Finding the Population Variance of Costs over the Solution Space of the TSP in Polynomial Time

PAUL JOHN SUTCLIFFE,ANDREW SOLOMON,JENNY EDWARDS
University of Technology, Sydney
Faculty of Information Technology
Broadway, Sydney, NSW 2007
AUSTRALIA

Abstract: We give a polynomial time algorithm to find the population variance of tour costs over the solution space of the symmetric Traveling Salesman Problem (TSP). In practical terms the algorithm provides a linear time method, on the number of edges of the problem, for determining the standard deviation of these costs. Application of the algorithm has produced empirical evidence that there is a clear relationship between the optimal tour cost and the standard deviation. This suggests that there may be a polynomial time algorithm to estimate the likely optimal tour cost of a TSP. The method for finding the variance also shows promise of being generalizable to higher statistical moments.

Key–Words: Traveling salesman problem, Combinatorial optimization, Variance, Standard deviation, Probability distributions, Moments

1 Introduction

1.1 The TSP

The traveling salesman problem (TSP) is a classic problem in combinatorial optimization. The problem is this: given a finite set of cities together with the distances between these cities, find the tour of shortest distance which visits each city just once and returns to the city of origin. Extensive general references include [9][2][18]. Linear programming reductions are surveyed in [15], while properties of frequently used local search heuristics are considered in [3].

It is natural to define the symmetric TSP in terms of a complete *undirected* graph $\Gamma = (V, E)$ with the vertices V representing cities, and the edges E representing the connections between cities. We label the set of n vertices as $\{1, 2, \ldots, n\}$, and an n-cycle permutation as a *tour* or *solution*, π. The set of all tours, the *solution space*, is denoted Θ. The distance between cities (or *cost* of an edge), is a function $\mathrm{c} : E \to \Re$ which we extend to the function,$\pi : \Omega \to \Re$, defined as the cost of a tour $\Omega(\pi) = \sum_{i=1}^{n} \mathrm{c}(\{\pi(i), \pi((i \mod n) + 1)\})$.

The TSP then, is to find some n-cycle permutation π of V for which $\Omega(\pi)$ is smallest, and such a permutation π^* is called a *global minimum tour* or *optimal tour*. The central difficulty in finding a global minimum tour is that the number of tours increases in proportion to the factorial of the number of cities. Specifically, if there are n cities then the number of tours is $|\Theta| = (n - 1)!/2$. It is well known that the existence of a polynomial time algorithm to find the global minimum cost of a TSP would imply the existence of a polynomial time algorithm to find a tour with this cost [18, p10].

1.2 Importance of this Research

The TSP is of interest for three interrelated reasons. Firstly the problem and its obvious variations have numerous practical applications, secondly it is known to be NP-complete [18, p9], and so is of theoretical interest, and lastly it is, comparatively, well studied and so makes a good test application for new optimization algorithms [14]. Novel optimization heuristics are typically motivated by observation of the behavior of complex systems, for example, evolution by natural selection is mimicked in genetic algorithms, the hardening of metals in simulated annealing [11, Chapter 6]. However our normal three dimensional world provides poor guidance to the characteristics of combinatorial problems. A good example is the properties of local minima of the TSP under the 2-opt move [18]. It is shown in [6] that each local minimum has a cost of no more than the average cost of solutions over the solution space. So whereas in the real world, we may have crater lakes on high mountains no such traps can occur in this landscape of the TSP (although many others can). The chief purpose of the research discussed here is to provide more insight into the solution space of the TSP.

The remainder of this paper is organized as follows. In Section 1.3 we briefly review the statistical concepts used. In Section 2 we prove that it is possible to find the variance of tour costs over the solution space of a TSP of size n cities in $O(n^2)$. Section 2.2 considers the generalization of our main result to higher statistical moments. Finally in Section 3.1 we apply our main result to an empirical study of the relationship between, the cost of optimal solutions, the standard deviation of costs and the problem size.

