Time-Dynamic Density and Mode Estimation with Application to Fast Mode Tracking

Peter Hall, Hans-Georg Müller, and Ping-Shi Wu

February 20, 2005

Abstract. We introduce a nonparametric time-dynamic kernel type density estimate for the situation where an underlying multivariate distribution evolves with time. Based on this time-dynamic density estimate, we propose nonparametric estimates for the time-dynamic mode of the underlying distribution. Our estimators involve boundary kernels for the time dimension so that the estimator is always centered at current time, and multivariate kernels for the spatial dimension of the time-evolving distribution. Under certain mild conditions, the asymptotic behavior of density and mode estimators, especially their uniform convergence in both time and space, is derived. A time-dynamic algorithm for mode tracking is proposed, including automatic bandwidth choices, and is implemented via a mean update algorithm. Simulation studies and real data illustrations demonstrate that the proposed methods work well in practice.

Key words and phrases: Boundary kernel, Mean-shift algorithm, Modal Evolution, Nonparametric estimation, Spatio-temporal modeling, Uniform convergence.

Peter Hall is at the Center for Mathematics and Its Applications, Australian National University, Canberra, ACT 0200, Australia (E-mail:halpstat@maths.anu.edu.au). Hans-Georg Müller (E-mail:mueller@wald.ucdavis.edu) and Ping-Shi Wu (E-mail:pswu@wald.ucdavis.edu) are at the Department of Statistics, University of California Davis, Davis, CA 95616. Research supported in part by NSF grants DMS02-04869 and DMS03-54448. We are grateful for the insightful remarks of two referees that led to improvements in procedures and presentation.

1 Introduction

While on-line density estimation has been well studied in the recursive density estimation framework (see, for example, Wegman and Davies 1995; Hall and Patil 1994; Hossjer and Holst 1995), time-dynamic density estimation so far has not received much attention in statistics. A major difference between time-dynamic density estimation and on-line density estimation is that in the former, the data collected at different times are independently but not identically distributed, since the distribution generating the data is changing over time, while in the latter it is assumed that the observed data are independently and identically distributed throughout. Recently, with the increasing prevalence of data-streaming sources, the importance of time-dynamic estimation with the goal of capturing changing trends and features in distributions is increasingly recognized (Elgammal et al. 2003; Wegman and Marchette 2003).

Real-time density estimation has been treated in Hall and Patil (1994), using univariate kernel density estimates. We address here the multivariate (especially bivariate) situation of streaming data where the distribution that generates the observed data is evolving as time progresses. Such situations occur in spatial statistics when time exerts an influence on the distributions of interest, and suitable spatio-temporal models are of interest. Examples are the occurrence of spatial event locations such as earthquakes, tornados, or diseases that are spatially localized and where we may suspect that the spatial distribution changes over time. The changing features of these distributions are then of interest. The methods we propose are designed for p-variate distributions that change over time.

It is often important to quickly capture trends of temporal changes in spatial distribu-

tions, and a natural characteristic to observe is the mode of the underlying distributions. As the distributions change, so too does the mode, and this leads us to consider the problem of time-dynamic mode-tracking. The modes that change over time form a modal evolution path, which is a p-variate curve that is indexed by time. It is of special interest to estimate current mode locations when the observed data are only available from the past, where the distribution that generates the data increasingly differs from the current distribution. We develop an algorithm for time-dynamic mode tracking and illustrate it with simulated data, stock market data and spatial epidemiological data.

The estimation of the density and its mode at current time uses data from a historical time window. These data do not come from the current distribution of interest but rather from earlier distributions. By using the assumption of slowly changing distributions, and employing temporal boundary kernels that are centered at the current time of interest, we extract information from the observations in the historical time window about current density and mode. The range of history to be included depends on a temporal bandwidth, the choice of which will be discussed below. Our algorithm also requires a spatial bandwidth. The smoothing weights that we will apply are derived from a product of temporal boundary and multivariate spatial kernel functions. For fast mode tracking, the mean shift algorithm (Fukunaga and Hostetler 1975; Cheng 1995; Comaniciu and Meer 2002) provides a convenient implementation that will be illustrated below. This algorithm enables fast updating of a current mode location by iteratively obtaining the sample mean of the observations in a suitable spatio-temporal neighborhood with the current mode location at its center.

This article is organized as follows. The construction of the proposed density estimate

and its uniform convergence to the true density function are described in section 2. Section 3 contains the time-dynamic mode estimators and a result on the uniform convergence in time of the estimated modal function to the target modal function. The mean shift implementation of dynamic mode-tracking is described in section 4, which also contains a proposal for automatic bandwidth selection. Simulations and data applications are the theme of section 5. The proofs of the theoretical results are collected in an appendix.

2 Time-Dynamic Density Estimation

The evolution of time is modeled over a discrete grid of asymptotically dense times $t_i = t_{i,n}$, which permits us to exploit the smoothness of the underlying distributions when viewed as functions of time. It is assumed that at each of these times a random number of observations is made, such as internet messages or new cases of an epidemic originating at that point of time. These observations form a sample from $F(\cdot, t_i)$, the distribution applicable at time t_i.

Formally, let $\{\mathbf{X}_{ij}\}_{j=1}^{M_i}$ be a sequence of independent and identically distributed multivariate random variables observed at time $t_i = \frac{i}{n},\ 1 \leq i \leq n$, $\mathbf{X}_{ij} \in \mathbb{R}^p$, and $\{\mathbf{X}_{ij}\}_{j=1}^{M_i}$ be distributed as $F(\cdot, t_i)$. For all i, M_i is a discrete random variable. We also make the following assumptions:

(A1) For $1 \leq i \neq i' \leq n$, $1 \leq j \leq M_i$ and $1 \leq j' \leq M_{i'}$, $\mathbf{X}_{ij}$ and $\mathbf{X}_{i'j'}$ are independent.

(A2) For each t_i, $\mathbf{X}_{ij}$ has a twice differentiable density function $f(\mathbf{x}, t_i) : \mathbb{R}^p \times [0, 1] \mapsto \mathbb{R}$.

(A3) $|f(\mathbf{x}, t_2) - f(\mathbf{x}, t_1)| \leq L|t_2 - t_1|$, for $\mathbf{x} \in \mathbb{R}^p,\ L \in \mathbb{R}^+$.

For constructing the kernel density estimate, we use a product kernel that consists of a

spatial kernel $K_{\mathrm{S},\mathbf{H}}$ which corresponds to a regular p-dimensional nonnegative kernel function, and a temporal kernel $K_{\mathrm{T},\lambda} = \lambda^{-1}K_{\mathrm{T}}(v/\lambda)$ which is a boundary kernel. Specifically, $K_{\mathrm{S},\mathbf{H}}(\mathbf{u}) = |\mathbf{H}|^{-1/2}K_{\mathrm{S}}(\mathbf{H}^{-1/2}\mathbf{u})$ is a spherically symmetric multivariate kernel function with a finite support, a bounded Hölder-continuous derivative, and satisfying several conditions stated below, where $\mathbf{H}$ is a $p \times p$ bandwidth matrix. For theorems 1 and 2 below we assume bandwidth matrices of diagonal type $\mathbf{H} = \mathrm{diag}\,(h_1^2, \ldots, h_p^2)$ without substantial loss of efficiency for investigating a unimodal density function; compare Wand and Jones (1993, 1995) regarding this simplifying choice. Practical implementation of the proposed method is based on smoothing matrices $\mathbf{H}$ that are not necessarily diagonal; see section 4.3.

The spatial kernel function is assumed to satisfy

(A4)
$$\int_{\mathbb{R}^p} K_{\mathrm{S}}(\mathbf{x})d\mathbf{x} = 1 \quad , \quad \lim_{\|\mathbf{x}\|\to\infty} \|\mathbf{x}\|^p K_{\mathrm{S}}(\mathbf{x}) = 0,$$
$$\int_{\mathbb{R}^p} \mathbf{x}^\alpha K_{\mathrm{S}}(\mathbf{x})d\mathbf{x} = m_\alpha < \infty,\ \alpha \in N_0 \quad , \quad \int_{\mathbb{R}^p} \mathbf{x}^\alpha K_{\mathrm{S}}(\mathbf{x})d\mathbf{x} = 0,\ \alpha \notin N_0,$$

where for the multiindex $\alpha = (\alpha_1, \ldots, \alpha_p)$, $\alpha_i \geq 0$ for $1 \leq i \leq p$, $\mathbf{x}^\alpha = \prod_{i=1}^{p} x_i^{\alpha_i}$ for $\mathbf{x} \in \mathbb{R}^p$, and N_0 is an index set in which all p entries of each member are either 0 or even integers. The temporal kernel is a boundary kernel in the sense that it has one-sided support with 0 as an endpoint and satisfies the usual moment conditions for a second order kernel.

