Communicated by Andrew Barto

An Analysis of the Elastic Net Approach to the Traveling Salesman Problem

Richard Durbin*
King's College Research Centre, Cambridge CB2 1ST, England

Richard Szeliski
Artificial Intelligence Center, SRI International, Menlo Park, CA 94025 USA

Alan Yuille
Division of Applied Sciences, Harvard University, Cambridge, MA 02138 USA

This paper analyzes the elastic net approach (Durbin and Willshaw 1987) to the traveling salesman problem of finding the shortest path through a set of cities. The elastic net approach jointly minimizes the length of an arbitrary path in the plane and the distance between the path points and the cities. The tradeoff between these two requirements is controlled by a scale parameter K. A global minimum is found for large K, and is then tracked to a small value. In this paper, we show that (1) in the small K limit the elastic path passes arbitrarily close to all the cities, but that only one path point is attracted to each city, (2) in the large K limit the net lies at the center of the set of cities, and (3) at a critical value of K the energy function bifurcates. We also show that this method can be interpreted in terms of extremizing a probability distribution controlled by K. The minimum at a given K corresponds to the *maximum a posteriori* (MAP) Bayesian estimate of the tour under a natural statistical interpretation. The analysis presented in this paper gives us a better understanding of the behavior of the elastic net, allows us to better choose the parameters for the optimization, and suggests how to extend the underlying ideas to other domains.

1 Introduction

The traveling salesman problem (Lawler *et al.* 1985) is a classical problem in combinatorial optimization. The task is to find the shortest possible tour through a set of N cities that passes through each city exactly once. This problem is known to be NP-complete, and it is generally believed

*Current address: Department of Psychology, Stanford University, Stanford, CA 94305 USA.

that the computational power needed to solve it grows exponentially with the number of cities. In this paper we analyze a recent parallel analog algorithm based on an elastic net approach (Durbin and Willshaw 1987) that generates good solutions in much less time.

This approach uses a fast heuristic method with a strong geometrical flavor that is based on the tea trader model of neural development (Willshaw and Von der Malsburg 1979). It will work in a space of any dimension, but for simplicity we will assume the two-dimensional plane in this paper. Below we briefly review the algorithm.

Let $\{\mathbf{X}_i\}$, $i = 1$ to N, represent the positions of the N cities. The algorithm manipulates a path of points in the plane, specified by $\{\mathbf{Y}_j\}$, $j = 1$ to M (M larger than N), so that they eventually define a tour (that is, eventually each city $\mathbf{X}_i$ has some path point $\mathbf{Y}_j$ converge to it). The path is updated each time step according to

$$\Delta \mathbf{Y}_j = \alpha \sum_i w_{ij}(\mathbf{X}_i - \mathbf{Y}_j) + \beta K(\mathbf{Y}_{j+1} + \mathbf{Y}_{j-1} - 2\mathbf{Y}_j)$$

where

$$w_{ij} = \frac{e^{-|\mathbf{X}_i - \mathbf{Y}_j|^2/2K^2}}{\sum_k e^{-|\mathbf{X}_i - \mathbf{Y}_k|^2/2K^2}},$$

α and β are constants, and K is the scale parameter. Informally, the α term pulls the path toward the cities, so that for each $\mathbf{X}_i$ there is at least one $\mathbf{Y}_j$ within distance approximately K. The β term pulls neighboring path points toward each other, and hence tries to make the path short. The update equations are integrable, so that $\Delta \mathbf{Y}_j = -K\partial E/\partial \mathbf{Y}_j$ for an "energy" function, E, given by

$$E(\{\mathbf{Y}_j\}, K) = -\alpha K \sum_i \log \sum_j e^{-|\mathbf{X}_i - \mathbf{Y}_j|^2/2K^2} + \beta \sum_j \{\mathbf{Y}_j - \mathbf{Y}_{j+1}\}^2 \quad (1.1)$$

For fixed K the path will converge to a (possibly local) minimum of E. At large values of K the energy function is smoothed and there is only one minimum. At small values of K, the energy function contains many local minima, all of which correspond to possible tours of the cities (we prove this later in the paper), and the deepest minimum is the shortest possible tour. The algorithm proceeds by starting at large K, and gradually reducing K, keeping to a local minimum of E (see Fig. 1). We would like this minimum that is tracked to remain the global minimum as K becomes small. Unfortunately, this can not be guaranteed (see section 3).