1.3 Statistical Preliminaries

We recall [20, Chapter 4][5, Chapter 5] that the *expected value*, Ex, or *mean* value of a discrete random variable X is $\mathrm{Ex}(X) = \sum_x x\mathrm{p}(x)$, where $\mathrm{p}(x)$ is the probability of X taking the value x. The *rth moment about the mean* is defined as

$$\mathrm{mm_r}(X) = \mathrm{Ex}((X - \mathrm{Ex}(X))^r). \tag{1}$$

The second moment about the mean, $\mathrm{mm}_2(X)$, is the well known *population variance*[1] and reflects the spread of values around the mean of X. The square root of the variance is the standard deviation. In terms of a TSP with solution space Θ, cost function Ω and mean tour cost μ, (1) can be written

$$\mathrm{mm_r}(\Theta) = \frac{\sum\limits_{\pi\in\Theta}((\Omega(\pi) - \mu)^r)}{|\Theta|}. \tag{2}$$

It is reported in [16], and a simple proof follows from Lemma 1 below, that the mean tour cost over the solution space of a problem of size n cities with edge set E is $\mu = \frac{2}{n-1}\sum\limits_{e\in E}\mathrm{c}(e)$.

2 Finding the Variance of Tour Costs in Polynomial Time

We begin with three technical lemmas concerning the number of edges in the solution space. Each is easily proved.

Lemma 1 *There are $(n-2)!$ tours containing a particular edge $\{u, v\}$ for a TSP of size $n \geq 3$ cities.* □

Lemma 2 *There are $(n-3)!$ tours containing the two adjacent edges $\{u, v\}$, $\{v, x\}$ for TSP of size $n \geq 3$ cities.* □

Lemma 3 *There are $2(n-3)!$ tours containing the two non-adjacent edges $\{u, v\}, \{x, y\}$, with u, v, x, y distinct, for a TSP of size $n > 3$ cities.* □

2.1 Proof of Variance Theorem

In order to prove our central theorem we provide some notational machinery. Let Θ be the solution space of a TSP with edge set E and cost function Ω. We index each π in Θ with an integer $k \in [1 \ldots |\Theta|]$, similarly we label the edges of E as e_i with $i \in [1 \ldots |E|]$. We define the function $[1 \ldots |\Theta|]\times[1 \ldots |E|] : t \to \{0, 1\}$ as

$$t_{ki} = \begin{cases} 1 & \text{if edge } e_i \text{ is in tour } k \\ 0 & \text{otherwise.} \end{cases}$$

Under this arrangement if k is the index of a tour π, then the cost of π is

$$\Omega(\pi) = t_{k1}\mathrm{c}(e_1) + t_{k2}\mathrm{c}(e_2) \ldots t_{k|E|}\mathrm{c}(e_{|E|}),$$

and specializing (2) to $r = 2$, the variance of the tour costs over a solution space with mean tour cost μ is

$$\mathrm{mm}_2(\Theta) = \frac{\sum\limits_{k=1}^{|\Theta|}((t_{k1}\mathrm{c}(e_1){+}t_{k2}\mathrm{c}(e_2)...t_{k|E|}\mathrm{c}(e_{|E|}){-}\mu)^2)}{|\Theta|}. \tag{3}$$

Now $|\Theta|$ is of course factorial on n and so this formulation is impractical for all but the smallest problems. In Theorem 4 we give a polynomial time solution to the problem.

Theorem 4 *The variance of tour costs over the solution space of a TSP of size n cities and with edge set, E is*

$$\mathrm{var} = \frac{2\beta_1}{(n-1)} - \frac{4\beta_1 + 2\beta_2}{(n-1)(n-2)} \tag{4}$$

with the functions β_1, β_2 being defined as

$$\begin{aligned} \beta_1 &= \sum_{e\in E}\mathrm{c}_0(e)^2 \\ \beta_2 &= \sum_{e\in E}\left[\mathrm{c}_0(e)\sum_{f\in A(e)}\mathrm{c}_0(f)\right] \end{aligned} \tag{5}$$

where $\mathrm{c}_0(e) = \mathrm{c}(e) - \mu/n$ and $A(e)$ is the set of edges adjacent to an edge e.