(A5)
$$\int_0^1 K_{\mathrm{T}}(t)dt = 1, \quad \int_0^1 tK_{\mathrm{T}}(t)dt = 0, \quad \int_0^1 t^2 K_{\mathrm{T}}(t)dt = v_T < \infty.$$

Various classes of such boundary kernels have been discussed in the literature (e.g., Gasser et al. 1985; Müller 1991; Jones 1993). A typical example is the kernel $K_{\mathrm{T}}(t) = 4-6t$, $t \in [0, 1]$.

Let

$$\hat{f}(\mathbf{z},t) = \frac{1}{N(t)\,\lambda\,|\mathbf{H}|^{1/2}} \sum_{1\le t_i\le t} K_{\mathrm{T}}\left(\frac{t-t_i}{\lambda}\right) \sum_{j=1}^{M_i} K_{\mathrm{S}}\left(\mathbf{H}^{-1/2}(\mathbf{z}-\mathbf{X}_{ij})\right) \tag{1}$$

denote the time-dynamic kernel density estimator of $f(\mathbf{z},t)$, where $N(t) = \sum_{1\le t_i\le t} M_i$. In the following, we simplify the estimator by stipulating equal spatial bandwidths $h_i = h$, obtaining:

$$\hat{f}(\mathbf{z},t) = \frac{1}{N(t)h^p\lambda} \sum_{1\le t_i\le t} K_{\mathrm{T}}\left(\frac{t-t_i}{\lambda}\right) \sum_{j=1}^{M_i} K_{\mathrm{S}}\left(\frac{\mathbf{z}-\mathbf{X}_{ij}}{h}\right). \tag{2}$$

Let $\mathcal{K}$ denote any compact subset of $\mathbb{R}^p$, and $\mathcal{I}$ any compact subset of $(0,1]$. We establish the following result regarding uniform convergence of our proposed time-dynamic kernel density. Uniform consistency is achieved over both spatial and temporal domains.

Theorem 1. *Assume (A1)-(A5), that K_{S} is a spherically symmetric and compactly supported kernel, with a bounded, Hölder-continuous derivative, that K_{T} is a second-order boundary kernel supported on a compact subset of $[0,\infty)$, that $M_1, M_2, \ldots$ are independent and identically distributed random variables with mean $m > 0$ and all moments finite, independent of the $\mathbf{X}_{ij}$'s, and that $f(\mathbf{z},t)$, viewed as a $(p+1)$-variate function on $\mathbb{R}^p \times (0,1]$, has two bounded derivatives. Take each $h_i = h$. If $h = h(n) \downarrow 0$, $\lambda = \lambda(n) \downarrow 0$ and $n^{1-\epsilon}h^p\lambda \to \infty$ for some $\epsilon > 0$, then*

$$\hat{f}(\mathbf{z},t) = f(\mathbf{z},t) + O\Big\{\left(nh^p\lambda\right)^{-1/2}(\log n)^{1/2} + h^2 + \lambda^2\Big\} \tag{3}$$

uniformly in $(\mathbf{z},t) \in \mathcal{K}\times\mathcal{I}$, with probability 1.

3 Time-Dynamic Mode Estimation

Mode estimation by taking the location of the maximum of the density estimate in univariate setup was first proposed by Parzen (1962). Among others, Epanechnikov (1969) and Konakov (1973) extended the univariate result to the multivariate situation. We propose here an extension of Parzen's idea of mode estimation to the case of a time-dynamic mode estimator. Assume $f(\cdot, t)$ is a family of unimodal density functions, indexed by time t, and define the time-dynamic mode at time t as

$$\theta_t = \inf_{x_1 \in \mathbb{R}} \cdots \inf_{x_p \in \mathbb{R}} \arg\max_{\mathbf{x} \in \mathbb{R}^p} f(\mathbf{x}, t).$$

The proposed time-dynamic mode estimator at time t will then be defined as:

$$\hat{\theta}_t = \inf_{x_1 \in \mathbb{R}} \cdots \inf_{x_p \in \mathbb{R}} \arg\max_{\mathbf{x} \in \mathbb{R}^p} \hat{f}(\mathbf{x}, t), \tag{4}$$

where the infima in the last expression are taken sequentially over each direction of $\mathbf{x} \in \mathbb{R}^p$, to define a unique location of the maximum. We refer to θ_t as the modal evolution over time.

Next we present a result on the convergence of the first derivative of the density function estimate. This leads to uniform convergence of the modal evolution estimate $\hat{\theta}_t$ towards the target θ_t over time, including rates of convergence. Let $\mathbf{D}_{e_j} f(\mathbf{z}, t)$ denote $(\partial/\partial z_j) f(\mathbf{z}, t)$, and analogously for the corresponding estimates.

Theorem 2. *If in addition to the conditions in Theorem 1, $f(\mathbf{z}, t)$ has three bounded derivatives, and $n^{1-\epsilon} h^{p+3} \to \infty$ for some $\epsilon > 0$, then*

$$\mathbf{D}_{e_j} \hat{f}(\mathbf{z}, t) = \mathbf{D}_{e_j} f(\mathbf{z}, t) + O\Big\{ \left(nh^{p+3}\right)^{-1/2} (\log n)^{1/2} + h^2 \Big\} \tag{5}$$

uniformly in $j = 1, \ldots, p$ and $(\mathbf{z}, t) \in \mathcal{K} \times \mathcal{I}$, with probability 1.

If moreover, for each $t \in (0,1]$, $f(\cdot,t)$ *has a unique mode* θ_t *and if, for all sufficiently small* $\epsilon > 0$, *the Hessian of* $f(\mathbf{z},t)$ *is bounded away from zero uniformly in* $\mathbf{z}$ *satisfying* $|\mathbf{z}-\theta_t| \le \epsilon$ *and in* $t \in (\epsilon,1]$, $\epsilon < \inf \mathcal{I}$, *choosing* h *to be a constant multiple of* $(n^{-1}\log n)^{1/(p+7)}$, *the mode estimates* $\hat{\theta}_t$ *of* $\hat{f}(\cdot,t)$ *satisfy*

$$\sup_{t\in\mathcal{I}} |\hat{\theta}_t - \theta_t| = O\big\{(n^{-1}\log n)^{2/(p+7)}\big\} \tag{6}$$

with probability 1.

4 Time-Dynamic Mode Tracking

4.1 Direct Implementation – Tracking Algorithm I

In the situation where the distribution which generates the data is evolving in time, the modal evolution θ_t corresponding to time-dynamic modes is of interest for tracking the shifts of the underlying distributions. The current mode location θ_t needs to be estimated from the data up to time t. We refer to this estimation problem as time-dynamic mode tracking. The mode provides information about current location of the mass of data, and tracking the evolving modes is of interest not least for visualizing the density evolution.

Implementing the concept in (4) and the corresponding time-dynamic density estimate (2) directly leads to the *time-dynamic mode tracking algorithm*-(I):

(Step 1). Initialize the starting evaluation time, $t_{i'}$, for some i', $1 \le i' \le n$.

(Step 2). Choose a temporal bandwidth $(\lambda_{i'})$ to determine the local time window, $\mathcal{T}(\lambda_{i'}) = \{\mathbf{X}_{ij} : \ t_{i'} - \lambda_{i'} \le t_i \le t_{i'}\}$.

(Step 3). Within $\mathcal{T}(\lambda_{i'})$, apply (4) and suitable spatial bandwidths to calculate $\hat{\theta}^{(i')}$.

(Step 4). Repeat (Step 2) and (Step 3) for $i = i' + 1, \ldots, n$.

(Step 5). Graphically connect $\hat{\theta}^{(i)}$ and $\hat{\theta}^{(i+1)}$, for example by straight lines, for $i = i', \ldots, n-1$.

In short, we move a local time window of varying size along the time axis from the starting evaluation time. Ignoring the issue of selecting the size of the moving window for the time being, the algorithm searches for the time-dynamic mode estimate within the local retrospective temporal window, by calculating the time-dynamic kernel density estimate using a spatial bandwidth, and then locating its maximizer as the mode estimate. Connecting the consecutive mode estimates in time, we obtain the estimated modal evolution path.