The elastic net approach is similar to a number of previously developed algorithms that use elastic matching (Burr 1981), energy-based matching (Terzopoulos *et al.* 1987), or topographic mapping (Kohonen 1988) to solve vision, speech, and neural development problems. Alternative parallel analog algorithms have also recently been proposed for solving the traveling salesman problem (Hopfield and Tank 1985; Angéniol *et al.* 1988). The method of Angéniol *et al.* is closely related to that discussed here, but is based on Kohonen's self-organization algorithm (Kohonen 1988). It is faster, but for large problems it is marginally less accurate than the elastic method.

An important contribution of this paper is to analyze the behavior of the energy function as the constant K changes and to describe the energy landscape. In particular, we prove results about the behavior of the function for large and small K, confirming the assertions made above about how the algorithm works. First, however, we show that

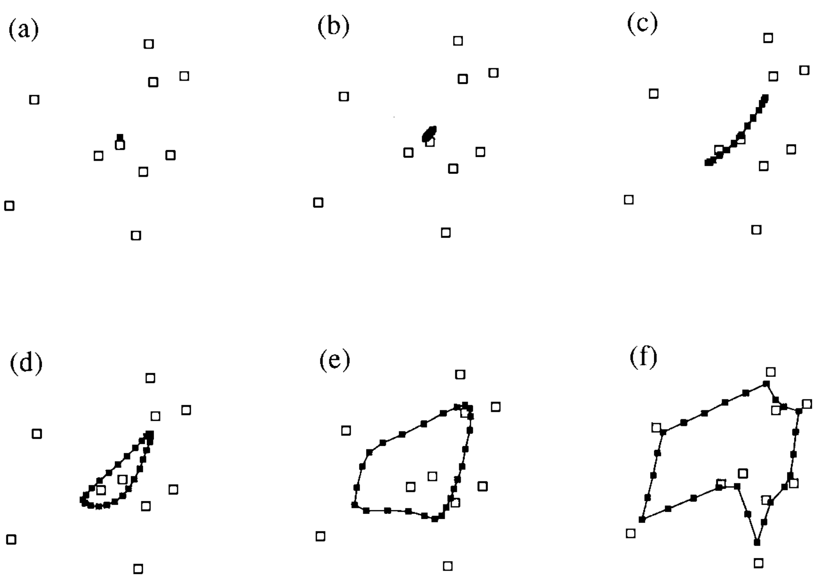

Figure 1: The convergence of the network as a function of K. The white and black squares represent the data (10 cities) and the network (25 path points), respectively. The six figures (a–f) show the configuration found by the network at values $K = 0.261, 0.26, 0.21, 0.20, 0.12, 0.04$. The data set is centered on (0.49, 0.46) and has second-order moments $(K_{xx}, K_{xy}, K_{yy}) = (0.75, -0.23, 0.70)$. We use $\alpha = 0.2$ and $\beta = 1.0$. The first bifurcation, when the origin becomes unstable, can be calculated to occur at $K = 0.2606$ (see Section 5), in agreement with the simulation. The second break temperature is between $K = 0.21$ and $K = 0.20$ for the simulation when the line spreads into a loop. This corresponds well to the point at which the second eigenvalue becomes negative ($K = 0.196$). The correspondence is not exact because nonlinear terms become significant after the first break.

minimizing this cost function can be interpreted in terms of maximizing a probability distribution.

2 The Probabilistic Interpretation

The energy function (1.1) can be related to a probability distribution by exponentiation. This is analogous to use of the Gibbs distribution in statistical mechanics.

$$\begin{aligned} L(\{\mathbf{Y}_j\}, K) &= \frac{1}{(2\pi)^N K^{2N} M^N} e^{-E/\alpha K} \\ &= \prod_{i=1}^{N} \frac{1}{M}\{\sum_{j=1}^{M} \frac{1}{2\pi K^2} e^{-|\mathbf{X}_i - \mathbf{Y}_j|^2/2K^2}\} \prod_{j=1}^{M} e^{-\frac{\beta}{\alpha K}\{\mathbf{Y}_j - \mathbf{Y}_{j+1}\}^2} \end{aligned}$$

Observe that minimizing E with respect to $\{\mathbf{Y}_j\}$ corresponds to maximizing L with respect to $\{\mathbf{Y}_j\}$.

