COMMUN. MATH. SCI.
Vol. 8, No. 1, pp. 217–233

DENSITY ESTIMATION BY DUAL ASCENT OF THE LOG-LIKELIHOOD*

ESTEBAN G. TABAK† AND ERIC VANDEN-EIJNDEN‡

Dedicated to the sixtieth birthday of Professor Andrew Majda

Abstract. A methodology is developed to assign, from an observed sample, a joint-probability distribution to a set of continuous variables. The algorithm proposed performs this assignment by mapping the original variables onto a jointly-Gaussian set. The map is built iteratively, ascending the log-likelihood of the observations, through a series of steps that move the marginal distributions along a random set of orthogonal directions towards normality.

Key words. Density estimation, machine learning, maximum likelihood.

AMS subject classifications. 34A50, 65C30, 65L20, 60H35.

1. Introduction and problem setting

Extracting information from data is a fundamental problem underlying many applications. Medical doctors seek to diagnose a patient's health from clinical data, blood tests and genetic information. Pharmaceutical companies analyze the results of massive *in vitro* tests of different compounds to select the best candidate for new drug development. Insurance companies assess, based on financial data, the probability that a number of credit-lines go into default within the same time-window. Using commercial data, market analysts attempt to quantify the effect that advertising campaigns have on sales. Weather forecasters extract from present and past observations the likely state of the weather in the near future. Climate scientists estimate long-time trends from observations over the years of quantities such as sea-surface temperature and the atmospheric concentration of CO_2.

In many of these applications, the fundamental "data problem" consists of estimating, from a sample of a set of interdependent variables, their joint probability distribution. Thus, the financial analyst dealing in credit derivatives seeks the probability of joint default of many debts over a specified time window; the medical doctor, the likelihood that a patient's test results are associated with a certain disease; the weather forecaster, the likelihood that the pattern of today's measurements anticipate tomorrow's rain.

For continuous variables, the density estimation problem can be posed as follows: Given a sample of m independent observations x^j of n variables x_i, one seeks a robust estimate of their underlying probability density, $\rho(x)$. This problem has been addressed in numerous ways. In a parametric approach, one considers a family of probability densities depending on a set of parameters, and maximizes the likelihood of the observations in the allowed parameter range; Gaussian mixtures [3] and smoothing splines [9] are popular choices. Another procedure for density estimation, widely used in the financial world, is the Gaussian copula, in which the set of the marginal densities of the individual variables are estimated and then combined into a joint Gaussian distribution [4]. Yet another approach is the so-called projection

*Received: September 22, 2008; accepted (in revised version): January 6, 2009.

†Courant Institute, New York University, 251 Mercer street, New York, NY 10012 (tabak@cims.nyu.edu).

‡Courant Institute, New York University, 251 Mercer street, New York, NY 10012(eve2@cims.nyu.edu).

pursuit [5], which seeks optimal directions for functional fitting. Within this latter framework, the Gaussianization procedure proposed in [6] has some commonality with the methodology developed here.

We propose to perform density estimation by mapping the x's into a new set of variables y with known probability density $\mu(y)$. Then the density $\rho(x)$ is given by

$$\rho(x) = J_y(x)\,\mu(y(x)), \tag{1.1}$$

where $J_y(x)$, the Jacobian of the map $y(x)$, is computed explicitly alongside the map. The map $y(x)$ is built as an infinite composition of infinitesimal transformations, i.e., by introducing a flow $z = \phi_t(x)$ such that

$$\phi_0(x) = x, \qquad \lim_{t\to\infty} \phi_t(x) = y(x). \tag{1.2}$$

Associated with the map $\phi_t(x)$ we can introduce the density $\tilde{\rho}_t(x)_t$ given by (1.1) but with $y(x)$ replaced by $\phi_t(x)$:

$$\tilde{\rho}_t(x) = J_{\phi_t}(x)\,\mu(\phi_t(x)). \tag{1.3}$$

If (1.2) holds, then from (1.1) the density $\tilde{\rho}_t(x)$ satisfies

$$\tilde{\rho}_0(x) = \mu(x) \qquad \lim_{t\to\infty} \tilde{\rho}_t(x) = \rho(x)\,.$$