The performance of the time-dynamic mode trajectory naturally depends on the size of the temporal and spatial bandwidths. These can be chosen subjectively, but often an automatic selection method will be preferred (see section 4.3).

4.2 Implementation via Mean-Shift – Tracking Algorithm II

The mean shift algorithm has been implicitly based on first order local sample moments as a way to implement mode climbing algorithms such as the *Gradient Clustering Algorithm* (Fukunaga and Hostetler 1975), and the *Mean Shift/Mode Seeking Algorithm* (Cheng 1995; Comaniciu et al. 2002). This algorithm also underlies recent data sharpening proposals (Choi and Hall 1999a; Choi and Hall 1999b; Hall et al. 2001). The mean shift concept is related to gradient methods for the location of maxima and to local moments (Müller and Yan 2001). Our implementation here follows the local moment approach, which is extended here to accommodate the time-evolving feature.

For unit vectors e_j pointing in the j-th coordinate direction, we define the time-dynamic estimates

$$\hat{\mu}_{e_j}(\mathbf{z},t) = \frac{1}{h_j} \frac{\sum\limits_{1\le t_i\le t} \sum_{k=1}^{M_i} [\mathbf{z}^{(j)} - \mathbf{X}_{ik}^{(j)}]\, K_{\mathrm{S}}\big\{(\mathbf{z}-\mathbf{X}_{ik})\,\mathbf{H}^{-1/2}\big\}\, K_{\mathrm{T}}\{(t-t_k)/\lambda\}}{\sum\limits_{1\le t_i\le t} \sum_{k=1}^{M_i} K_{\mathrm{S}}\big\{(\mathbf{z}-\mathbf{X}_{ik})\,\mathbf{H}^{-1/2}\big\}\, K_{\mathrm{T}}\{(t-t_k)/\lambda\}}, \; 1\le j\le p, \tag{7}$$

which provide weighted local means taking the observations in a spatio-temporal window around a current starting point $\mathbf{z}$. The computation of the sample mean in a window around a given starting point provides the next iterate and is the core iteration step for the mean shift algorithm, which corresponds to an implementation of the steepest ascent method (Fukunaga and Hostetler 1975).

The time-dynamic mean shift estimate, evaluated at point $\mathbf{z}$ and time t, is then

$$\hat{m}(\mathbf{z},t) = (\hat{\mu}_{e_1}(\mathbf{z},t), \ldots, \hat{\mu}_{e_p}(\mathbf{z},t)). \tag{8}$$

Noting that the original Mean Shift Algorithm (see Cheng 1995) leads to $y_0 \to y_1 = y_0 + m(y_0) \to \ldots \to y_c$, we define the *Real-time Mean Shift Algorithm* as follows:

$$y_0(t) \to y_1(t) = y_0(t) + \hat{m}(y_0(t),t) \to \ldots \to y_c(t). \tag{9}$$

The *time-dynamic mode tracking algorithm*-(II) is then implemented as follows:

(Step 1). Initialize the starting evaluation time, $t_{i'}$ for some i', $1 \le i' \le n$.

(Step 2). Choose a temporal bandwidth to define the local time window, $\mathcal{T}(\lambda_{i'}) = \{\mathbf{X}_{ij} : t_{i'} - \lambda_{i'} \le t_i \le t_{i'}\}$.

(Step 3). Within $\mathcal{T}(\lambda_{i'})$, randomly choose a starting point, $y_c(t^{(0)})$ and, choosing a spatial bandwidth, run the Real-Time Mean Shift algorithm until it stops at the local

mode estimate, $y_c(t^{(1)})$, in $\mathcal{T}(\lambda_{i'})$.

(Step 4). Repeat (Step 2) and (Step 3), but use the previous local mode estimate evaluated at time t_{k-1}, i.e. $y_c(t^{(k-i')})$, as the starting point for the time-dynamic Mean Shift Algorithm to get $y_c(t^{(k-i'+1)})$ at time t_k, where $i' \leq k \leq n$.

(Step 5). The trajectory connecting all the $n - i' + 2$ time-dynamic mode estimates through the time frame, $y_c(t^{(0)}) \rightarrow \ldots \rightarrow y_c(t^{(n-i'+1)})$, then is the proposed estimate of the time-dynamic local modal evolution.

This algorithm generally gives very similar results to the direct implementation version (Tracking Algorithm I). However, it is considerably faster as current modes are the starting point for the next iterative gradient climbing, and since the mode for the next time considered will usually be close to the previous one. Moreover, an evaluation of the entire time-dynamic density with subsequent mode search, as required for Tracking Algorithm I, is costly and unnecessary if one focuses on the modes. Another appealing aspect of Tracking Algorithm II is the smoothness of estimated mode trajectories at low cost. Tracking Algorithm I can also be tuned to increase smoothness of estimated modal paths, if one evaluates at a denser grid of time points which however will substantially increase computational costs above the already much higher computational cost of Tracking Algorithm I as compared to Tracking Algorithm II.

To illustrate the costs, for sample data consisting of 1500 observations, each one sampled at a different time point, both algorithms were applied with the same bandwidth specifications and kernel functions. The mesh size for Algorithm I on which to evaluate the density was chosen as 30 by 30. We found that the computation time for Algorithm I was almost 30

times more than that of Algorithm II. The advantage of the direct implementation scheme of Algorithm I of being conceptually straightforward is clearly outweighed by added cost and reduced smoothness.

4.3 Automatic Bandwidth Selection

We discuss here bandwidth selectors for the case $p = 2$. The ideas can be easily extended to higher dimensional situations. The smoothing parameters involved in constructing the time-dynamic kernel density and mode estimate are the temporal bandwidth and the spatial bandwidth matrix which is assumed to be of full matrix form; the complexity of these tuning parameters renders subjective choices difficult and automatic choices desirable. The bandwidths for time-dynamic density estimation and mode tracking may vary in dependency with time, and automatic choices should reflect this.

The range of historical data to be considered is determined by the temporal bandwidth which defines the epoch prior to current time for which data are included. Given a temporal bandwidth, the spatial data to be considered are determined and one has at hand a problem closely related to two-dimensional density estimation. For such problems, various bandwidth matrix selectors have been proposed previously, including the normal reference method (Scott, 1992), multivariate cross validation (Sain et al, 1994), a multivariate rule of thumb (Silverman, 1986), or a multivariate plug-in method (Wand & Jones 1995; Duong and Hazelton 2003).

A natural proposal therefore is to couple the selection of the spatial bandwidth to temporal bandwidth choice; this idea is embodied in the proposed Time Dynamic Least Squares Cross Validation (TDLSCV) scheme, which provides "local" bandwidths for each time t. We

found it sufficient to choose a very simple reference method for the spatial smoothing matrix selection. Assuming a Gaussian kernel is used for the spatial smoothing step, then given a temporal bandwidth at time t, say $\lambda(t)$, the corresponding spatial bandwidth matrix is chosen as

$$\mathbf{H}_\lambda(t) = n^{-2/(4+p)}\, S_\lambda(t), \tag{10}$$

where $S_\lambda(t)$ is the sample covariance matrix of the data that fall in the time window $[t-\lambda(t), t]$.

Other established algorithms for bandwidth choice for spatial densities (such as described in Duong and Hazelton, 2003, see also http://www.maths.uwa.edu.au/∼duongt/research/index.html) were substantially slower and did not provide improvements over our algorithm. To use equation (10) for $\mathbf{H}_\lambda(t)$ for other (non-Gaussian) spatial kernels, one needs to multiply by suitable adjustment factors that depend on the kernel (Silverman 1986).