Proof: Each tour has only n edges, so for a given tour k, only n of the t_{ki} are 1, the remaining are 0. This gives (3) as

$$\mathrm{var}(\Theta) \;=\; \frac{\sum\limits_{k=1}^{|\Theta|}((t_{k1}\mathrm{c}_0(e_1){+}t_{k2}\mathrm{c}_0(e_2)...t_{k|E|}\mathrm{c}_0(e_{|E|}))^2)}{|\Theta|},$$

[1] For the remainder of this paper we will simply refer to this statistic as the *variance*

$$= \frac{\sum_{k=1}^{|\Theta|} \sum_{i=1}^{|E|} \sum_{j=1}^{|E|} t_{ki} t_{kj} c_0(e_i) c_0(e_j)}{|\Theta|} .$$

Expanding the numerator we have

$$\sum_{k=1}^{|\Theta|} \sum_{i=1}^{|E|} \sum_{j=1}^{|E|} t_{ki} t_{kj} c_0(e_i) c_0(e_j) \qquad (6)$$

$$\begin{aligned} = & \; \alpha_{11} c_0(e_1) c_0(e_1) + \ldots + \alpha_{1|E|} c_0(e_1) c_0(e_{|E|}) \\ + & \; \alpha_{21} c_0(e_2) c_0(e_1) + \ldots + \alpha_{2|E|} c_0(e_2) c_0(e_{|E|}) \\ & \vdots \\ + & \; \alpha_{|E|1} c_0(e_{|E|}) c_0(e_1) + \ldots + \alpha_{|E||E|} c_0(e_{|E|}) c_0(e_{|E|}) \end{aligned}$$

where $\alpha_{ij} = \sum_{k=1}^{|\Theta|} t_{ki} t_{kj}$, is the number of tours which contain both edges e_i and e_j.

Let e_i and e_j be any two edges of a TSP, either these edges are equal, they are adjacent or they are neither equal nor adjacent. Thus each constant α_{ij} is one of three values.

case 1 α_{ii}. By Lemma 1 each edge e_i appears in $(n-2)!$ tours over the solution space, thus $\alpha_{ii} = (n-2)!$.

case 2 α_{ij} such that edge e_i is adjacent to e_j. By Lemma 2 any two adjacent edges appear in $(n-3)!$ tours so in this case $\alpha_{ij} = (n-3)!$.

case 3 α_{ij} such that edge e_i is non adjacent to e_j. By Lemma 3 the two edges appear in $2(n-3)!$ tours so $\alpha_{ij} = 2(n-3)!$.

We recall that $A(e)$ is the set of edges adjacent to an edge e and we define $N(e)$ to be the set of edges neither adjacent to e nor equal to e. So (6) becomes,

$$\begin{aligned} = & \; (n-2)! \left(\sum_{e \in E} c_0(e)^2 \right) \\ + & \; (n-3)! \left(\sum_{e \in E} \left[c_0(e) \sum_{f \in A(e)} c_0(f) \right] \right) \\ + & \; 2(n-3)! \left(\sum_{e \in E} \left[c_0(e) \sum_{f \in N(e)} c_0(f) \right] \right) , \end{aligned}$$

$$= (n-2)!\beta_1 + (n-3)!\beta_2 + 2(n-3)!\beta_3,$$

where

$$\beta_3 = \sum_{e \in E} \left[c_0(e) \sum_{f \in N(e)} c_0(f) \right] .$$

Giving the variance as

$$\begin{aligned} \mathrm{var}(\Theta) & = \tfrac{2((n-2)!\beta_1 + (n-3)!\beta_2 + 2(n-3)!\beta_3)}{(n-1)!} \\ & = \tfrac{2\beta_1}{(n-1)} + \tfrac{2\beta_2 + 4\beta_3}{(n-1)(n-2)} \end{aligned}$$