Given a sample x^j, $j = 1,\ldots,m$, a measure of the quality of the estimated density $\tilde{\rho}_t(x)$ is the log-likelihood of the sample with respect to this density,

$$L[\phi_t] = \frac{1}{m}\sum_{j=1}^{m} \log \tilde{\rho}_t(x^j) = \frac{1}{m}\sum_{j=1}^{m} \big(\log(J_{\phi_t}(x^j)) + \log(\mu(\phi_t(x^j)))\big)\,. \tag{1.4}$$

This suggests constructing the flow ϕ_t by following a direction of ascent of $L[\phi_t]$, so that the log-likelihood is always increasing,

$$\frac{d}{dt} L[\phi_t] \geq 0, \tag{1.5}$$

and that the map $y(x) = \lim_{t\to\infty} \phi_t(x)$ is a (local) maximizer of the log-likelihood function of the sample with respect to $\rho(x) = \tilde{\rho}_\infty(x)$:

$$L[y] = \frac{1}{m}\sum_{j=1}^{m} \log \tilde{\rho}_\infty(x^j) = \frac{1}{m}\sum_{j=1}^{m} \big(\log(J_y(x^j)) + \log(\mu(y(x^j)))\big)\,. \tag{1.6}$$

The methodology proposed here builds on the realization that such a direction of ascent can be determined locally in time, based on the current values of $z^j = \phi_t(x^j)$, i.e., without reference to the original sample. The original values of x^j can be thought of as Lagrangian markers for a flow that carry the particles $z^j = \phi_t(x^j)$ toward a state with probability density μ.

It will emerge in the discussion below that the most natural choice for the target distribution μ is an isotropic Gaussian. This choice allows one to build the map ϕ_t from the composition of single variable transformations, which are much easier to determine.

The remainder of this paper is structured as follows. In section 2, we consider the ideal situation where the sample consists of infinitely many observations. In this case

the procedure gives rise to a nonlinear diffusive equation for the probability density ρ_t of the particles $z=\phi_t(x)$. Section 3 shows numerical examples of the solution to this partial differential equation, which displays fast convergence and robust "probability fronts." We prove in section 4 that the procedure makes ρ_t always converge to the target μ, and hence the estimate $\tilde{\rho}_t(x)$ converges to the actual density $\rho(x)$. Section 5 shows how the procedure can be reduced, still in the case with infinitely many observations, to the one-dimensional descent of each marginal density toward the corresponding marginal of μ.

Section 6 translates all these results into a procedure for density estimation from samples of finite size. This involves the following new ingredients:

- Random rotations, which allows one to consider the marginals along all directions in rapid succession.
- The introduction of a family of maps depending on only a handful of parameters, to act as building blocks for the flow ϕ_t. These maps have carefully controlled length-scales, so as not to over-resolve the density and not turn it into a set of delta-functions concentrated on the observational set.
- A straightforward procedure to discretize the time associated with the particle flow.

Section 7 presents one- and two-dimensional examples of applications of the algorithm to synthetic data. Real data scenarios, typically in much higher dimensions, will be discussed elsewhere, in the context of specific applications, e.g., to medical diagnosis from genetic and clinical data.

2. The continuous case

As the number of observations m tends to infinity, the log-likelihood function (1.4) becomes

$$L_\rho[\phi_t]=\int (\log(J_{\phi_t}(x))+\log(\mu(\phi_t(x))))\,\rho(x)\,dx. \tag{2.1}$$

Its first variation with respect to ϕ_t can be computed exactly and is given by

$$\frac{\delta L_\rho}{\delta \phi_t}=J_{\phi_t}(x)\left(\frac{\nabla_z \mu(z)}{\mu(z)}\rho_t(z)-\nabla_z \rho_t(z)\right), \tag{2.2}$$

where $z=\phi_t(x)$ and

$$\rho_t(z)=\frac{\rho(x)}{J_{\phi_t}(x)}. \tag{2.3}$$

This function (not to be confused with $\tilde{\rho}_t(x)$ in (1.3)) is the probability density of the variable $z=\phi_t(x)$ given that x is distributed according to $\rho(x)$.