We can interpret L in terms of Bayes' theorem, which states that

$$P(\mathbf{Y}|\mathbf{X}) = \frac{P(\mathbf{X}|\mathbf{Y})P(\mathbf{Y})}{P(\mathbf{X})} \tag{2.1}$$

where $P(\mathbf{Y}|\mathbf{X})$ is the probability of a tour $(\mathbf{Y})$ given a set of cities $(\mathbf{X})$. Our algorithm maximizes $P(\mathbf{Y}|\mathbf{X})$ over all possible tours $(\mathbf{Y})$, so the value of $P(\mathbf{X})$ is irrelevant.

The distribution

$$P(\mathbf{Y}) = \prod_{j=1}^{M} e^{-\beta/\alpha K\{\mathbf{Y}_j - \mathbf{Y}_{j+1}\}^2} \tag{2.2}$$

is the a priori probability of a given tour. This distribution is a correlated gaussian that assigns greater prior probability to shorter tours. The distribution

$$P(\mathbf{X}|\mathbf{Y}) = \prod_{i=1}^{N} \frac{1}{M}\{\sum_{j=1}^{M} \frac{1}{2\pi K^2} e^{-|\mathbf{X}_i - \mathbf{Y}_j|^2/2K^2}\} \tag{2.3}$$

is the probability of the cities being at $(\mathbf{X})$ given that the tour points are at $(\mathbf{Y})$.

$P(\mathbf{X}|\mathbf{Y})$ is the product of N independent probability distributions

$$P(\mathbf{X}_i|\mathbf{Y}) = \frac{1}{M} \sum_{j=1}^{M} \frac{1}{2\pi K^2} e^{-|\mathbf{X}_i - \mathbf{Y}_j|^2/2K^2}$$

This equation is equivalent to assuming that the measured position of city $\mathbf{X}_i$ was actually derived from one of the tour points in $\{\mathbf{Y}_j\}$ with a two-dimensional gaussian error of variance K^2, but without knowing which tour point $\mathbf{X}_i$ corresponds to. Thus equation 2.1 shows that the elastic net algorithm is computing the most probable tour (finding the

Bayesian MAP estimate) given a prior model (2.2) that favors short tours and a sensor model (2.3) with two-dimensional position uncertainty. Our method thus has an obvious extension to surface interpolation and three-dimensional surface modeling where the correspondence between surface points and measured data points is unknown.

3 Tracking a Minimum

The algorithm devised by Durbin and Willshaw minimizes E at large K and then tracks the minimum energy solution down to small K. At a local minimum (or any extremum), we have

$$\frac{\partial E}{\partial Y_i^\mu} = 0$$

where μ is 1 or 2 and Y_i^1 and Y_i^2 are the x and y components of the position vector $\mathbf{Y}_i$. As we follow the extrema as K changes we get the equation

$$\frac{d}{dK}\left(\frac{\partial E}{\partial Y_i^\mu}\right) = 0$$

which becomes

$$\frac{\partial^2 E}{\partial K \partial Y_i^\mu} + \frac{\partial^2 E}{\partial Y_i^\mu \partial Y_j^\nu}\frac{dY_j^\nu}{dK} = 0$$

To obtain the trajectory we must solve this equation for dY_j^ν/dK. When we are at a true minimum, the Hessian $\partial^2 E/\partial Y_i^\mu \partial Y_j^\nu$ is a positive definite matrix and can be inverted, enabling us to compute dY_j^ν/dK. Bifurcations occur when the Hessian has zero eigenvalues. In this case dY_j^ν/dK is underdetermined and there are several possible solutions.