However it's easy to see that $\beta_3 = -\beta_1 - \beta_2$, since for any e in E we have $E = A(e) \bigcup N(e) \bigcup \{e\}$ and $\sum_{e \in E} c_0(e) = 0$ (by the definition of c_0). Therefore we have

$$\mathrm{var}(\Theta) = \tfrac{2\beta_1}{(n-1)} - \tfrac{4\beta_1 + 2\beta_2}{(n-1)(n-2)} \qquad (7)$$

as required. □

Corollary 5 *It is possible to calculate the variance of tour costs over the solution space of any TSP with n cities in $O(n^2)$.*

Proof: The function β_1 above can clearly be found in $O(n^2)$, since $|E| = (n^2 - n)/2$. Let I_x be the set of edges incident to a city x. Let $e = \{x, y\}$ be an edge, then $A(e) = (I_x - \{e\}) \bigcup (I_y - \{e\})$. Thus to find β_2 first compute and store $S_x = \sum_{f \in I_x} c_0(f)$ for each of the n cities of the instance. The complexity of this phase is $O(n^2)$. The sum β_2 is then

$$\beta_2 = \sum_{e = \{x,y\} \in E} \left[c_0(e)(S_x + S_y - 2c_0(e)) \right] , \qquad (8)$$

and this can clearly be performed in $O(n^2)$. □

2.2 Generalization to Higher Moments

The square term in the calculation of the variance ensures that it is sensitive only to the *magnitude* of a value's difference from the mean. The third moment is significant since it is sensitive to both the *sign* and *magnitude* of the difference. It thus encapsulates information about the symmetry of a distribution about the mean.

In terms of generalizing Theorem 4 to the rth moment, Lemmas 2 and 3 would need to be expanded to consider the number of tours in which r edges can occur, taking into account their various adjacency relationships. This seems a tractable task and will be the subject of future investigation.

3 The Relationship between the Depth and Problem Size

By the *depth* of a solution, π, we mean the number of standard deviations from the cost, $\Omega(\pi)$, of the solution to the mean cost of all tours. This motivates the following definition: the depth of a solution π of a TSP with mean tour costs μ and standard deviation of costs σ is

$$\operatorname{depth}(\pi) = \frac{\mu - \Omega(\pi)}{\sigma}. \tag{9}$$

3.1 Real World Problems

We examine two sets of real world problems. In the first case, 95 problem instances were taken from the well known TSPLIB database[17]. The second set, of 39 instances, originates from the genomics community and arises from the requirement to sort Radiation Hybrid (RH) data into likely gene sequences [7, Chapter 16][1]. The RH data problems are significant because they have no low dimensional embedding and because the problem domain requires large numbers of good, although not necessarily optimal solutions to be generated. The specific data set used was obtained from the RHDF9000 dog radiation hybrid panel [10][4], with each of the 39 TSP instances corresponding to the RH data over a single canine chromosome.

3.1.1 Empirical Results on TSPLIB Instances

Each of the problems we considered is under 6000 cities in size and has an approximate embedding on a two dimensional surface. In each case the mean and standard deviation of tour costs were determined by application of the methods discussed above.

Known optimal solution costs from [21] for the 95 cases were used to compare the depth of the optimal tours to the size of the problem. Figure 1 shows the results of this survey. It indicates a striking linear relationship between the depth of the optimal solutions and the square root of the problem sizes,$\sqrt{n}$. Indeed the Spearman's correlation coefficient between the two is 0.989 with a two tailed level of statistical significance of less than 0.001. It is also significant that this correlation is *stronger* than that observed between μ/σ and $\sqrt{n}$. The results of curve fitting using a linear least squares regression model are summarized in Table 1.