To address the remaining problem of temporal bandwidth selection, we adopt Least Squares Cross Validation Criterion (LSCV, see Silverman 1986). Let $\hat{f}^{(-ij)}(\mathbf{x}_{ij}, t; \lambda)$ be the version of $\hat{f}(\mathbf{x}_{ij}, t; \lambda)$ constructed when omitting x_{ij} from the sample, defined as

$$\hat{f}^{(-ij)}(\mathbf{x}_{ij}, t; \lambda) = \frac{1}{\big(N(t)-1\big)\lambda\,|\mathbf{H}_\lambda(t)|^{1/2}} \sum_{(k,l)\in A_{-(i,j)}} \sum K_{\mathrm{T}}\big((t-t_k)/\lambda\big) K_{\mathrm{S}}\big(\mathbf{H}_\lambda(t)^{-1/2}(\mathbf{x}_{ij}-\mathbf{X}_{kl})\big),$$

where $A_{-(i,j)} = \big\{(k,l)|0 \le t_k \le t \text{ and } (k,l) \ne (i,j)\big\}$, and for given t, define

$$LSCV(\lambda|t) \;=\; \int_0^t \int \hat{f}(\mathbf{x}, v; \lambda)^2 d\mathbf{x} dv - \frac{2}{N_\lambda(t)} \sum_{t-\lambda\le t_i\le t} \sum_{j=1}^{M_i} \hat{f}^{(-ij)}(\mathbf{x}_{ij}, t; \lambda), \tag{11}$$

where $N_\lambda(t) = \sum_{t-\lambda\le t_i\le t} M_i$ is the number of observations falling into the time window $[t-\lambda, t]$ .The time-dynamic temporal bandwidth at t is found by minimizing $LSCV(\lambda)$,

$$\lambda^*(t) \;=\underset{\lambda\in[0\,,1]}{\arg\min}\; LSCV(\lambda|t), \tag{12}$$

This minimizer is selected as the time-dynamic temporal bandwidth λ_t at time t, and the associated value of the bandwidth matrix $\mathbf{H}^*(t) = \mathbf{H}_{\lambda^*(t)}(t)$ is the chosen for the spatial smoothing step.

The *TDLSCV* criterion thus amounts to minimization of only a univariate data-dependent function, while providing simultaneous selection of temporal bandwidth and spatial bandwidth smoothing matrix, and accordingly is computationally fast. Alternatives would be plug-in methods or bootstrap methods (compare Silverman 1986, or Jones et al. 1996) but these methods are either very complicated numerically much more demanding. As we will demonstrate, the *TDLSCV* approach works very well in practice, produces smooth and reliable modal evolution paths, and is computationally fast, which is a particularly useful feature in possible online applications.

4.4 Large Sample Evaluation of Bandwidth Selection

For evaluating the behavior of the proposed bandwidth selector (TDLSCV), we compared its Integrated Mean Square Error (IMSE) with several fixed temporal bandwidth choices of $\lambda =$ 0.0067, 0.0167, 0.0334, 0.0667, 0.1334, and 0.2668 for 500 Monte Carlo runs. Each single data set was generated from bivariate normal distribution with means $\mu(t_i) = \big(3\cos(2\pi t_i), \sin(2\pi t_i)\big)$ and covariances

$$\Sigma(t_i) = \frac{1}{10}\begin{bmatrix} 1+t_i & 1/2 \\ 1/2 & (1+t_i)/2 \end{bmatrix},$$

changing over time, where the t_i, $i = 1, \ldots, 1500$, are 1500 equally spaced time points in [-0.25,1]. We used the Gaussian kernel function as the spatial kernel function, and for the runs with fixed temporal bandwidth we used equation (10) for the spatial bandwidth matrix

choice. The evaluation starts at the 450-th time point, t_{450}.

The observed values of IMSE and also Integrated Square Bias (ISB) and Integrated Variance (IVar) for each temporal bandwidth choice scheme are listed in Table 1.

Table 1: Evaluation of different bandwidth choices, comparing Time-Dynamic Least Squares Cross Validation (TDLSCV) with fixed temporal smoothing windows of $\lambda = 0.0067,\ 0.0167,\ 0.0334,\ 0.0667,\ 0.1334,$ and 0.2668, where the spatial bandwidth matrix is chosen according to (10).

Scheme ($\lambda =$)	TDLSCV	0.0067	0.0167	0.0334	0.0667	0.1334	0.2668
ISB	0.0658	0.0143	0.0314	0.0654	0.1329	0.2661	0.5279
IVar	0.0753	0.1481	0.0974	0.0747	0.0571	0.0435	0.0415
IMSE	0.1411	0.1624	0.1287	0.1401	0.1901	0.3096	0.5694

We find that the proposed automatic bandwidth selector works quite satisfactorily even though it does not attain the lowest IMSE. Furthermore, the averages over all simulation runs of estimated modal paths corresponding to different temporal bandwidth choices fell perfectly on the track of the true modal function except lagging behind the current mode by different amounts of time lag. The time lag between the true and estimated mode paths increases as the temporal bandwidth increases. Our overall conclusion is that the proposed bandwidth selector *TDLSCV* provides a practical, appealing and reliable temporal bandwidth choice.

We illustrate the time-dynamic density estimation in this simulation setting, using *TDLSCV* and selecting one run. The aim is to estimate the entire time-dynamic density. Estimated and true densities are depicted as contour plots in Figure 1 for four time points, at $t_{450},\ t_{800},\ t_{1150},\ t_{1500}$. The proposed time-dynamic density estimate is seen to capture the

true density quite well in terms of its geometric structure and mode location, while indicating some flattening of the estimated density as is typical for smoothing procedures.

5 Applications

5.1 Simulation Studies

Since it is the case of highest practical importance, we only consider the bivariate situation with $p = 2$ in our applications. We generated data from different bivariate normal distributions $\mathrm{BVN}\big(\mu(t_i), \Sigma(t_i)\big)$ by using different mean and covariance functions as detailed below. At each t_i, $i = 1, \ldots, 1500$, we generated one observation according to the current distribution $F(\cdot, t_i)$, thus choosing $M_i = 1$. Note that both the mean vector and covariance matrix are functions of time.

We track the modal evolution path by the estimate produced by Tracking Algorithm II, as described in section 4.2. We used a bivariate Gaussian kernel function, $K_{\mathrm{S}}(\mathbf{x}) = (2\pi)^{-1}\exp(-\frac{1}{2}\mathbf{x}\mathbf{x}')$, as spatial kernel function and the kernel $K_{\mathrm{T}}(t) = 4 - 6t$ as temporal kernel function. This is a well-known simple boundary kernel (e.g., Müller 1991) which has a minimum variance property. For a large class of boundary kernels that can be used alternatively compare for example Müller and Wang (1994). For temporal and time-dynamic spatial bandwidths choices, we implemented TDLSCV (10)-(12), providing bandwidths indexed by current time t. The time-dynamic mode tracking results are shown after the 450-th time point, t_{450}. We repeated the estimation for each of 500 runs, and show 10 randomly chosen estimated modal paths, overlapped with the corresponding data, in Figure 2 and 3.

For these simulations, the covariance matrix as a function of time t was chosen as

$$\Sigma(t) = \frac{1}{10} \begin{bmatrix} 1+t & 1/2 \\ 1/2 & 1+t \end{bmatrix},$$

except for the case of the ellipse, where the right lower element was $(1+t)/2$. The mean functions were $\mu(t) = (10t, 10t)$ for the straight line, $\mu(t) = (10t, (10t-2.5)^2)$ for the parabola, $\mu(t) = (10t, \sin(10t))$ for the sine curve, and $\mu(t) = (3\cos(2\pi t), \sin(2\pi t))$ for the ellipsoid. For all cases, the time points t_i were chosen as 1500 equally spaced points in [0, 1].

In each panel of Figure 2 and 3, the solid black line denotes the true modal evolution path, the colored lines represent 10 randomly chosen estimated modal paths, and the gray dots in the background are the data corresponding to the chosen Monte Carlo runs. The color bar at the side shows the coloring scheme used in the plots, which serves to indicate the progression of time. Overall, the estimated modal evolution paths are seen to track the geometric features of the underlying true paths quite well even though there is a minor time lag.

5.2 Tracking of Fox Rabies Epidemic

Data on the occurrence of fox rabies (http://lib.stat.cmu.edu/datasets/Andrews/T50.2) were analyzed by Sayers et al. (1977) using pattern analysis. The original wildlife rabies epidemic study was performed over a 133×133 km^2 region from January 1963 to March 1971 in southern Germany. The monthly occurrence frequencies (there are 96 months included in the data we analyzed) were reported from 1024 uniformly distributed stations. Details about data collection can be found in Moegle et al. (1974).

Here the time-evolution is of major interest, and the data fit our model with random number of observations at each time point M_i, where $i = 1, \ldots, 96$. The modal evolution path can be thought of as representing direction and speed of the geographic movement of the epidemic. The time-dynamic feature will be important in current-time epidemiological applications. The modal evolution path provides a useful tool for visualizing the spatio-temporal evolution of incidence of the epidemic and would aid in the optimal deployment of resources to combat the its spread.

