periodic component.

51. Propagating Periodic Component: Analysis

- *Reminder:* for each component, we have
$$\Delta y_{p\ell} = \sum_i \sum_j C_{ij} \cdot A_{\ell i} \cdot \cos(\omega_\ell \cdot t_{ij} + \varphi_{\ell i}).$$
- It is reasonable to assume that different phrases $\varphi_{\ell i}$ are independent (and uniformly distributed).
- Thus, by the Central Limit Theorem, the distribution of $\Delta y_{p\ell}$ is close to normal, with 0 mean.
- The variance of $\Delta y_{p\ell}$ is $\dfrac{1}{2} \cdot \sum\limits_i A^2_{\ell i} \cdot (K^2_{ci} + K^2_{si})$.
- Each amplitude $A_{\ell i}$ can take any value from 0 to the known bound $P_{\ell i}$.
- Thus, the variance is bounded by $\dfrac{1}{2} \cdot \sum\limits_i P^2_{\ell i} \cdot (K^2_{ci} + K^2_{si})$.
- So, we arrive at the following algorithm.

52. Propagating Periodic-Induced Component: Algorithm

- First, we apply the algorithm f to the measurement results $\widetilde{x}_i(t_{ij})$ and get the estimate $\widetilde{y}$.
- Then, we select a small value δ and for each sensor i, we do the following:
  - take $x_i^{(ci)}(t_{ij}) = \widetilde{x}_i(t_{ij}) + \delta \cdot \cos(\omega_\ell \cdot t_{ij})$ for all moments j;
  - for other sensors $i' \neq i$, take $x_{i'}^{(ci)}(t_{i'j}) = \widetilde{x}_i(t_{i'j})$;
  - substitute the resulting values $x_{i'}^{(ci)}(t_{i'j})$ into the data processing algorithm f and get the result $y^{(ci)}$;
  - then, take $x_i^{(si)}(t_{ij}) = \widetilde{x}_i(t_{ij}) + \delta \cdot \sin(\omega_\ell \cdot t_{ij})$ for all moments j;
  - for all other $i' \neq i$, take $x_{i'}^{(si)}(t_{i'j}) = \widetilde{x}_i(t_{i'j})$;
  - substitute the resulting values $x_{i'}^{(si)}(t_{i'j})$ into the data processing algorithm f and get the result $y^{(si)}$.

53. Algorithm (cont-d)

- *Reminder:*
  - First, we apply the algorithm f to the measurement results $\widetilde{x}_i(t_{ij})$ and get the estimate $\widetilde{y}$.
  - Then, for each sensor i, we simulate cosine terms and get the results $y^{(ci)}$.
  - Third, for each sensor i, we simulate sine terms and get the results $y^{(si)}$.
- Finally, we estimate the desired bound $\sigma_{p\ell}$ on the standard deviation of $\Delta y_{p\ell}$ as

$$\sigma_{p\ell} = \sqrt{\frac{1}{2} \cdot \sum_i P_{\ell i}^2 \cdot \left(\left(\frac{y^{(ci)} - \widetilde{y}}{\delta} \right)^2 + \left(\frac{y^{(si)} - \widetilde{y}}{\delta} \right)^2 \right)}.$$

54. Appendix to Chapter 3

- We know:
  - a fuzzy set $\mu_i(x)$ describing the users' ideal value;
  - the fuzzy set $\mu_a(x)$ describing the users' impression of the actual value.
- For crisp sets, the solution is possibly satisfactory if some of the possibly actual values is also desired.
- In the fuzzy case, we can only talk about the degree to which the proposed solution can be desired.
- A possible decision is satisfactory if either:
  - the actual value is x_1, and this value is desired,
  - or the actual value is x_2, and this value is desired,
  - ...
- Here x_1, x_2, ..., go over all possible values of the desired quantity.

