Proceedings of the 3rd International Symposium on Uncertainty Quantification and Stochastic Modeling
D.A. Rade; A. M. G. Lima (Editors), ABCM, Maresias, SP, Brazil, February 15-19, 2016

Bayesian Triangle Bound Smoothing

Ivan Sendin[1]**,Thiago Alves de Queiroz**[2]**, and Marcos Aurélio Batista**[3]

[1]Faculty of Computing, Federal University of Uberlândia, Uberlândia, Minas Gerais, Brazil

[2]Institute of Mathematics and Technology, Federal University of Goiás - Campus Catalão,Catalão, Goiás, Brazil.

[3]Department of Computer Science, Federal University of Goiás - Campus Catalão, Catalão, Goiás, Brazil.

Abstract: Triangle bound smoothing is used to reduce the uncertainty from a set of Euclidean distances: an iterative process uses triangle inequality to remove the slack on those distances, providing a new set of distances. In this work we introduce a Bayesian version of triangle bound smoothing that also produce statistical information on those distances. The method was applied to a set of distances that simulate a protein on an NMR experiment and the resulting set of distances obtained a mean squared error value lower than the initial.

Keywords: interval distances

1 INTRODUCTION

In this paper, we address the issue of obtaining a better set of distances given an initial set of distances with some uncertainty. The presence of distances in scientific problems occur quite often, to name a few:

Protein Structure Determination The Nuclear Magnetic Resonance (NMR) experiment is used to determine the protein structure. The information obtained by NMR experiment is the imprecise distance of some pair of atoms, together with some exact distances obtained through molecular geometry, the protein structure is determined [9]. The computational problem related with protein structure determination using NMR distances is called *graph embedding problem*: given $G(V, E)$ one ask if there is a $x_i \in R^k$ such that $|x_i - x_j| = e_{ij}, \forall e_{ij} \in E$. This problem is known to be NP-Complete for $k = 1$ and NP-hard for $k \geq 1$[17]. For a sufficient dense graph with exact distances, the problem can be solved in polynomial time [6]. When interval distances are involved this problem is known as Molecular Distance Geometry Problem (MDGP)[14];

Global Positioning System(GPS) To compute the current position and velocity using GPS a unit receiver measures this distance to several GPS satellites [16];

Nanostructures As in the case of MDGP, the process to obtain a three-dimensional representation of nanostructures begins with obtaining the distances between pairs of atoms using NMR. In this case, the information of which pair each distance is associated is not available. This problem is called Unassigned Distance Geometry Problem and can be addressed using Liga Algorithm [11, 10];

Sensor Network localization This problem consist on a large number of wireless sensors scattered around an area. Each sensor is capable to communicate with nearby peers using radio signal and infer their distance using radio signal timing. Some sensors have this position precisely determined by GPS. The problem consists in determine the location of each sensor [1, 3, 21].

One difficulty that must be handled in the search for solutions is that the information is not precise: errors and limitations on measurements, models and computation produce inexact distances. In some cases only a minimum and maximum values for each distance can be determined, these distances are called interval distances.

The existence of interval distances increases the search space and hinders the build of the solution. Some addresses the problem using particles filter [18, 19], discretization[13], optimization [20] or mathematical modeling[2]. One approach to minimize the interval distance issue is simple produce narrower intervals using triangle bound smoothing.

In this work, we introduce a Bayesian bound smoothing that in addition to narrowing the intervals also produces statistical information about them. This paper is organized as follow: in Section 2 we present the traditional triangle bound smoothing algorithm , in Section 3 the statistical motivation to use a different approach to triangle smoothing is explained. In Section 4, we present result of this new approach presenting some result obtained trough computational experiments. Finally, Section 5 presents our conclusions.

2 TRIANGLE BOUND SMOOTHING

Triangle smoothing or bound smoothing [4] is a method used to reduce the slack on a system composed of interval distances, it is based on the triangle inequality: given a,b and c the size of the sides of an triangle the following holds:

$$a \leq b + c. \tag{1}$$

As stated before, imprecise distances can be modeled in form of intervals, the correct value of one distance a is unknown, but we known upper and lower bounds for a and we will use these values to solve some problem, the interval value is represented by $\bar{a} = [l_a, u_a]$, with $a \geq l_a$ and $a \leq u_a$.