3.1.2 Empirical Results on Canine RH Instances

In each of the 39 cases the RH panel data was converted to a TSP using the CarthaGène package[19]. The resulting TSP instances have sizes ranging between 68 and 588 cities. For each instance the optimal solution cost was estimated using the well known Lin-Kernighan algorithm. Given the size of the problems it is probable that all of the solutions found are in fact optimal, and that any that are not, are within a few percent of optimal. This data coupled with the mean and standard deviation of tour costs was used to approximate the depth of the optimal solutions in each case. Again a striking correlation between the depth of the best solution seen and the square root of the problem size was found, the relationship being near linear as is evident in Figure 1. The Spearman's correlation coefficient between the two is 0.988 with a two tailed level of statistical significance of less than 0.001. Again the correlation was *stronger* than that between μ/σ and $\sqrt{n}$. The results of curve fitting with a least squares regression using a power model on $\sqrt{n}$ are summarized in Table 1.

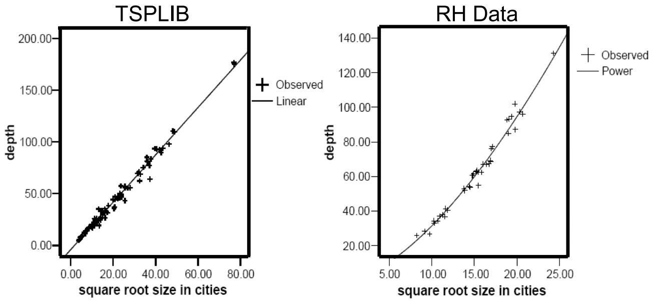

Figure 1: The relationship between the depth of the optimal solution and the square root of the problem size. Left, 95 instances from the TSPLIB problem set. Right, 39 instances each originating from canine RH panel data.

Problem	best fit	df.	ρ^2
TSPLIB	$\text{depth} = -4.27 + 2.29n^{0.5}$	93	0.988
canine RH	$\text{depth} = 0.806n^{0.796}$	37	0.982

Table 1: The results of curve fitting using least squares regression showing the relationship between the depth of the optimal solution and problem size in cities, n. In the case of the TSPLIB data a linear model on $\sqrt{n}$ was applied. In the case of the canine RH problems a near linear power model on $\sqrt{n}$ provided the best fit. The resulting best fit expressions are presented here in terms of n. In both cases the statistical significance is better than 0.001.

3.2 Application of these Relationships

The relationships noted above open the possibility of producing fast estimates of the likely optimal solution cost, particularly where there is knowledge of the application domain or other previously analyzed similar instances. This situation certainly arises in many application domains. It is anticipated that additional information about the skew of the distribution given by the third moment about the mean will improve the models provided above. In practical terms direct calculation of the third moment may prove to be too expensive for large problems and it may be necessary to resort to sampling techniques to estimate this statistic. However an expression giving the third moment would certainly aid the refinement of such an approach.

4 Conclusions

In this paper we have demonstrated that the TSP is well enough constrained to be amenable to statistical analysis. We have proved that the variance of tour costs over the solution space of the TSP with n cities may be found in time $O(n^2)$. The method provided is therefore linear on the number of edges of the instance. The combinatorial techniques employed invite generalization to determining other, higher order statistical moments and also show promise of being generalizable to variations of the TSP.

In the case of typically occurring 2 dimensional problem instances, we provide empirical evidence that there is a simple linear relationship between the square root of a TSP's size and the depth of its optimal solution. The *depth* in this case being, the number of standard deviations the optimal tour cost is below the mean. In other instances with no low dimensional embedding we find a near linear relationship between these parameters. In terms of direct application, we believe refinement of the results herein will provide a fast method to estimate the likely optimal solution cost of a problem.

Acknowledgements: The authors of this paper would like to thank Simon de Givry, Thomas Schiex and Christophe Hitte, for their kind provision of the canine RH data set and for its translation into TSP instances. In addition this paper benefited from the assistance of Jill Maddox and Tara Matise. The first author would like to extend his thanks to David Plyler, Len Hill and Neil Thomson.