55. Derivation of the d-Formula (cont-d)

- For each value x_k, we know:
  - the degree $\mu_a(x_k)$ with which this value is actual, and
  - the degree $\mu_i(x_k)$ to which this value is desired.
- Let us use $\min(a, b)$ to describe "and" – the simplest possible choice of an "and"-operation.
- Then we can estimate the degree to which the value x_k is both actual *and* desired as

$$\min(\mu_a(x_k), \mu_i(x_k)).$$

- Let us use $\max(a, b)$ to describe "or" – the simplest possible choice of an "or"-operation.
- Then, we can estimate the degree d to which the two fuzzy sets match as

$$d = \max_x \min(\mu_a(x), \mu_i(x)).$$

56. Fuzzy Target Approach: How Are Membership Functions Elicited?

- In many applications (e.g., in fuzzy control), the shape of a membership function does not affect the result.
- Thus, it is reasonable to use the simplest possible membership functions – symmetric triangular ones.
- To describe a symmetric triangular function, it is sufficient to know the support $[\underline{x}, \overline{x}]$ of this function.
- So, e.g., to get the membership function $\mu_i(x)$ describing the desired situation:
  - we ask the user for all the values $a_1, \ldots, a_n$ which, in their opinion, satisfy the requirement;
  - we then take the smallest of these values as $\underline{a}$ and the largest of these values as $\overline{a}$;
  - finally, we form symmetric triangular $\mu_i(x)$ whose support is $[\underline{a}, \overline{a}]$.

57. Analyzing the Problem

- *Reminder:* all we elicit from the experts is two intervals:
  - an interval $[\underline{a}, \overline{a}] = [\widetilde{a} - \Delta_a, \widetilde{a} + \Delta_a]$ describing the set of all *desired* values, and
  - an interval $[\underline{b}, \overline{b}] = [\widetilde{b} - \Delta_b, \widetilde{b} + \Delta_b]$ describing the set of all the values which are *possible*.
- Based on these intervals, we build triangular membership functions $\mu_i(x)$ and $\mu_a(x)$ centered in $\widetilde{a}$ and $\widetilde{b}$.
- For these membership functions,

$$d = \max_x \min(\mu_a(x), \mu_i(x)) = 1 - \frac{|\widetilde{b} - \widetilde{a}|}{\Delta_a + \Delta_b}.$$

- This is the formula that we need to justify.

58. Our Main Idea

- If we knew the exact values of a and b, then we would conclude $a = b$, $a < b$, or $b < a$.
- In reality, we know the values a and b with uncertainty.
- Even if the actual values a and b are the same, we may get approximate values which are different.
- It is reasonable to assume that if the actual values are the same, then $\text{Prob}(a > b) = \text{Prob}(b > a)$, i.e.,

$$\text{Prob}(a > b) = 1/2.$$

- If the probabilities that $a > b$ and that $a < b$ differ, this is an indication that the actual value differ.
- Thus, it's reasonable to use $|\text{Prob}(a > b) - \text{Prob}(b > a)|$ as the degree to which a and b may be different.

59. How To Estimate $\text{Prob}(a > b)$ and $\text{Prob}(b > a)$

- If we knew the exact values of a and b, then we could check $a > b$ by comparing $r \stackrel{\text{def}}{=} a - b$ with 0.
- In real life, we only know a and b with interval uncertainty, i.e., we only know that
$$a \in [\widetilde{a} - \Delta_a, \widetilde{a} + \Delta_a] \text{ and } b \in [\widetilde{b} - \Delta_b, \widetilde{b} + \Delta_b].$$
- In this case, we only know the range $\mathbf{r}$ of possible values of $r = a - b$; interval arithmetic leads to
$$\mathbf{r} = [(\widetilde{a} - \widetilde{b}) - (\Delta_a + \Delta_b), (\widetilde{a} - \widetilde{b}) + (\Delta_a + \Delta_b)].$$
- We do not have any reason to assume that some values from $\mathbf{r}$ are more probable and some are less probable.
- It is thus reasonable to assume that all the values from $\mathbf{r}$ are equally probable, i.e., r is *uniformly* distributed.
- This argument is widely used in data processing; it is called *Laplace Principle of Indifference*.

60. How To Estimate Probabilities (cont-d)

- We estimate Prob$(a > b)$ as Prob$(a - b > 0)$.
- We estimate Prob$(a < b)$ as Prob$(a - b < 0)$.
- We assumed that $r = a - b$ is uniformly distributed on
$$[(\widetilde{a} - \widetilde{b}) - (\Delta_a + \Delta_b), (\widetilde{a} - \widetilde{b}) + (\Delta_a + \Delta_b)].$$
- We can compute Prob$(a - b > 0)$, Prob$(a - b < 0)$, and
$$|\mathrm{Prob}(a > b) - \mathrm{Prob}(b > a)| = \frac{|\widetilde{a} - \widetilde{b}|}{\Delta_a + \Delta_b}.$$
- Since $d = 1 - \dfrac{|\widetilde{b} - \widetilde{a}|}{\Delta_a + \Delta_b}$, we get
$$d = 1 - |\mathrm{Prob}(a > b) - \mathrm{Prob}(b > a)|.$$
- We have produced a new justification for the d-formula.
- This justification that does not use any simplifying assumptions about memb. f-s and/or "and"-operations.