The triangle inequality (Equation 1) holds for interval values and the following inequations can be obtained[1]:

$$u_a \leq u_b + u_c \tag{2}$$

and

$$l_a \geq l_b - u_c, \tag{3}$$

for one triangle with sides $\bar{a}, \bar{b}$ and $\bar{c}$.

In case of one inequation do not hold, the values of u_a or l_a are updated, reducing the slack from interval values.

For a system composed of several interconnected triangles the slack removal process is performed iteratively until a limit is reached [5, 15].

As usually presented, triangle bound smoothing is deterministic and conservative: they hold wider intervals, even when they are unlikely; and computationally expensive [5, 15].

3 BAYESIAN TRIANGLE BOUND SMOOTHING

The proposed method is an application of Bayesian inference[8] in a set of interval distances. Initially the *prior* of each distance is a uniform distribution whose support is defined by the limits of its interval. Using Bayesian inference narrower limits are generated and statistical information about intervals are obtained. The inference is based on the following: given $\bar{a}, \bar{b}$ and c the distances of a triangle, $\bar{a}, \bar{b} \in R^2$, and $c \in R$. For $x, x \in \bar{a}$, we have

$$P(\dot{a} > x) \propto P(\dot{b} > x - c) \tag{4}$$

,where $\dot{a}$ and $\dot{b}$ are samples from $\bar{a}$ and $\bar{b}$ and $\propto$ means *proportional to*.

3.1 Illustrative Example

To illustrate the rationale of the method, lets take for example one triangle with following vertex distances: $\bar{a} = [3, 7], \bar{b} = [2, 4]$ and $c = 2$, those distances were obtained by a imprecise method, e.g NMR, and we known that the interval distances contains the correct distance. Using traditional triangle smoothing one can remove the slack of the measurement and update $\bar{a}$ to $[3, 6]$.

Using Equation 4, one can state, for example, the following:

1. $P(a \geq 3) \propto P(b \geq 1)$
2. $P(a \geq 4) \propto P(b \geq 2)$
3. $P(a \geq 5) \propto P(b \geq 3)$
4. $P(a \geq 6) \propto P(b \geq 4)$
5. $P(a \geq 7) \propto P(b \geq 5)$

While traditional triangle bound smoothing produces a new interval ($\bar{a} = [3, 6]$), the Bayesian approach produces new interval, with same bounds, and statistical information about it. For example, in statements 4 and 5 the probability related wit the a distance is zero, because the probability on the right term is also zero. Analyzing statement 3 one could conclude that $a > 5$ is unlikely because its probability is proportional to $P(b \geq 3)$. In Figure 1 we show the likelihood obtained analytically using statistical inference this example.

[1] Also, any permutation on $\bar{a}, \bar{b}$ and $\bar{c}$

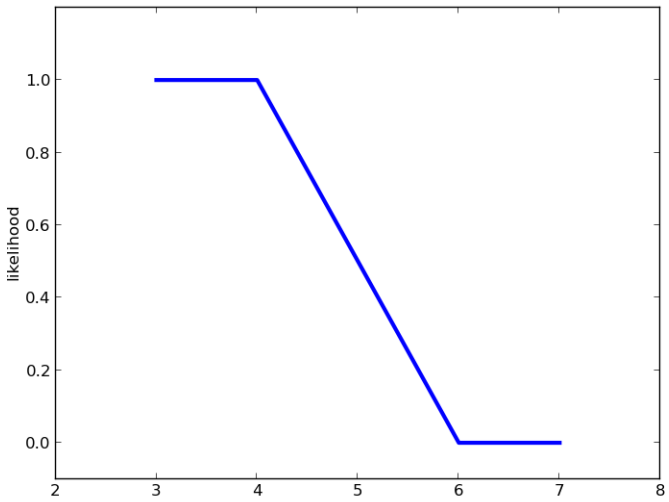

Figure 1 – Likelihood function for distance a - with prior $\bar{a} = [3, 7]$ - forming a triangle with others sides $\bar{b} = [2, 4]$ and $c = 2$. The function is normalized to have unit maximum.