Computer simulations and our calculations in the large K limit (see below) show that such a bifurcation occurs as the tour initially spreads out from a point. After this, our simulations suggest that the minimum tracks smoothly with K. Other minima of the energy function also appear as K is reduced. For the configuration shown in Figure 1, the number of minima increases rapidly from 1 at $K = 0.12$ to 3 at $K = 0.10$, 9 at $K = 0.08$ (shown in Fig. 2), and very many at $K < 0.05$. The minimum found by tracking K from large to small values is *not* necessarily the optimal (global) minimum (Fig. 2). Nevertheless, empirically the minima found are within a few percent of optimal (Durbin and Willshaw 1987). One possible improvement would be to pick up and track nearby minima by local random perturbation as in simulated annealing (Kirkpatrick *et al.* 1983).

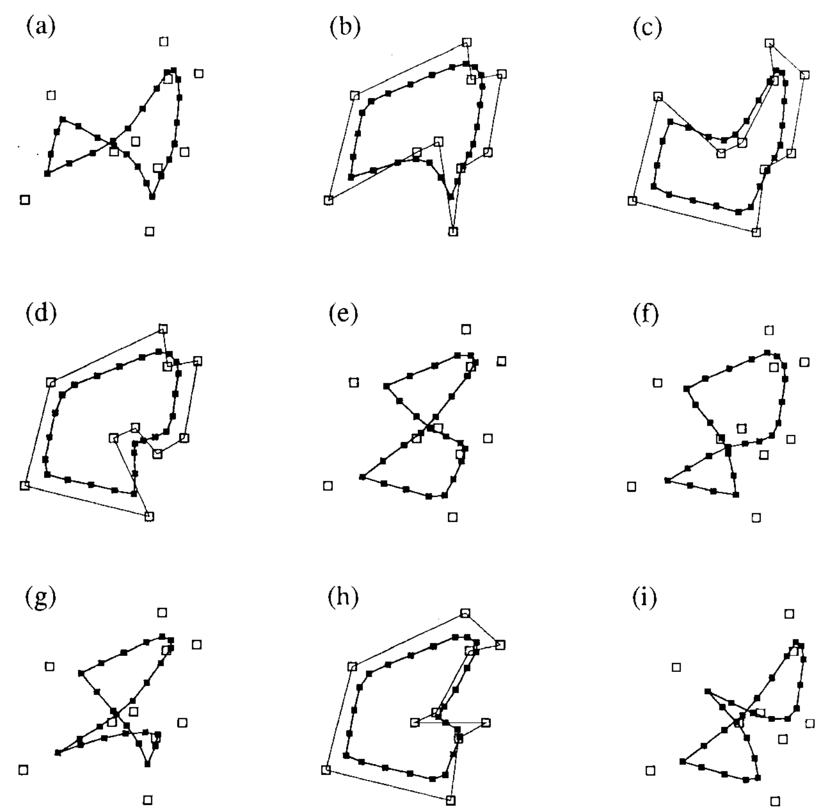

Figure 2: Possible minima of the energy function for the data in Figure 1. To investigate the energy minima we started 1000 simulations at $K = 0.08$ with random initial configurations, hence *without* using the solutions for larger K as initial conditions. In cases that lead to a sensible tour on subsequent slow reduction of K, the lines joining the white boxes (data points) show the tour found by the network. We found nine distinct groups of minima with the following frequencies: (a) 260, (b) 183 (tour length 3.356), (c) 140 (tour length 3.288), (d) 130 (tour length 3.350), (e) 95, (f) 72, (g) 48, (h) 36 (tour length 3.420), and (i) 36. We have probably found all the major forms of minima, since the least frequent happened 36 times. Note that the most frequent pattern (a) is neither the one obtained from tracking K (b) nor the optimal one (c).

4 The Small K Limit

We now consider the behavior of the extrema and the Hessian at these extrema as $K \to 0$. We will show that the only stable extrema occur when each $\{\mathbf{X}_i\}$ has at least one $\{\mathbf{Y}_j\}$ arbitrarily close to it. Thus, the extrema all correspond to possible tours.