Since the disease reporting stations are equally spaced, we re-coded the position of each station by assigning an integer value from 1 to 32 for both longitude and altitude, reflecting the relative position on the lattice into which the study region has been mapped. Then we applied the proposed time-dynamic mode tracking algorithm starting from t_{30}, implementing Tracking Algorithm II with TDLSCV and a randomly chosen starting point. Constructing the modal evolution path for this application took 26 seconds on a PC running WIN/XP with 3.4G Hz CPU and 1G RAM.

From the estimated modal evolution path, displayed in Figure 4, the spatio-temporal course of the evolution of the epidemic is clearly discernible, with an initial movement from north-west to south-east, followed by a brief period of movement towards the south-west. While the north-west to south-east movement has been described in previous analyses of these data, the change in direction towards the south-west appears to be a new finding enabled by dynamic mode-tracking. In time-dynamic applications the change in the direction would provide valuable information for disease control. In addition to the direction of modal evolution, Figure 5 also provides information about the speed of the spread of the disease

through the color-coding of the path. We find that the speed of disease spread is fairly homogeneous over time.

5.3 Tracking of Bivariate Stock Market Indices

Daily closing prices of two technology stock indices, the AMEX biotech index (BTK) and the XPH semiconductor index (SOXX), recorded for the period from August 2 1999 to July 30 2000, with a total of 995 daily records, provide a second example where useful information can be obtained with dynamic mode tracking. The data feature one pair of closing prices for each business day during the evaluation period, i.e. $M_i = 1$, $i = 1, \ldots, 995$. Although this type of data is usually viewed as a series of stock market indices that have been studied mostly in a time series framework, the proposed dynamic mode tracking serves as a nonparametric alternative and at the least it will provide a graphical tool for visualization of the bivariate movements of the stock indices. The time-dynamic feature is clearly of interest for online tracking of current movements in financial markets.

In estimation, we applied the proposed Mode Tracking Algorithm II with TDLSCV and starting from t_{100}. The result is the estimated modal evolution path in Figure 5. This path reveals interesting patterns of the time-dynamic stock-index evolution. Both indices are on an increasing trend in the beginning, then index SOXX starts to decrease while index BTK stays at a roughly constant level, followed by a decline in both indices. At the end of the time period, the index BTK starts to rebound. There are also some loops. For this application it took only 10 seconds to run the whole course of application on a PC running WIN/XP with 3.4G Hz CPU and 1G RAM.

As before, speed information is provided through the color-coding of the modal paths. We find substantially faster stock market movements throughout the first half of the observed time period and slower movements during the second half. During the evaluation period, biotechnology stocks as reflected in index BTK clearly fared better than semiconductor stocks reflected in index SOXX. The time-dynamic modal evolution path provides insight and visualization beyond what could be achieved with classical methods of time series analysis.

Animated pictures for these examples and also the simulations are available at http://anson.ucdavis.edu/∼mueller/modetrack/modetrack.html. They provide additional insight into the working of the algorithm and a sense of how the proposed procedures would perform in online situations.

Appendix: Proofs

First we give a proof of Theorem 1. Let var′ denote variance conditional on $\mathcal{M}$, write $C, C_1, C_2, \ldots$ for a sequence of positive constants not depending on $\mathbf{z}$ or on t, and note that

$$\begin{aligned} Q = Q(\mathbf{z},t) &\equiv \hat{f}(\mathbf{z},t) - E\{\hat{f}(\mathbf{z},t) \mid \mathcal{M}\} \\ &= \frac{1}{N(t)h^{p+1}} \sum_{1\le t_i\le t} K_{\mathrm{T}}\Big(\frac{t-t_i}{h}\Big) \sum_{j=1}^{M_i} \Big[K_{\mathrm{S}}\Big(\frac{\mathbf{z}-\mathbf{X}_{ij}}{h}\Big) - E\Big\{K_{\mathrm{S}}\Big(\frac{\mathbf{z}-\mathbf{X}_{ij}}{h}\Big)\Big\}\Big] \end{aligned}$$

and $\mathrm{var}'(Q) \le C_1/N(t)h^{p+1}$, uniformly in $(\mathbf{z},t) \in \mathbb{R}^p \times \mathcal{I}$. Conditional on $\mathcal{M}$, Q equals a sum of $N(t)$ independent random variables with zero mean and with absolute values bounded uniformly by $C_2/N(t)h^{p+1}$, for all realizations and all $(\mathbf{z},t) \in \mathbb{R}^p \times \mathcal{I}$. Using these results and Bennett's inequality (see e.g. Pollard 1984, pp. 193) we deduce that with $\ell_n = \log n$ and

$\xi = C\,(\ell_n/N(t)h^{p+1})^{1/2}$,

$$P(|Q| > \xi \mid \mathcal{M}) \le 2\,\exp\bigl\{ - C_3\,C^2\,\ell_n/(1+\xi)\bigr\}\,,$$

where C_3 depends only on C_1 and C_2. Therefore, if $C_4 > 0$ is given, and we choose $C^2 > 3C_4/C_3$, then

$$\sup_{(\mathbf{z},t)\in\mathbb{R}^p\times\mathcal{I}} P\Bigl\{|Q(\mathbf{z},t)| > C\,\bigl(\ell_n/N(t)h^{p+1}\bigr)^{1/2}\Bigr\} = O\bigl(n^{-3C_4}\bigr)\,.$$

(Note that, since we assumed all moments of the distribution of M_i are finite, then $P(|N(t) - ntm| > n^{(1/2)+\epsilon}) = O(n^{-B})$ for each $\epsilon, B > 0$.) Hence, if the sets $\mathcal{K}_n \subset \mathcal{K}$ and $\mathcal{I}_n \subset \mathcal{I}$ each have at most $O(n^{C_4})$ elements, then

$$P\Biggl\{\sup_{(\mathbf{z},t)\in\mathcal{K}_n\times\mathcal{I}_n} |Q(\mathbf{z},t)| > C\,\bigl(\ell_n/N(t)h^{p+1}\bigr)^{1/2}\Biggr\} = O\bigl(n^{-C_4}\bigr)\,.$$

This result, and the Hölder continuity of K (or of its derivative, when we are proving the analogous result for $\mathbf{D}_{e_j}f$ rather than f) allow us to show that if C_4 (and hence also, C) is chosen sufficiently large,

$$P\Biggl\{\sup_{(\mathbf{z},t)\in\mathcal{K}\times\mathcal{I}} |Q(\mathbf{z},t)| > C\,\bigl(\ell_n/nh^{p+1}\bigr)^{1/2}\Biggr\} = O\bigl(n^{-C_4}\bigr) \tag{13}$$

Moreover, it is easy to show that $E\{\hat{f}(\mathbf{z},t) \mid \mathcal{M}\} = f(\mathbf{z},t) + O(h^2)$ with probability 1. Theorem 1 follows from this result and (13).

Next we outline the derivation of Theorem 2. An argument similar to that above gives (5) in that theorem. Let $\delta = (nh^{p+3})^{-1/2}\,(\log n)^{1/2} + h^2$ denote the order of magnitude of the remainder term on the right-hand side of (5). Now, θ_t and $\hat{\theta}_t$ are, respectively, solutions of the simultaneous equations $\mathbf{D}_{e_j}f(\mathbf{z},t) = 0$, for $1 \le j \le p$, and $\mathbf{D}_{e_j}\hat{f}(\mathbf{z},t) = 0$, $1 \le j \le p$. It follows from this property, using (5) and Taylor expansion, that with probability 1,

$$0 = \mathbf{D}_{e_j}\hat{f}(\hat{\theta}_t,t) = \mathbf{D}_{e_j}f(\hat{\theta}_t,t) + O(\delta) = M_t\,(\hat{\theta}_t - \theta_t) + O\left(\|\hat{\theta}_t - \theta_t\|^2 + \delta\right)\,, \tag{14}$$

where the $p \times p$ matrix M_t denotes the Hessian of $f(\mathbf{z}, t)$, evaluated at $\mathbf{z} = \theta_t$. Since, by assumption, M_t is nonsingular, then (14) entails

$$\hat{\theta}_t - \theta_t = O(\delta) \tag{15}$$

with probability 1. By assumption in Theorem 2, h is a constant multiple of $(n^{-1} \log n)^{1/(p+7)}$, and therefore δ equals a constant multiple of $(n^{-1} \log n)^{2/(p+7)}$. Result (6) follows from this property and (15).

REFERENCES

Cheng, Y. (1995), "Mean Shift, Mode Seeking, and Clustering," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **17**, 790-799.

Comaniciu, D., and Meer, P. (2002), "Mean Shift: A Robust Approach Toward Feature Space Analysis." *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **24**, 1-18.