Data: Distances $\bar{a}, \bar{b}, c$; k number of steps
Result: Weighted particles in $\bar{a}$
$a_t = \emptyset$;
for k *steps* **do**
  Sample $\dot{a}$ from $\bar{a}$;
  Sample $\dot{b}$ from $\bar{b}$;
  if $\dot{a} \leq \dot{b} + c$ **then**
    Save $\dot{a}$ in a_t;
  end
  Update $\bar{a}$ with a_t;
end

Algorithm 1: Updating the importance of particles in $\bar{a}$

4 COMPUTATIONAL EXPERIMENTS

The Bayesian inference process was implemented in Python using two different approaches:

Particles Based Each interval distance is associated with a set of particles sampled from a uniform distribution limited by the associated range. To apply Bayesian inference triangles are sampled and the importance of the particle of each interval distance is updated using triangular inequality;

Explicit Probabilities Each interval is associated with a uniform distribution represented by k slices, each of then receiving the value of $\frac{1}{k}$, the initial probability of each slice.The calculation of the probability of a region is made with the sum of the slices of interest - Algorithm 3. Upon execution of the method, the value of each slice is updated according to Equation 4.

The proposed method was applied to two experiments: first is verified if the method can generate an expected distribution for a set of randomly generated triangles, secondly is verified if the information generated by the method is meaningful.

4.1 Random Triangles

The perform this test, random triangles were generated according with [7]. From the set of random triangles we selected those - by triangular similarity - compatible with the triangle with sides 2,[2,4],[3,7]. We use a set of particles sampled uniformly [3,7], we used the Bayesian process to update the importance of those particles according with their compatibility with the distances 2 and a sample from [2,4]. The distance distribution obtained from random triangles of the side [3,7] is presented in red in Figure 2 and in blue the final distribution.

Data: Set of distances S; k number of steps, k' number of steps
Result: Weighted particles in S
for $k1$ *steps* **do**
  Randomnly select a trinagle with wides $(\bar{a}, \bar{b}, c)$ from S;
  Particles($\bar{a}, \bar{b}, c, k'$);
  Particles($\bar{b}, \bar{a}, c, k'$);
end

Algorithm 2: Performs the update of a set of interval distances using particles.

Data: l,min,max,k,w[]
Result: $P(d \leq l)$
if $l < min$ **then**
  return 0;
end
if $l > max$ **then**
  return 1;
end
i $\leftarrow$ index of l in $w[]$;
return $\sum_{j=0}^{i} w[j]$;

Algorithm 3: Calculation of the probability of an interval distance represented by slices.

4.2 Protein NMR Data

Using a method for creating artificial backbones [12], a set of points is created, representing a backbone of a protein, where each point is equivalent to an atom. Precise and interval distances are obtained of some atom pairs in the same way that we would expect in an NMR experiment. The interval distances pass through a Bayesian inference process.

Table 1 shows the results obtained in applying the method on each interval distance of a protein with a backbone of 10 atoms

The MSE(Mean Squared Distance) column is calculated sampling from distribution and comparing with the correct distance:

$$\frac{1}{k}\sum_{i=1}^{k}(r - a_i)^2 \tag{5}$$

, where r is the correct distance and a_i is one sample from posteriori distribution or uniform distribution - for comparison purposes. The values shown were obtained with $k = 1000$ samples. The Correct column shows whether the new range contains the correct value. In brackets, the distance from interval to the correct value is showed if it does not contain. The Shrinkage show the amount of shrinkage of each interval.

The overral MSE was analyzed and Wilcoxon Signed-Ranks Test indicated that the MSE calculated from uniform distribution, $Mdn = 0.34$, were statistically significantly higher than MSE obtained with the proposed method, $Mdn = 0.30$, $Z = 56$, $p < .03$.

Table 2 shows the average results of the method when applied to proteins with sizes 100, 150 and 200.

5 CONCLUSIONS

We proposed a Bayesian method to reduce the slack from a set of interval distances and produce statistical information about them.

We performed computational experiments and was possible to reproduce a similar distribution expected in a set of triangles randomly generated.

When the method was applied to a set of distances simulating an NMR experiment it was observed that the average squared error decreased compared to initial uniform distribution.

Acknowledgements.