61. Appendix to Chapter 4: Machine Learning

- In machine learning:
  - we know how to classify several known objects, and
  - we want to learn how to classify new objects.
- For example, in a biomedical application:
  - we have microarray data corresponding to healthy cells and
  - we have microarray data corresponding to different types of tumors.
- Based on these samples, we would like to be able, given a microarray data, to decide
  - whether we are dealing with a healthy tissue or with a tumor, and
  - if it is a tumor, what type of cancer does the patient have.

62. Machine Learning: A Formal Description

- Each object is characterized by the results $x = (x_1, \ldots, x_n)$ of measuring several (n) different quantities.
- So, in mathematical terms, machine learning can be described as a following problem:
  - we have K possible labels $1, \ldots, K$ describing different classes;
  - we have several vectors $x(j) \in R^n$, $j = 1, \ldots, N$;
  - each vector is labeled by an integer $k(j)$ ranging from 1 to K;
  - vectors labeled as belonging to the k-th class will be also denoted by $x(k, 1), \ldots, x(k, N_k)$;
  - we want to use these vectors to assign, to each new vector $x \in R^n$, a value $k \in \{1, \ldots, K\}$.

63. Machine Learning: Original Idea

- Often, each class C_k is *convex*: if $x, x' \in C_k$ and $\alpha \in (0,1)$, then $\alpha \cdot x + (1-\alpha) \cdot x' \in C_k$.
- It all C_k are convex, then we can separate them by using linear separators.
- For example, for $K = 2$, there exists a linear function $f(x) = c_0 + \sum\limits_{i=1}^{n} c_i \cdot x_i$ and a threshold value y_0 such that:
  - for all vectors $x \in C_1$, we have $f(x) < y_0$, while
  - for all vectors $x \in C_2$, we have $f(x) > y_0$.
- This can be used to assign a new vector x to an appropriate class: $x \to C_1$ if $f(x) < y_0$, else $x \to C_2$.
- For $K > 2$, we can use linear functions separating different pairs of classes.

64. Machine Learning: Current Development

- In practice, the classes C_k are often not convex.
- As a result, we need *nonlinear* separating functions.
- The first such separating functions came from simulating (non-linear) biological neurons.
- Even more efficient algorithms originate from the Taylor representation of a separating function:

$$f(x_1,\ldots,x_n)=c_0+\sum_{i=1}^n c_i\cdot x_i+\sum_{i=1}^n\sum_{j=1}^n c_{ij}\cdot x_i\cdot x_j+\ldots$$

- This expression becomes linear if we add new variables $x_i\cdot x_j$, etc., to the original variables $x_1,\ldots,x_n$.
- The corresponding *Support Vector Machine* (SVM) techniques are the most efficient in machine learning.
- For example, SVM is used to automatically diagnose cancer based on the microarray gene expression data.

65. There Is Room for Improvement

- In SVM, we divide the original samples into a training set and a training set.
- We train an SVM method on the training set.
- We test the resulting classification on a testing set.
- Depending on the type of tumor, 90 to 100% correct classifications.
- 90% is impressive, but it still means that up to 10% of all the patients are misclassified.
- How can we improve this classification?

66. Our Idea

- Efficient linear algorithms are based on an assumption that all the classes C_k are convex.
- In practice, the classes C_k are often not convex.
- SVM uses (less efficient) general nonlinear techniques.
- Often, while the classes C_k are *not exactly convex*, they are *somewhat* convex:
  - for many vectors x and x' from each class C_k and for many values α,
  - the convex combination $\alpha \cdot x + (1 - \alpha) \cdot x'$ still belongs to C_k.
- In this talk, we use fuzzy techniques to formalize this imprecise idea of "somewhat" convexity.
- We show that the resulting machine learning algorithm indeed improves the efficiency.