Duong, T., and Hazelton, M. (2003), "Plug-in Bandwidth Matrices for Bivariate Kernel Density Estimation," *Journal of Nonparametric Statistics*, **15**, 17-30.

Eddy, W. F. (1980), "Optimal Kernel Estimators of the Mode," *The Annals of Statistics*, **8**, 870-882.

Elgammal, A., Duraiswami, R., and Davis, L. S. (2003), "Efficient Kernel Density Estimation using the Fast Gauss Transform with Applications to Color Modeling and Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**, 1499-1504.

Epanechnikov, V.A. (1969), "Nonparametric Estimation of a Multidimensional Probability Density," *Theory of Probability and its Applications*, **14**, 153-158.

Fukunaga, K., and Hostetler, L. D. (1975), "The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition," *IEEE Transactions on Information Theory*, **21**, 32-40.

Gasser, T., Müller, H. G., and Mammitzsch, V. (1985), "Kernels for Nonparametric Curve Estimation," *Journal of the Royal Statistical Society Ser. B*, **47**, 238-256.

Hall, P., and Patil, P. (1994), "On the Efficiency of On-line Density Estimators," *IEEE Transactions on Information Theory*, **40**, 1504-1512.

Hall, P., and Choi, E. (1999a), "Nonparametric Approach to Analysis of Space-Time Data on Earthquake Occurrences," *Journal of Computational and Graphical Statistics*, **8**, 733-748.

—— (1999b). "Data Sharpening as a Prelude to Density Estimation," *Biometrika*, **86**, 941-947.

Hall, P., Poskitt, D. S., and Presnell, B. (2001), "A Functional Data-analytic Approach to Signal Discrimination," *Technometrics*, **43**, 941-947.

Hossjer, O., and Holst, U. (1995), "Online Density Estimators with High-Efficiency," *IEEE Transactions on Information Theory*, **41**, 829-833.

Jones, M. C. (1993), "Simple Boundary Correction for Kernel Density-estimation," *Statistics and Computing*, **3**, 135-146.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996), "A Brief Survey of Bandwidth Selection for Density Estimation," *Journal of the American Statistical Association*, **91**, 401-407.

Konakov, V. D. (1973), "On the Asymptotic Normality of the Mode of Multidimensional Distributions," *Theory of Probability and its Applications*, **18**, 794-799.

Müller, H. G. (1991), "Smooth Optimum Kernel Estimators near Endpoints," *Biometrika*, **78**, 521-530.

Müller, H.G., and Wang, J.L. (1994), "Hazard Rate Estimation Under Random Censoring With Varying Kernels and Bandwidths," *Biometrics* **50**, 61-76.

Müller, H. G., and Yan, X. (2001), "On Local Moments," *Journal of Multivariate Analysis*, **76**, 90-109.

Moegle, H., Knorpp, F., Bögel, K., Arata, A., Dietz, P., and Diethelm, P. (1974), "Zur Epidemiologie der Wildtiertollwut, Untersuchungen in südlichen Teil der Bundesrepublik Deutschland," *Zentralblatt Veterinärmedizin B*, **21**, 647.

Parzen, E. (1962), "On Estimation of a Probability Density Function and Mode," *Annals of Mathematical Statistics*, **33**, 1065-1076.

Pollard, D. (1984), *Convergence of Stochastic Processes*, New York: Springer.

Sayers, B., Men Mansourian, B. G., Phan Tan, T., and Bögel, K. (1977), "A Pattern-analysis Study of a Wild-life Rabies Epizootic," *Medical Informatics*, **2**, 11-34.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: Wiley.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman & Hall.

Wand, M. P., and Jones, M. C. (1993), "Comparison of smoothing parameterizations in

bivariate kernel density estimation," *Journal of the American Statistical Association*, **88**, 520-528.

Wand, M. P., and Jones, M. C. (1995), *Kernel Smoothing*, London: Chapman & Hall.

Wegman, E. J., and Davies, H. I. (1979), "Remarks on Some Recursive Estimators of a Probability Density," *The Annals of Statistics*, **7**, 316-327.

Wegman, E. J., and Marchette, D. J. (2003), "On Some Techniques for Streaming Data: A Case Study of Internet Packet Headers," *Journal of Computational and Graphical Statistics*, **12**, 893-914.

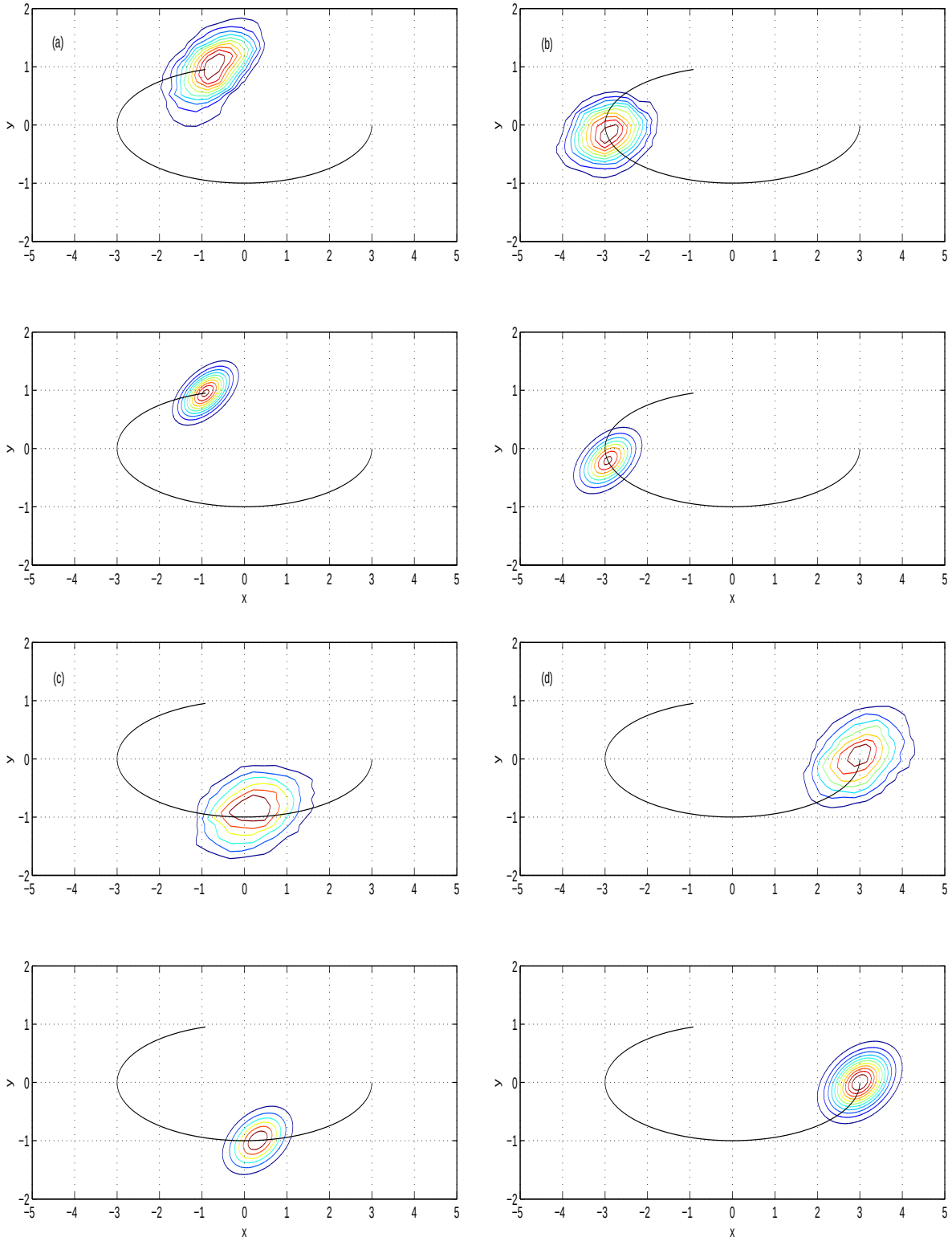


Figure 1: Estimated (above) and true (below) contour plots for Time-dynamic Density Estimation at $(a)\ t_{450},\ (b)\ t_{800},\ (c)\ t_{1150},\ (d)\ t_{1500}$, obtained with Tracking Algorithm I and the TDLSCV bandwidth selector.