References:

[1] Agarwala R.,Applegate, D.L., Maglott D., Schuler G.D., and Schaffer, A.A., A fast and scalable radiation hybrid map construction and integration strategy, *Genome Research* 10-8,2000,pp. 350-364.

[2] Giorgio Ausiello ,M. Protasi, A. Marchetti-Spaccamela ,G. Gambosi ,P. Crescenzi and V. Kann, *Complexity and Approximation: Combinatorial Optimization Problems and Their Approximability Properties*,Springer-Verlag New York 1999

[3] B.W. Colletti and J.W. Barnes, Local search structure in the symmetric travelling salesperson problem under a general class of rearrangement neighborhoods *Applied Mathematics Letters.*,14-1,2001,pp. 105-108.

[4] T. Faraut, S. de Givry, P. Chabrier, T. Derrien, F. Galibert, C. Hitte and T. Schiex, A comparative genome approach to marker ordering, *Proc. of ECCB-06*,2007,p7.

[5] John E. Freund, *Mathematical Statistics,* 2nd edn, Prentice Hall,1972

[6] L.K. Grover, Local search and the local structure of NP-complete problems, *Operations Research Letters*,12,1992,pp 235–243.

[7] Dan Gusfield, *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, Cambridge University Press 1997

[8] Gregory Gutin, Exponential neighbourhood local search for the traveling salesman problem *Comput. Oper. Res.*,26–4,1999,pp. 313–320.

[9] Gregory Gutin and Abraham P. Punnen, *Traveling Salesman Problem and Its Variations*, Kluwer Academic Publishers 2002

[10] Hitte C., Madeoy J., Kirkness EF., Priat C., Lorentzen T.D., Senger F., Thomas D., Derrien T., Ramirez C., Scott C., Evanno G., Pullar B., Cadieu E., Oza V., Lourgant K., Jaffe D.B.,Tacher S., Dreano S., Berkova N., Andre C., Deloukas P., Fraser C., Lindblad-Toh K.,Ostrander E.A., Galibert F. Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping, *Nature reviews. Genetics.* 6-8,2005,pp. 643-648.

[11] Juraj Hromkovič, *Algorithms for Hard Problems*,2nd edn,Springer,1998

[12] S. Kirkpatrick, D. Gelatt Jr. and M. P. Vecchi, Optimization by Simulated Annealing., *Science*,220–4598,1983,pp. 671-680

[13] Korte Bernhard, Vygen Jens, *Combinatorial Optimization*,Springer 2002

[14] E. L. Lawler ,J. K. Lenstra ,A. H. G. Rinnooy Kan and D. B. Shmoys, *The Traveling Salesman Problem*, John Wiley & Sons Ltd. 1985

[15] A. J. Orman and H. P. Williams, A Survey of Different integer programming formulations of the travelling salesman problem, *Research report 04.67* Department of Operational Research,London School of Economics and Political Science,2004.

[16] A. Punnen,F. Margot and S. Kabadi, TSP heuristics: Domination analysis and complexity, *Research report 2001-06, Dept. of Mathematics, Univ. of Kentucky*, March 2001

[17] Gerhard Reinelt, A Traveling Salesman Problem Library, *ORSA Journal on Computing* 3–4,1991,pp. 376–384.

[18] Gerhard Reinelt, *The traveling salesman: Computational solutions for TSP applications,* LNCS 840,Springer Verlag 1994

[19] de Givry, Simon, Bouchez Martin, Chabrier Patrick, Milan Denis, and Schiex Thomas, CARTHAGENE: multipopulation integrated genetic and radiated hybrid mapping, *Bioinformatics*,21–8,2005,pp. 1703–1704.

[20] Stirzaker David, *Elementary probability,* 2nd edn,Cambridge Univ. Press 2003

[21] Best known solutions for symmetric TSPs, `http://www.informatik.uni-heidelberg.de/groups/comopt/software/TSPLIB95/STSP.html`, March 17 2006