67. Need to Use Degrees

- "Somewhat" convexity means that if $x, x' \in C_k$, then $\alpha \cdot x + (1-\alpha) \cdot x' \in C_k$ with some degree of confidence.
- Let $\mu_k(x)$ denote our degree of confidence that $x \in C_k$.
- We arrive at the following fuzzy rule: If $x, x' \in C_k$ *and* convexity holds, then $\alpha \cdot x + (1-\alpha) \cdot x' \in C_k$.
- If we use product for "and", we get

$$\mu_k(\alpha \cdot x + (1-\alpha) \cdot x') \geq r \cdot \mu_k(x) \cdot \mu_k(x').$$

- So, if x'' is a convex combination of two sample vectors, then $\mu_k(x'') \geq r \cdot 1 \cdot 1 = r$.
- For combination of three sample vectors, $\mu_k(x'') \geq r^2$.
- For $y = \sum\limits_{j=1}^{N_k} \alpha_j \cdot x(k,j)$, we have $\mu_k(y) \geq r^{\|\alpha\|_0 - 1}$, where $\|\alpha\|_0$ is the number of non-zero values α_j.

68. Using Closeness

- If $y \in C_k$ and x is close to y, then $x \in C_k$ with some degree of confidence.
- In probability theory, Central Limit Theorem leads to Gaussian degree of confidence.
- We this assume that the degree of confidence is described by a Gaussian expression $\exp\left(-\frac{\|x-y\|_2^2}{\sigma^2}\right)$.
- As a result, for every two vectors x and y, we have

$$\mu_k(x) \geq \mu_k(y) \cdot \exp\left(-\frac{\|x-y\|_2^2}{\sigma^2}\right).$$

69. Combining Both Formulas

- Resulting formula: $\mu_k(x) \ge \widetilde{\mu}_k(x)$, where:

$$\widetilde{\mu}_k(x) \stackrel{\text{def}}{=} \max_{\alpha} \exp\left(-\frac{\left\| x - \sum\limits_{j=1}^{N_k} \alpha_j \cdot x(k,j) \right\|_2^2}{\sigma^2}\right) \cdot r^{\|\alpha\|_0 - 1}.$$

- To classify a vector x, we:
  - compute $\widetilde{\mu}_k(x)$ for different classes k, and
  - select the class k for which $\widetilde{\mu}_k(x)$ is the largest.
- This is equivalent to minimizing $L_k(x) = -\ln(\widetilde{\mu}_k(x))$:

$$L_k(x) = \mathcal{C} \cdot \left\| x - \sum_{j=1}^{N_k} \alpha_j \cdot x(k,j) \right\|_2^2 + \|\alpha\|_0.$$

70. Towards an Efficient Algorithm

- *Reminder:* we minimize $\mathcal{C}\cdot\left\|x-\sum_{j=1}^{N_k}\alpha_j\cdot x(k,j)\right\|_2^2+\|\alpha\|_0$.
- *Lagrange multipliers:* this is equiv. to minimizing $\|\alpha\|_0$ under the constraint $\left\|x-\sum_{j=1}^{N_k}\alpha_j\cdot x(k,j)\right\|_2\le C$.
- *Problem:* minimizing $\|\alpha\|_0$ is, in general, NP-hard.
- *Good news:* often, minimizing $\|\alpha\|_0$ is equivalent to minimizing $\|\alpha\|_1\stackrel{\text{def}}{=}\sum_{j=1}^{N_k}|\alpha_j|$.
- *Resulting algorithm:* minimize

$$\mathcal{C}'\cdot\left\|x-\sum_{j=1}^{N_k}\alpha_j\cdot x(k,j)\right\|_2^2+\|\alpha\|_1.$$

71. Taking the Specific Problem into Account

- For microarray analysis, the actual values of the vector x depend on the efficiency of the microarray technique.
- In other words, with a less efficient technique, we will get $\lambda \cdot x$ for some constant λ.
- From this viewpoint, it is reasonable to use:
  - not just *convex* combinations, but also
  - arbitrary *linear* combinations of the original vectors $x(k, j)$.