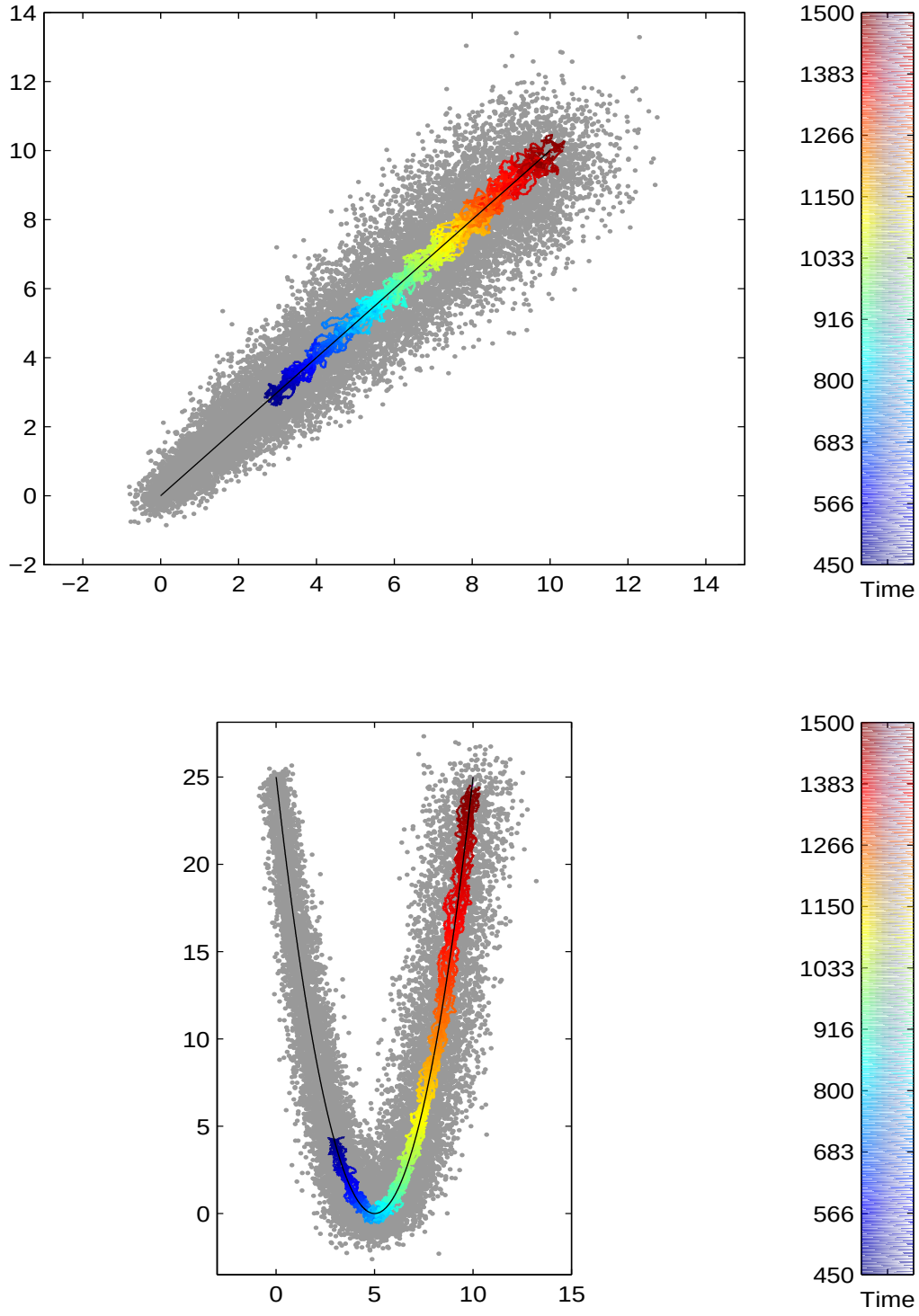


Figure 2: Applications of Tracking Algorithm II with TDLSCV to simulated bivariate normal data sets, each of size 1500 for the case of linear (above) and quadratic (below) modal evolution paths. Ten randomly chosen estimated modal paths out of 500 time-dynamic mode paths are plotted along with the corresponding data. The solid black line indicates the true modal evolution path. The color of the path indicates the progression of time.

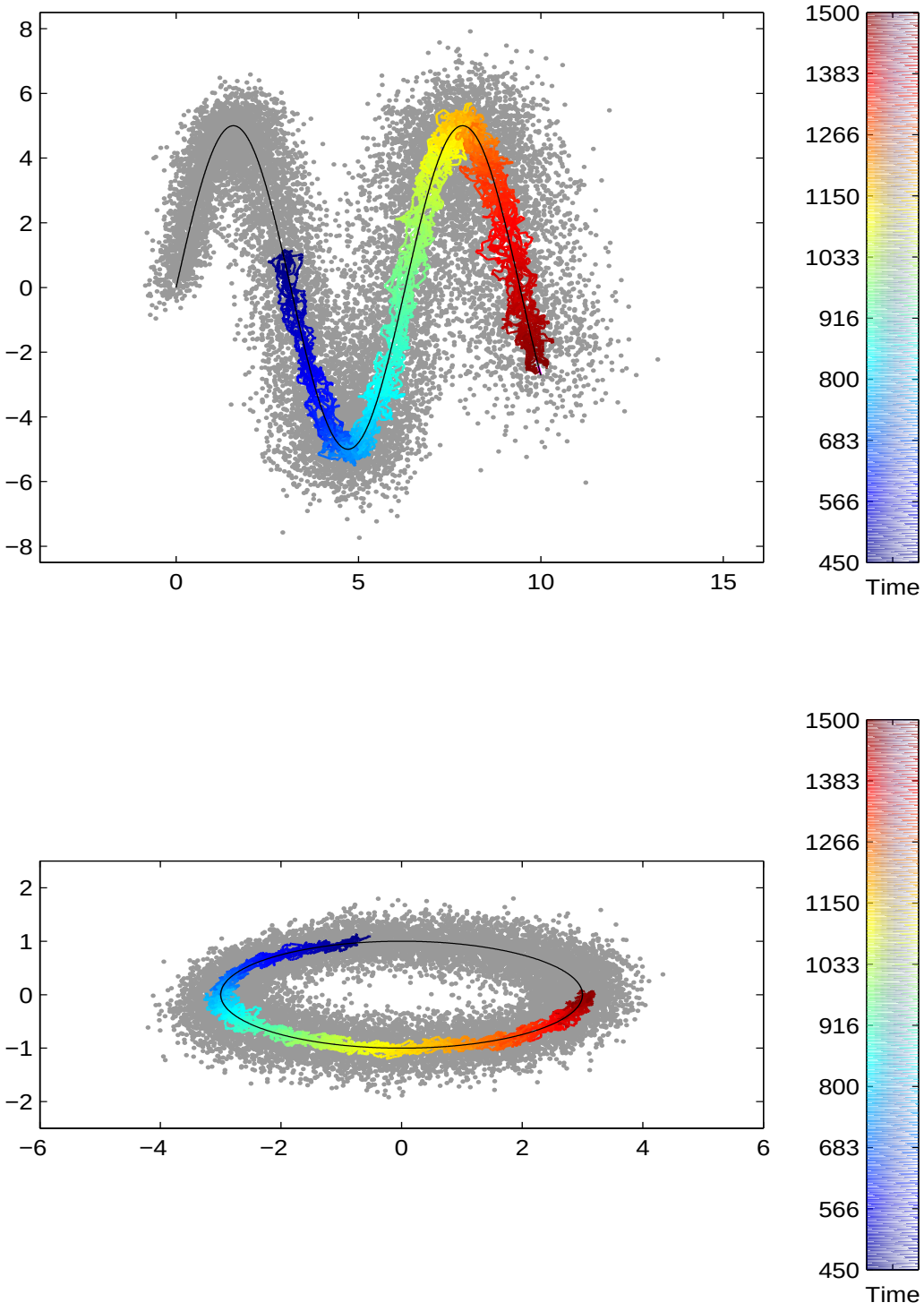


Figure 3: Same as Figure 2, for sinusoidal (above) and ellipsoidal (below) model paths.