The authors would like to thank the Brazilian National Council for Scientific and Technological Development (CNPq), Research Support Foundation of Goiás State (FAPEG) and Federal University of Uberlândia (grant 04/1014/074-PROPP).

REFERENCES

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. Wireless sensor networks: A survey. *Comput. Netw.*, 38(4):393–422, March 2002.

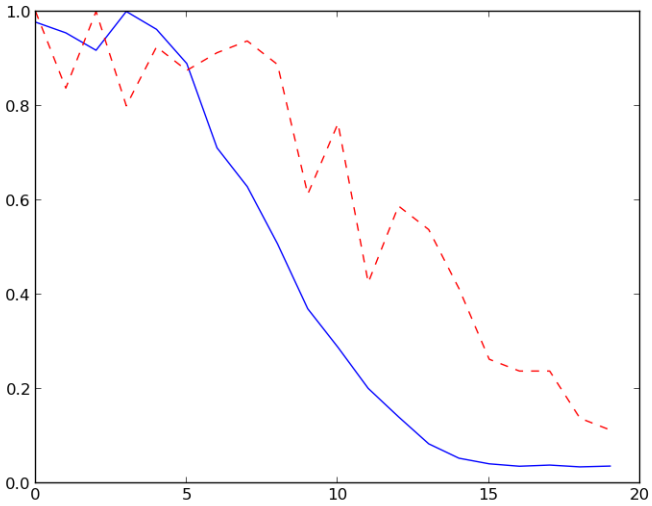

Figure 2 – In red, the size distribution to side [3,7] obtained with randomly generated triangles with other sides [2,4] and 2. In blue, the distance distribution obtained with particles.

[2] Babak Alipanahi, Nathan Krislock, Ali Ghodsi, Henry Wolkowicz, Logan Donaldson, and Ming Li. Determining protein structures from NOESY distance constraints by semidefinite programming. *Journal of Computational Biology*, 20(4):296–310, 2013.

[3] P. Biswas and Yinyu Ye. Semidefinite programming for ad hoc wireless sensor network localization. In *Information Processing in Sensor Networks, 2004. IPSN 2004. Third International Symposium on*, pages 46–54, April 2004.

[4] G. M. Crippen and T. F. Havel. *Distance Geometry and Molecular Conformation*. England: Research Studies Press Ltd, 1988.

[5] Narsingh Deo and Paulius Micikevicius. Coarse-grained parallelization of distance-bound smoothing for the molecular conformation problem. In SajalK. Das and Swapan Bhattacharya, editors, *Distributed Computing*, volume 2571 of *Lecture Notes in Computer Science*, pages 55–66. Springer Berlin Heidelberg, 2002.

[6] Qunfeng Dong and Zhijun Wu. A geometric build-up algorithm for solving the molecular distance geometry problem with sparse distance data. *Journal of Global Optimization*, 26(3):321–333, 2003.

[7] Alan Edelman and Gilbert Strang. Random triangle theory with geometry and applications. *Foundations of Computational Mathematics*, 15(3):681–713, 2015.

[8] N.J. Gordon, D.J. Salmond, and A.F.M. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, 140(2):107–113, Apr 1993.

[9] Peter Guntert. Structure calculation of biological macromolecules from nmr data. *Quarterly Reviews of Biophysics*, 31:145–237, 5 1998.

[10] P. Juh, L. Granlund, P. M. Duxbury, W. F. Punch, and S. J. L. Billinge. The Liga algorithm for ab initio determination of nanostructure. *Acta Crystallographica Section A*, 64:631–640, 2008.

[11] P. Juhs, D. M. Cherba, P. M. Duxbury, W. F. Punch, and S. J. L. Billinge. Ab initio determination of solid-state nanostructure. *Nature*, 440:655 – 658, 2006.

[12] Carlile Lavor. On generating instances for the molecular distance geometry problem. In *Global Optimization: From Theory to Implementation, Nonconvex Optimization and Its Application Series*, pages 405–414. Springer, 2006.

[13] Leo Liberti, Carlile Lavor, and Nelson Maculan. A branch-and-prune algorithm for the molecular distance geometry problem. *International Transactions in Operational Research*, 15(1):1–17, 2008a.