72. Towards an Efficient Algorithm (cont-d)

- We repeat ℓ_1-minimization for each of K classes.
- While ℓ_1-minimization is efficient, it still takes a large amount of computation time; so:
  - instead of trying to represent the vector x as a linear combination of vectors from each class,
  - let us look for a representation of x as a linear combination of *all* sample vectors, from all classes:

$$\mathcal{C}' \cdot \left\| x - \sum_{j=1}^{N} \alpha_j \cdot x(j) \right\|_2^2 + \|\alpha\|_1 \to \min .$$

- Then, for each class k, we only take the components belonging to this class, and select k for which

$$\left\| x - \sum_{j:k(j)=k} \alpha_j \cdot x(j) \right\|_2 \to \min .$$

73. Interesting Observation

- This time-saving idea not only increased the efficiency, it also improve the quality of classification.
- We think that this improvement is related to the fact that all the data contain measurement noise.
- On each computation step, we process noisy data.
- Hence, the results get noisier and noisier with each computation step.
- From this viewpoint, the longer computations, the more noise we add.
- By speeding up computation, we thus decrease the noise.
- This compensates a minor loss of optimality, when we replacing K minimizations with a single one.

74. Results

- The probability p of correct identification increased:
  - for brain tumor, p increased from 90% for the best SVM techniques to 91% for our method;
  - for prostate tumor, the probability p similarly increased from 93% to 94%.
- Our method has an additional advantage:
  - to make SVM efficient, we need to select appropriate nonlinear functions;
  - if we select arbitrary functions, we usually get not-so-good results;
  - in contrast, our sparse method has only one parameter to tune: the parameter C'.
- Our technique is this less subjective, more reliable – and leads to better (or similar) classification results.

75. Appendix to Chapter 4: Quantum Computing

- Even after all *algorithmic* speed-ups are implemented, the computation time is still often too long.
- In such situations, the only remaining way to speed up computations is to speed up *hardware*.
- Such ideas range from *available* (e.g., *parallelization*) to *futuristic* (e.g., *quantum computing*).
- *Parallelization* has been largely well-researched.
- The use of *futuristic* techniques in uncertainty estimation is still largely an open problem.
- In the last section of Ch. 4, we show how *quantum computing* can be used to speed up computations.

76. Reliability of Interval Data

- *Usual assumption:* all measuring instruments (MI) functioned correctly.
- *Conclusion:* the resulting intervals $[\widetilde{x}-\Delta,\widetilde{x}+\Delta]$ contain the actual value x.
- *In practice:* a MI can malfunction, producing way-off values (outliers).
- *Problem:* outliers can ruin data processing.
- *Example:* average temperature in El Paso
  - based on measurements, $\dfrac{95+100+105}{3}=100.$
  - with outlier, $\dfrac{95+100+105+\underline{\mathbf{0}}}{4}=75.$
- *Natural idea:* describe the probability p of outliers.
- *Solution:* out of n results, dismiss $k \stackrel{\text{def}}{=} p\cdot n$ largest values and k smallest.

77. Need to Gauge the Reliability of Interval Data

- *Ideal case:* all measurements of the same quantity are correct.
- *Fact:* resulting intervals $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ contain the same (actual) value x.
- *Conclusion:* $\bigcap\limits_{i=1}^{n} \mathbf{x}^{(i)} \neq \emptyset$.
- *Reality:* we have outliers far from x, so $\bigcap\limits_{i=1}^{n} \mathbf{x}^{(i)} = \emptyset$.
- *Expectation:* out of n given intervals, $\geq n-k$ are correct – and hence have a non-empty intersection.
- *Conclusion:*
  - to check whether our estimate p for reliability is correct,
  - we must check whether out of n given intervals, $n-k$ have a non-empty intersection.

78. Need to Gauge Reliability of Interval Data: Multi-D Case

- In practice, a measuring instrument often measure several different quantities $x_1, \ldots, x_d$.
- Due to uncertainty, after the measurement, for each quantity x_i, we have an interval $\mathbf{x}_i$ of possible values.
- So, the set of all possible values of the tuple $x = (x_1, \ldots, x_d)$ is a *box*

  $$X = \mathbf{x}_1 \times \ldots \times \mathbf{x}_d = \{(x_1, \ldots, x_d) : x_1 \in \mathbf{x}_1, \ldots, x_d \in \mathbf{x}_d\}.$$

- Thus:
  - to check whether our estimate p for reliability is correct,
  - we must be able to check whether out of n given boxes, $n - k$ have a non-empty intersection.