In order to increase the log-likelihood, we evolve $\phi_t(x)$ according to

$$\dot{\phi}_t(x)=u_t(\phi_t(x)) \tag{2.4}$$

where

$$u_t(z)=\frac{\nabla_z \mu(z)}{\mu(z)}\rho_t(z)-\nabla_z \rho_t(z). \tag{2.5}$$

From (2.2), the velocity $u_t(z)$ is simply the gradient of the log-likelihood function divided by the (positive) Jacobian $J_{\phi_t}(x)$. This guarantees that the evolution (2.4)

follows a direction of ascent (though not steepest ascent) of the log-likelihood function and hence increases the value of this function. To understand what dropping the factor $J_{\phi_t}(x)$ amounts to, note that

$$\frac{\delta L_\rho[\varphi\circ\phi_t]}{\delta\varphi}\Big|_{\varphi=id} = \frac{\nabla_z\mu(z)}{\mu(z)}\rho_t(z) - \nabla_z\rho_t(z), \tag{2.6}$$

where $z=\phi_t(x)$. Thus, (2.4) corresponds to the evolution by steepest ascent on a modified log-likelihood function in which, at time t, one uses $z=\phi_t(x)$ as the current sample rather than the original x.

It is also useful to write the dual of (2.4) by looking at the evolution of the density $\rho_t(z)$. This function satisfies the Liouville equation

$$\frac{\partial\rho_t}{\partial t} + \nabla\cdot\left(\rho_t\, u_t\right) = 0, \tag{2.7}$$

or, explicitly using (2.5),

$$\frac{\partial\rho_t}{\partial t} = \nabla\cdot\left(\left(\nabla\rho_t - \frac{\nabla\mu}{\mu}\rho_t\right)\rho_t\right), \tag{2.8}$$

Thus, as the particles flow from x to $y(x)$ via $\phi_t(x)$, their probability density ρ_t evolves from the (unknown) initial ρ toward the target μ. At the same time, the current estimate for the density of the markers, $\tilde{\rho}_t$, evolves from μ towards ρ. This is what we refer to as dual ascent.

Finally, note that the Liouville equation 2.8 can be re-written in the form

$$\frac{\partial\rho_t}{\partial t} = \nabla\cdot\left(\mu^2\nabla\left(\frac{1}{2}\left(\frac{\rho_t}{\mu}\right)^2\right)\right). \tag{2.9}$$

This is a nonlinear diffusion equation frequently used to model flows in porous media. The form (2.9) clearly has the desired target $\rho_t=\mu$ as a stationary solution. Furthermore, we shall prove in section 4 that all initial probability densities ρ_0 converge to μ. Before doing this, however, we develop some tools for solving the PDE (2.9) numerically.

3. Numerical solution of the PDE

3.1. The one-dimensional case. When x, and hence y and z, are one dimensional, (2.9) becomes

$$\frac{\partial\rho_t}{\partial t} = \frac{\partial}{\partial z}\left(\mu^2\frac{\partial}{\partial z}\left(\frac{1}{2}\left(\frac{\rho_t}{\mu}\right)^2\right)\right). \tag{3.1}$$

This equation adopts a simpler and numerically more tractable form if one makes a change of variable from z to the cumulative distribution associated with the target density, $w=\int_{-\infty}^{z}\mu(s)\,ds$:

$$\frac{\partial r_t}{\partial t} = \frac{\partial}{\partial w}\left(\mu^3\frac{\partial}{\partial w}\left(\frac{1}{2}r_t^2\right)\right), \tag{3.2}$$

where $r_t=\rho_t/\mu$ and for simplicity we have assumed that $\mu>0$ on $\mathbb{R}$ (notice that r still integrates to 1, since $\int_{\mathbb{R}} r_t\,dw=\int_{\mathbb{R}}\rho_t\,dz$.)