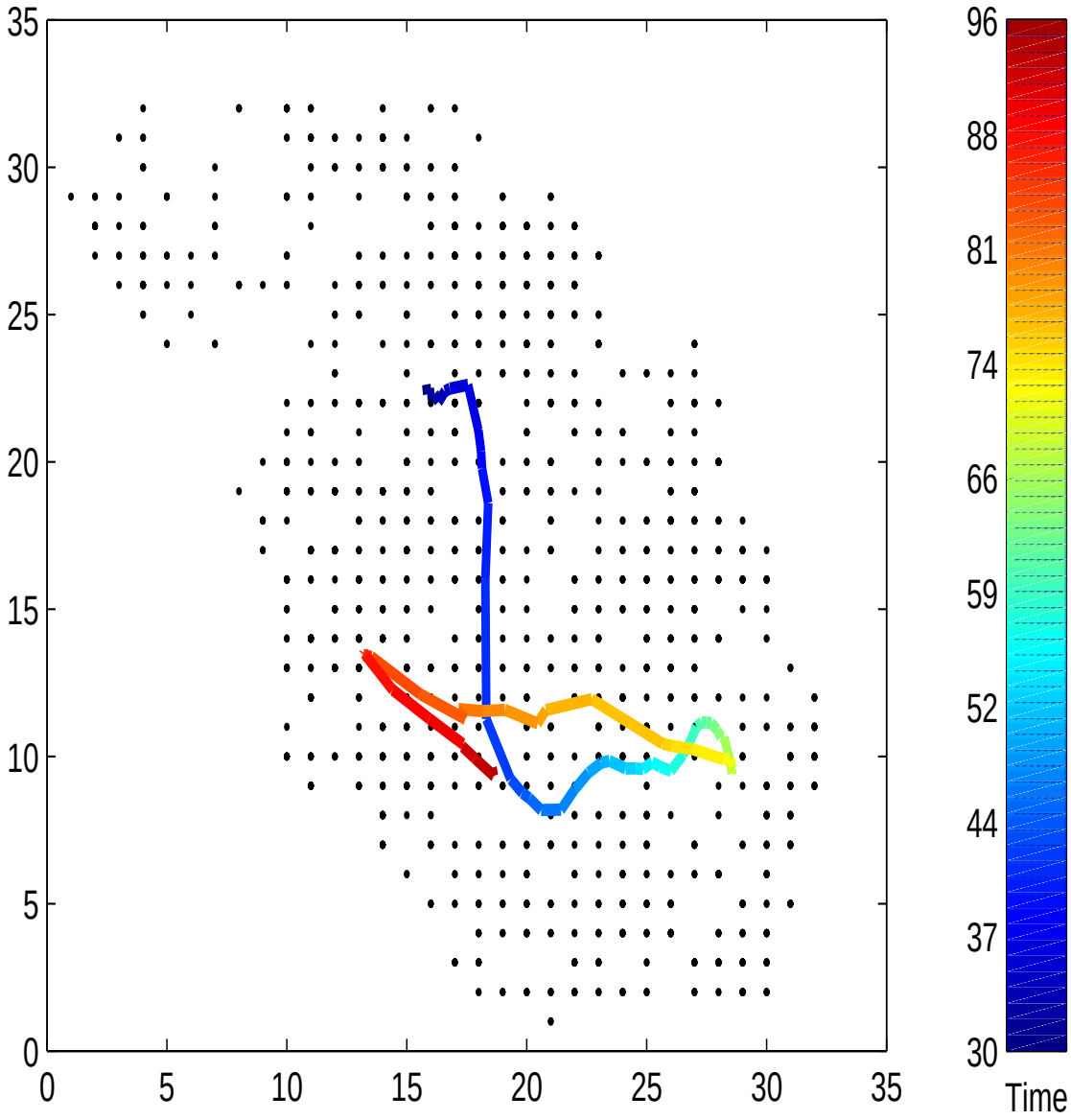


Figure 4: Modal evolution path for the Fox rabies epidemic. Observations are monthly occurrences of fox rabies recorded by 1024 equidistant from January 1963 to March 1971 in southern Germany, where for each month the number of new cases (M_i, $i = 1, \ldots, 96$) ranges from 1 to 72. The position of each station is coded from 1 to 32 for both longitude and latitude, reflecting the relative position on the lattice onto which the study region has been mapped, and is indicated by a black dot. The estimated modal evolution path denoted by the solid line is obtained by Tracking Algorithm II, implemented with TDLSCV. The time progression of the epidemic is color-coded.

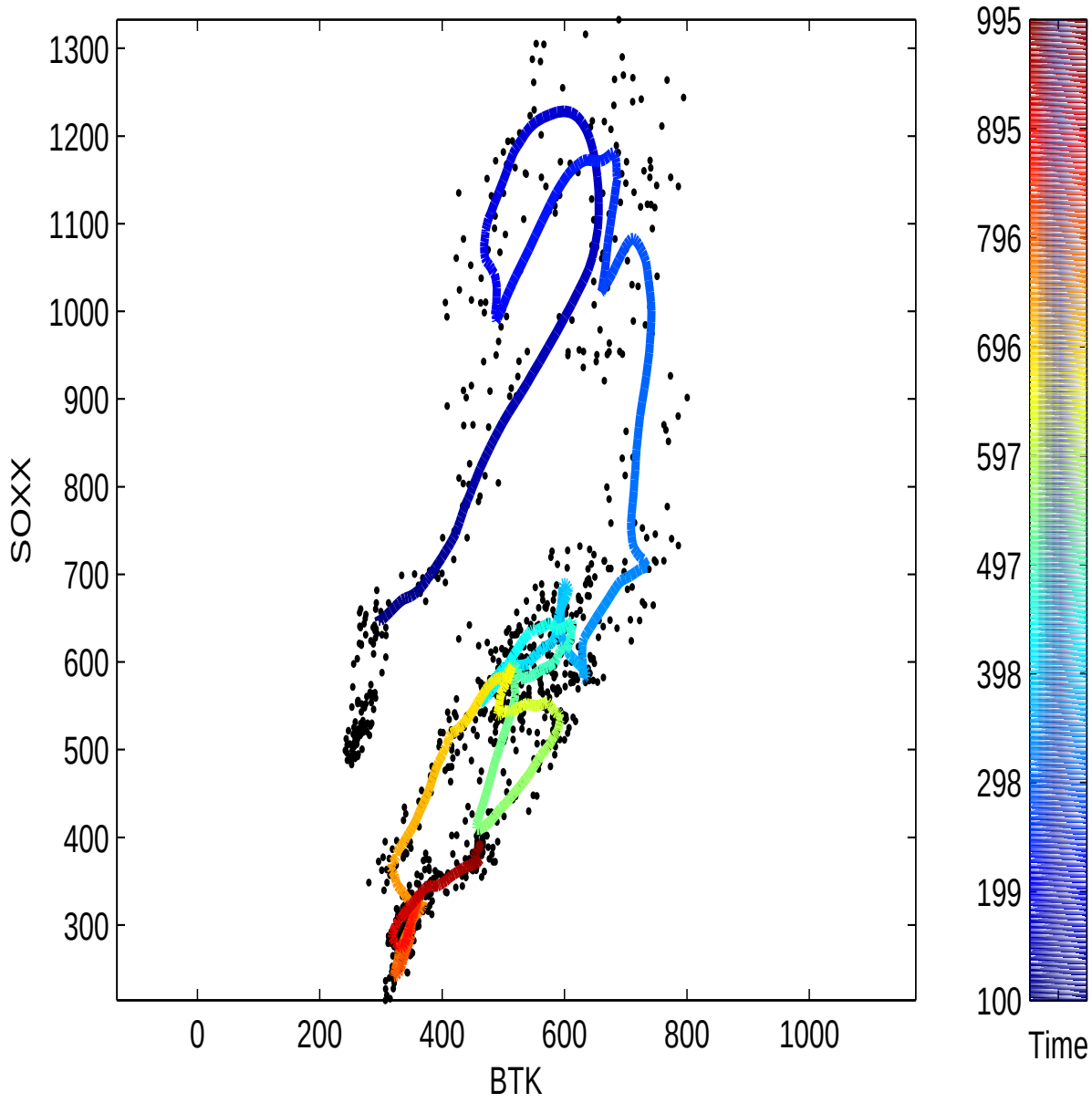


Figure 5: Modal evolution path for bivariate stock market indices. The data are the daily closing prices of two technology stock indices, the AMEX biotech index (BTK) and the XPH semiconductor index (SOXX), recorded for the period from August 2 1999 to July 30 2000, with a total of 995 daily records. The estimated modal evolution path indicated by the solid line is estimated by Tracking Algorithm II, implemented with TDLSCV. The time progression of the path is color-coded and clearly shows a deceleration towards the end of the time period with less movement in stock prices.