79. Resulting Computational Problem: Box Intersection Problem

Thus, both in the interval and in the fuzzy cases, we need to solve the following computational problem:

- Given:
  - integers d, n, and k; and
  - n d-dimensional boxes

$$X^{(j)} = [\underline{x}_1^{(j)}, \overline{x}_1^{(j)}] \times \ldots \times [\underline{x}_n^{(j)}, \overline{x}_n^{(j)}],$$

  $j = 1, \ldots, n$, with rational bounds $\underline{x}_i^{(j)}$ and $\overline{x}_i^{(j)}$.

- *Check* whether
  - we can select $n - k$ of these n boxes
  - in such a way that the selected boxes have a non-empty intersection.

80. Results

- *First result:* in general, the above computational problem is NP-hard.
- *Meaning:* no algorithm is possible that solves all particular cases of this problem in reasonable time.
- *In practice:* the number of d of quantities measured by a sensor is small: e.g.,
  - a GPS sensor measures 3 spatial coordinates;
  - a weather sensor measures (at most) 5:
    - temperature,
    - atmospheric pressure, and
    - the 3 dimensions of the wind vector.
- *Second result:* for a fixed dimension d, we can solve the above problem in polynomial time $O(n^d)$.

81. Algorithm: Description and Need for Speed Up

- *Lemma:* if a set of boxes has a common point, then there is another common vector whose all components are endpoints.
- *Proof:* move to an endpoint in each direction.
- *Number of endpoints:* n intervals have $\le 2n$ endpoints.
- *Bounds on computation time:* we have $\le (2n)^d$ combinations of endpoints, i.e., polynomial time.
- *Remaining problem:* n^d is too slow;
  - for $n = 100$ and $d = 5$, we need 10^{10} computational steps – very long but doable;
  - for $n = 10^4$ and $d = 5$, we need 10^{20} computational steps – which is unrealistic.

82. Use of Quantum Computing

- *Idea:* use Grover's algorithm for quantum search.
- *Problem:* search for a desired element in an unsorted list of size N.
- *Without using quantum effects:* we need – in the worst case – at least N computational steps.
- *A quantum computing algorithm* can find this element much faster – in $O(\sqrt{N})$ time.
- *Our case:* we must search $N = O(n^d)$ endpoint vectors.
- *Quantum speedup:* we need time $\sqrt{N} = O(n^{d/2})$.
- *Example:* for of $n = 10^4$ and $d = 5$,
  - the non-quantum algorithm requires a currently impossible amount of 10^{20} computational steps,
  - while the quantum algorithm requires only a reasonable amount of 10^{10} steps.

83. Quantum Computing: Conclusion

- In traditional interval computations, we assume that
  - the interval data corresponds to guaranteed interval bounds, and
  - that fuzzy estimates provided by experts are correct.
- In practice, measuring instruments are not 100% reliable, and experts are not 100% reliable.
- We may have estimates which are "way off", intervals which do not contain the actual values at all.
- Usually, we know the percentage of such outlier unreliable measurements.
- It is desirable to check that the reliability of the actual data is indeed within the given percentage.

84. Quantum Computing: Conclusions (cont-d)

In this section, we have shown that:

- in general, the problem of checking (gauging) this reliability is computationally intractable (NP-hard);
- in the reasonable case
  - when each sensor measures a small number of different quantities,
  - it is possible to solve this problem in polynomial time;
- quantum computations can drastically reduce the required computation time.

85. Appendix to Chapter 5: Fourier Analysis

- We want to find is the probability density $\rho(z)$ describing the distribution of the measurement error $z \stackrel{\text{def}}{=} \Delta x$.
- In order to find the unknown probability density, we will first find its Fourier transform
$$F(\omega) = \int \rho(z) \cdot e^{\mathrm{i}\cdot\omega\cdot z}\,dz = E\left[e^{\mathrm{i}\cdot\omega\cdot z}\right].$$
- Such a mathematical expectation is also known as a *characteristic function* of the random variable z.
- Based on the observed values of the difference $z^{(1)} - z^{(2)}$, we can estimate its characteristic function
$$D(\omega) = E\left[e^{\mathrm{i}\cdot\omega\cdot(z^{(1)}-z^{(2)})}\right].$$
- It is known that $D(\omega) = F(\omega)\cdot F^*(\omega) = |F(\omega)|^2$.
- How can we reconstruct the complex-valued function $F(\omega)$ if we only know its absolute value?