For any given i let

$$B(K) = \min_j |X_i - Y_j|$$

Then

$$\sum_j e^{-|\mathbf{X}_i - \mathbf{Y}_j|^2/2K^2} \leq M e^{-B^2(K)/2K^2}$$

and

$$-\alpha K \log \sum_j e^{-|\mathbf{X}_i - \mathbf{Y}_j|^2/2K^2} \geq -\alpha K \log M + \frac{B^2(K)\alpha}{2K}$$

Thus, for the energy to be bounded we must have

$$\lim_{K \to 0} \frac{\{B^2(K) - 2K^2 log M\}\alpha}{2K} = C$$

where C is a constant and hence $B(K) = O(K^{1/2})$.

Thus, in the limit as $K \to 0$, configurations with unmatched **X**s will have arbitrarily high energy, and so will not be found by the algorithm. This means that the minima will all correspond to possible tours. Although all the **X**s are matched there is no requirement that all the **Y**s are matched. Indeed, with correct choice of parameters α and β it can be shown that there will be only one tour point at each city. The remaining tour points space themselves evenly in the intercity intervals. A sufficient requirement on the parameters for this to happen is that

$$\frac{\alpha}{\beta} < A$$

where A is the shortest distance from a tour point being attracted to some city to any tour point not being attracted to that city.

To derive this condition on the parameters consider a single city situated at the origin. Define w_j by

$$w_j = \frac{e^{-|\mathbf{Y}_j|^2/2K^2}}{\sum_k e^{-|\mathbf{Y}_k|^2/2K^2}}$$

Assume the $\mathbf{Y}_j$ are at equilibrium. We wish to consider the stability of an equilibrium in which there are two w_j that stay significantly nonzero as $K \to 0$ (to have a single tour point converge to each city we require instability). We can choose K sufficiently small that there is no significant interaction between these two tour points and any other cities. Consider

perturbations $\mathbf{Y}'_j = \mathbf{Y}_j + \varepsilon_j$. We want to find eigenvectors such that

$$\begin{aligned}
\lambda\varepsilon_j &= -K\frac{\partial E}{\partial \mathbf{Y}'_j} \\
&= -\alpha w'_j \mathbf{Y}'_j + \beta K(\mathbf{Y}'_{j+1} + \mathbf{Y}'_{j-1} - 2\mathbf{Y}'_j) \\
&= -\alpha\{w_j\varepsilon_j + \mathbf{Y}_j \sum_k (\frac{\partial w_j}{\partial \mathbf{Y}_k} \cdot \varepsilon_k)\} + \beta K O(\epsilon) + O(\epsilon^2) \\
&= -\alpha\{w_j\varepsilon_j + \sum_k \mathbf{Y}_j \frac{1}{K^2}(w_j w_k \mathbf{Y}_k - \delta_{jk} w_j \mathbf{Y}_j)\cdot\varepsilon_k\} + \beta K O(\epsilon) + O(\epsilon^2) \\
&= -\frac{\alpha w_j}{K^2}\{K^2\varepsilon_j + \mathbf{Y}_j \sum_k w_k(\mathbf{Y}_k \cdot \varepsilon_k) - \mathbf{Y}_j(\mathbf{Y}_j \cdot \varepsilon_j)\} + \beta K O(\epsilon) + O(\epsilon^2)
\end{aligned}$$

where $\epsilon = \max_j(|\varepsilon_j|)$. In the limit of small K we can ignore the β term as well as the higher order ϵ term. Clearly then for each j $\bar{\varepsilon}_j = \epsilon_j \mathbf{Y}_j$ for some scalar ϵ_j. Let $\mu = -\lambda K^2/\alpha$, and consider the case where only w_1, w_2 are significant. This leads to the eigenvalue equation

$$\begin{pmatrix} w_1\left[K^2 - (1 - w_1)Y_1^2\right] - \mu & w_1 w_2 Y_2^2 \\ w_1 w_2 Y_1^2 & w_2\left[K^2 - (1 - w_2)Y_2^2\right] - \mu \end{pmatrix} \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} = 0$$