[14] Antonio Mucherino, Carlile Lavor, Leo Liberti, and Nelson Maculan. *Distance geometry: theory, methods, and applications*. Springer Science & Business Media, 2012.

[15] Kumar Rajan and Narsingh Deo. Computational experience with a parallel algorithm for tetrangle inequality bound smoothing. *Bulletin of Mathematical Biology*, 61(5):987–1008, 1999.

#	$Pair$	$Shrinkage$	$Correct$	$MSE/Particles$	$MSE/Uniform$
1	(0,3)	0.11	Y	0.30	0.33
2	(0,4)	0.09	Y	0.26	0.45
3	(0,5)	0.33	Y	0.12	0.41
4	(1,4)	0.08	Y	0.45	0.33
5	(1,5)	0.17	Y	0.32	0.52
6	(1,6)	0.55	Y	0.27	0.50
7	(2,5)	0.08	Y	0.44	0.36
8	(2,6)	0.18	N(0.05)	0.58	0.56
9	(2,7)	0.12	Y	0.19	0.24
10	(2,8)	0.02	Y	0.21	0.22
11	(2,9)	0.08	Y	0.33	0.43
12	(3,6)	0.36	Y	0.25	0.38
13	(3,7)	0.09	N(0.04)	0.31	0.26
14	(3,8)	0.06	Y	0.11	0.17
15	(3,9)	0.02	Y	0.30	0.42
16	(4,7)	0.21	Y	0.38	0.37
17	(4,8)	0.12	Y	0.31	0.30
18	(4,9)	0.03	Y	0.34	0.30
19	(5,8)	0.35	Y	0.22	0.28
20	(5,9)	0.08	Y	0.17	0.31
21	(6,9)	0.06	Y	0.38	0.34

Table 1 – Results obtained with the use of particle on Bayesian inference on a protein backbone structure with 10 atoms.

[16] Simo Sarkka. *Bayesian Filtering and Smoothing*. Cambridge University Press, New York, NY, USA, 2013.

[17] J. Saxe. Embeddability of weighted graphs in k-space is strongly np-hard. *Proceedings of 17th Allerton Conference in Communications, Control and Computing*, pages 480–489, 1979.

[18] Ivan Sendin and Siome Goldenstein. Proteins structure determination with imprecise distances. In *Workshop on Distance Geometry and Applications (DGA)*, 2013.

[19] Ivan Sendin, Siome Goldenstein, and Carlile Lavor. Protein structure reconstruction with data uncertainties. In *Uncertanties*, 2012.

[20] Changzhi Wu, Chaojie Li, and Qiang Long. A dc programming approach for sensor network localization with uncertainties in anchor positions. *Journal of Industrial and Management Optimization*, 10(3):817–826, 2014.

[21] Zhenzhen Zheng, Xinlong Luo, and Zhijun Wu. Geometric buildup algorithms for sensor network localization. *Mathematical Problems in Engineering*, 2012:1–18, 2012.

	$Distances$		Shrinkage		MSE(avg)			p value	
	Total	Interval	Particles	Direct	Uniform	Particles	Direct	Particles	Direct
Size 100									
1	2762	394	0.473	0.027	0.353	0.313	0.351	$< 10^{-41}$	0.172
2	2450	394	0.445	0.026	0.363	0.320	0.359	$< 10^{-42}$	0.193
3	3032	394	0.499	0.027	0.355	0.313	0.351	$< 10^{-44}$	0.135
Size 150									
1	3650	594	0.465	0.027	0.358	0.312	0.354	$< 10^{-63}$	0.055
2	5492	594	0.544	0.028	0.362	0.318	0.358	$< 10^{-86}$	0.009
3	5738	594	0.019	0.027	0.353	0.353	0.348	0.808	0.009
Size 200									
1	8108	794	0.556	0.027	0.355	0.310	0.351	$< 10^{-138}$	0.003
2	6096	794	0.496	0.027	0.357	0.317	0.353	$< 10^{-85}$	0.011
3	5148	794	0.446	0.027	0.354	0.308	0.350	$< 10^{-87}$	0.013

Table 2 – Results for three instances for each of protein sizes: 100, 150 and 200 using particles and explicit probability.