86. Is It Possible to Estimate Accuracy?

- Theoretically, we can consider all possible values of the difference $z^{(1)} - z^{(2)}$.
- In practice, we can only get values proportional to the smallest measuring unit h (e.g., $h = 1$ cm).
- In the 1-D case, the Fourier transform takes the form $F(\omega) = \sum\limits_{k=0}^{N} p_k \cdot s^k$, where $s \stackrel{\text{def}}{=} e^{\mathrm{i}\cdot\omega\cdot h}$.
- In the multi-D case, we have $z = (k_1 \cdot h_1, k_2 \cdot h_2, \ldots)$, and $F(\omega) = \sum\limits_{k_1=0}^{N_1} \sum\limits_{k_2=0}^{N_2} \ldots p_{k_1,k_2,\ldots} \cdot s_1^{k_1} \cdot s_2^{k_2} \cdot \ldots$
- In terms of polynomials, the question takes the following form:
  - we know the values $D(s) = |P(s)|^2 = P(s) \cdot P^*(s)$ for some polynomial $P(s)$,
  - we need to reconstruct this polynomial $P(s)$.

87. Is It Possible to Estimate Accuracy (cont-d)

- In 1-D case, each complex-valued polynomial of degree N has, in general, N complex roots $s^{(1)}$, $s^{(2)}$, etc.
- Thus, $P(s) = \text{const} \cdot (s - s^{(1)}) \cdot (s - s^{(2)}) \cdot \ldots$ and $|P(s)|^2 = \text{const} \cdot (s - s^{(1)}) \cdot (s - s^{(1)})^* \cdot \ldots$
- There are many factors, so there are many ways to represent it as a product – reconstruction is not unique.
- In the multi-D case, a generic polynomial *cannot* be represented as a product of polynomials.
- E.g., to describe a polynomial $\sum\limits_{k=0}^{n} \sum\limits_{l=1}^{n} c_{kl} \cdot s_1^k \cdot s_2^l$ of degree n, we need $(n+1)^2$ coefficients.
- When polyn. multiply, degrees add: $s^m \cdot s^{m'} = s^{m+m'}$.
- Thus, if $P(s)$ is a product of two polynomials, one has a degree $m < n$, and the other degree $n - m$.

88. Proof (cont-d)

- If $P(s)$ is a product of two polynomials, one has a degree $m < n$, and the other degree $n - m$.
- In general:
  - we need $(m+1)^2$ coefficients to describe a polynomial of degree m and
  - we need $(n-m+1)^2$ coefficients to describe a polynomial of degree $n-m$,
  - so to describe arbitrary products of such polynomials, we need $(m+1)^2 + (n-m+1)^2$ coefficients.
- In general, the total number of coefficients is smaller than $(n+1)^2$.
- So, a general polynomial cannot be represented as a product of two polynomials.

89. Conclusion

- We have shown that a general polynomial cannot be represented as a product of two polynomials.
- Thus, $D(s) = P(s) \cdot P^*(s) = Q(s) \cdot Q^*(s)$ implies that $Q(s) = P(s)$ or $Q(s) = P^*(s)$.
- In the first case, we get $\rho(x)$.
- In the second case, we get $\rho(-x)$.
- So, in general, only $\rho(x)$ and $\rho(-x)$ are consistent with the observed differences $\Delta x^{(1)} - \Delta x^{(2)}$.
- Thus, we *can* reconstruct the distribution $\rho(x)$ of measurement errors – modulo $x \to -x$.

90. Practical Question: How to Gauge the Accuracy

- *Problem:* find a function $\rho(z)$ which satisfies the following two conditions:
  - $\rho(z) \geq 0$ for all z, and
  - $|F(\omega)|^2 = D(\omega)$ for given $D(\omega)$.
- *Method:* of successive projections.
- We start with an arbitrary function $\rho^{(0)}(z)$.
- On the k-th iteration, starting with the result $\rho^{(k-1)}(z)$ of the previous iteration, we:
  - find the closest function $\rho'(x)$ to $\rho^{(k-1)}(z)$ which satisfies the 1st condition;
  - then, find the closest function $\rho^{(k)}(x)$ to $\rho'(z)$ which satisfies the 2nd condition.
- We continue this process until it converges.