The criterion for instability is that at least one λ is positive, and hence that the corresponding μ is negative. For large K both eigenvalues μ_1 and μ_2 are clearly positive. Hence for instability we require that K should be below the value at which one eigenvalue (and hence their product) goes negative. The product is

$$\begin{aligned}
\mu_1\mu_2 &= w_1 w_2\left[(K^2 - w_2 Y_1^2)(K^2 - w_1 Y_2^2) - w_1 w_2 Y_1 Y_2\right] \\
&= w_1 w_2 K^2\left[K^2 - (w_2 Y_1^2 + w_1 Y_2^2)\right]
\end{aligned}$$

Therefore we require that

$$K^2 < w_2 Y_1^2 + w_1 Y_2^2$$

Since $w_1 + w_2 = 1$ we will be safe if $\min|\mathbf{Y}_j| > K$. At equilibrium at small K we have $\mathbf{Y}_j = (\beta K/\alpha w_j)\mathbf{A}_j$ where $\mathbf{A}_j = \mathbf{Y}_{j+1} + \mathbf{Y}_{j-1}$. When, as will be usual, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are neighbors, then $|\mathbf{A}_j|$ is just the distance to the next path point not converging on the city, and a sufficient requirement is that this distance must be greater than α/β.

Since an average tour on N cities has length $(N/2)^{1/2}$ a safe estimate for $A_{\min}$ would be $0.2(N/2)^{1/2}/M$. Alternatively one could choose α to be a decreasing function of K, such as K^p where p is a fixed exponent between zero and one.

5 At Large K

For large K, the energy function (1.1) has a minimum corresponding to the net lying at the center of the cities. At a critical value of K this minimum becomes unstable and the system bifurcates. The initial movement of the net depends only on the second-order moments of the cities.

To show this, we first calculate the first and second derivatives of E with respect to $\mathbf{Y}$.

$$\frac{\partial E}{\partial Y_k^\mu} = \frac{\alpha}{K}\sum_{i=1}^{N}\frac{(Y_k^\mu - X_i^\mu)e^{-|\mathbf{X}_i-\mathbf{Y}_k|^2/2K^2}}{(\sum_{j=1}^{M} e^{-|\mathbf{X}_i-\mathbf{Y}_j|^2/2K^2})} + 2\beta\{2Y_k^\mu - Y_{k+1}^\mu - Y_{k-1}^\mu\} \tag{5.1}$$

$$\begin{aligned}\frac{\partial^2 E}{\partial Y_k^\mu \partial Y_l^\nu} = {} & \frac{\alpha}{K}\sum_{i=1}^{N}\frac{\delta^{\mu\nu}\delta_{kl}e^{-|\mathbf{X}_i-\mathbf{Y}_k|^2/2K^2}}{(\sum_{j=1}^{M} e^{-|\mathbf{X}_i-\mathbf{Y}_j|^2/2K^2})} \\ & + \frac{\alpha}{K^3}\sum_{i=1}^{N}\frac{(Y_k^\mu - X_i^\mu)(Y_l^\nu - X_i^\nu)e^{-|\mathbf{X}_i-\mathbf{Y}_k|^2/2K^2}e^{-|\mathbf{X}_i-\mathbf{Y}_l|^2/2K^2}}{(\sum_{j=1}^{M} e^{-|\mathbf{X}_i-\mathbf{Y}_j|^2/2K^2})^2} \\ & - \frac{\alpha}{K^3}\sum_{i=1}^{N}\frac{(Y_k^\mu - X_i^\mu)(Y_k^\nu - X_i^\nu)e^{-|\mathbf{X}_i-\mathbf{Y}_k|^2/2K^2}\delta_{kl}}{(\sum_{j=1}^{M} e^{-|\mathbf{X}_i-\mathbf{Y}_j|^2/2K^2})} \\ & + 2\beta\delta^{\mu\nu}\{2\delta_{kl} - \delta_{k+1\,l} - \delta_{k-1\,l}\}\end{aligned} \tag{5.2}$$

By substituting $\mathbf{Y}_j = 0$ into (5.1) we find that

$$\frac{\partial E}{\partial Y_k^\mu} = \frac{\alpha}{K}\sum_{i=1}^{N}\frac{(-X_i^\mu)e^{-|\mathbf{X}_i|^2/2K^2}}{(\sum_{j=1}^{M} e^{-|\mathbf{X}_i|^2/2K^2})} = -\frac{\alpha}{KM}\sum_{i=1}^{N} X_i^\mu$$

The origin is thus always an extremum, provided that it is chosen at the center of the $\mathbf{X}$s (i.e., $\sum_{i=1}^{N}\mathbf{X}_i = 0$).

To show that the center is a minimum for very large K, we must calculate the eigenvalues of the Hessian. As K decreases, this minimum becomes unstable and a bifurcation occurs. Knowing the value of K at which this occurs will give us a useful starting value for K when we are running the elastic net algorithm. At the origin the Hessian can be written as

$$\begin{aligned}\frac{\partial^2 E}{\partial Y_k^\mu \partial Y_l^\nu} = {} & \frac{\alpha N}{MK}\delta^{\mu\nu}\delta_{kl} + \frac{\alpha}{K^3 M^2}\sum_{i=1}^{N} X_i^\mu X_i^\nu \\ & - \frac{\alpha}{K^3 M}\delta_{kl}\sum_{i=1}^{N} X_i^\mu X_i^\nu + 2\beta\delta^{\mu\nu}\{2\delta_{kl} - \delta_{k+1\,l} - \delta_{k-1\,l}\}\end{aligned} \tag{5.3}$$

For large K, the eigenvalues of the Hessian are clearly all positive. By inspection of equation 2, we see that throughout the region $|\mathbf{Y}_j| \ll K$ the Hessian is positive definite and so the origin is a unique minimum. For small K, the dominant terms (the second and third terms on the right-hand side of 5.2) have negative trace, so there are some negative eigenvalues. Thus, the origin is a stable state for large K but then becomes unstable as K decreases. To see how this occurs we must explicitly calculate the eigenvectors.

If we compute the eigenvalues of the Hessian (Durbin *et al.* 1989), we find that smallest eigenvalue is

$$\lambda_{\text{min}} = \frac{\alpha N}{KM} - \frac{\alpha\lambda}{K^4 M} + 8\beta\sin^2\frac{\pi}{M} \tag{5.4}$$

where λ is the principal eigenvalue of the city covariance matrix. The center then becomes unstable and breaks at K s.t. $\lambda_{\text{min}} = 0$. This can be calculated from equation 5.4. Since the eigenvectors depend only on the second-order moments of the distribution of the cities the global minimum for K just below the critical value will also depend chiefly on the second-order moments. As K decreases further the higher order moments will become important. These theoretical results are confirmed by computer simulations (see Fig. 1). The net stays at the origin until the critical value of K and then forms a line along the principal axis of the city distribution. Near the second critical value of K (when the eigenvalue determined by substituting the minor eigenvalue of the city covariance matrix into equation 5.4 becomes negative) the line spreads into a loop.

6 Conclusion

In summary, we have obtained several theoretical results concerning the elastic net method. First, we have shown how the elastic net solution can be interpreted as a *maximum a posteriori* estimate of an unknown tour (circular curve), where some points along the tour have been measured with gaussian uncertainty in position. Second, we have proved that for small K every point $\mathbf{X}_i$ is matched, and that each point must be within $O(K^{1/2})$ of a tour point. Third, we have found a condition on the parameters α, β under which each city becomes matched by only one tour point. Fourth, we have shown that at large K, a single minimum exists for the energy function, with all of the tour points lying at the center of gravity of the cities. Fifth, we have shown how to calculate the bifurcation points for the elastic net as K is reduced.

The first result is particularly interesting since it suggests that this approach can be applied to other interpolation, approximation, and matching problems (such as surface interpolation in computer vision). The important feature here is that we do not need to prespecify which model point matches a data point, allowing "slippery" matching. The second result proves the "correctness" of the elastic net method, in that any final solution must be a valid tour. The third, fourth, and fifth results can be used in selecting parameter values, a starting configuration for the net, and a starting value for K. The elastic net method that we have analyzed provides a simple, effective, and intuitively satisfying algorithm for generating good traveling salesman tours. We believe that similar continuation-based algorithms can be applied to a wide range of optimization and approximation problems.