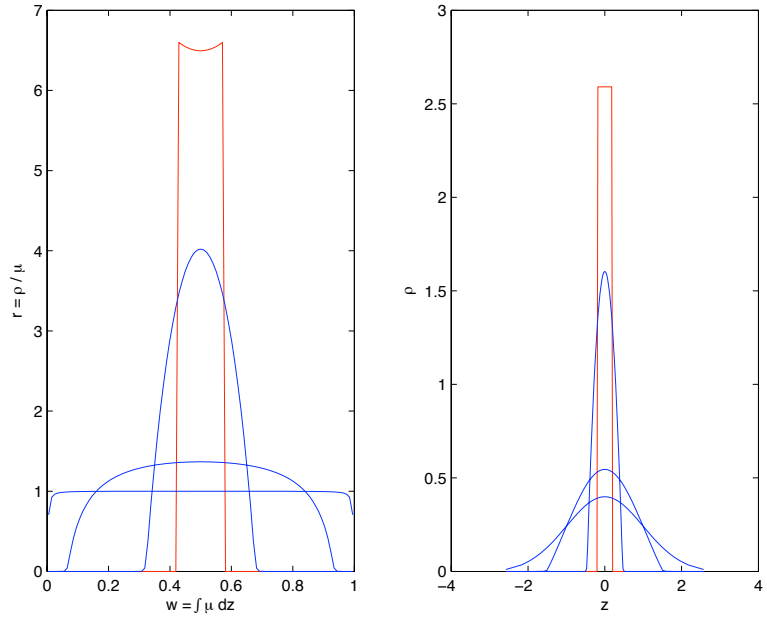

FIG. 3.1. *Numerical solution of (3.2), with a uniform distribution on* $[-0.2,0.2]$ *evolving toward the target Gaussian. The left panel shows snapshots of* r_t *as a function of* w*; the right panel translates this evolution back to the original variables, showing* ρ_t *as a function of* x*.*

Numerically, the form (3.2) is advantageous since it has a finite spatial domain, $0 \le w \le 1$, and a target density $r = 1$ which is uniform in w, making the simplest choice for a numerical grid, the uniform one, also the most effective. At the boundary points $w = 0$ and $w = 1$, one has the no-flux conditions

$$\lim_{w \to 0,1} \mu^3 \frac{\partial}{\partial w} \left(\frac{1}{2} r_t^2 \right) = 0. \tag{3.3}$$

A numerical solution of the PDE (3.2) is displayed in figure 3.1, together with the evolution of the density $\rho_t(z)$. In this run, the initial data $\rho_0(x)$ is concentrated in the interval $|x| < 0.2$, where it is distributed uniformly, and the choice for the target distribution $\mu(y)$ is a standard Gaussian. A noticeable feature of the solution, in addition to its fast convergence from $\rho_0(z)$ to $\mu(z)$, is the persistence of sharp density fronts. Such fronts, which occur when the support of ρ is finite, are ubiquitous in nonlinear diffusive equations.

3.2. Extension to general dimensions. For certain μ's, it is straightforward to extend the one-dimensional procedure to more dimensions. In Cartesian coordinates, (2.9) reads

$$\frac{\partial \rho_t}{\partial t} = \sum_{i=1}^{n} \frac{\partial}{\partial z_i} \left(\mu^2 \frac{\partial}{\partial z_i} \left(\frac{1}{2} \left(\frac{\rho_t}{\mu} \right)^2 \right) \right), \tag{3.4}$$

where each term on the right-hand side has exactly the same form as in the one-dimensional case. If the target density $\mu(z)$ factorizes as a product of one-dimensional

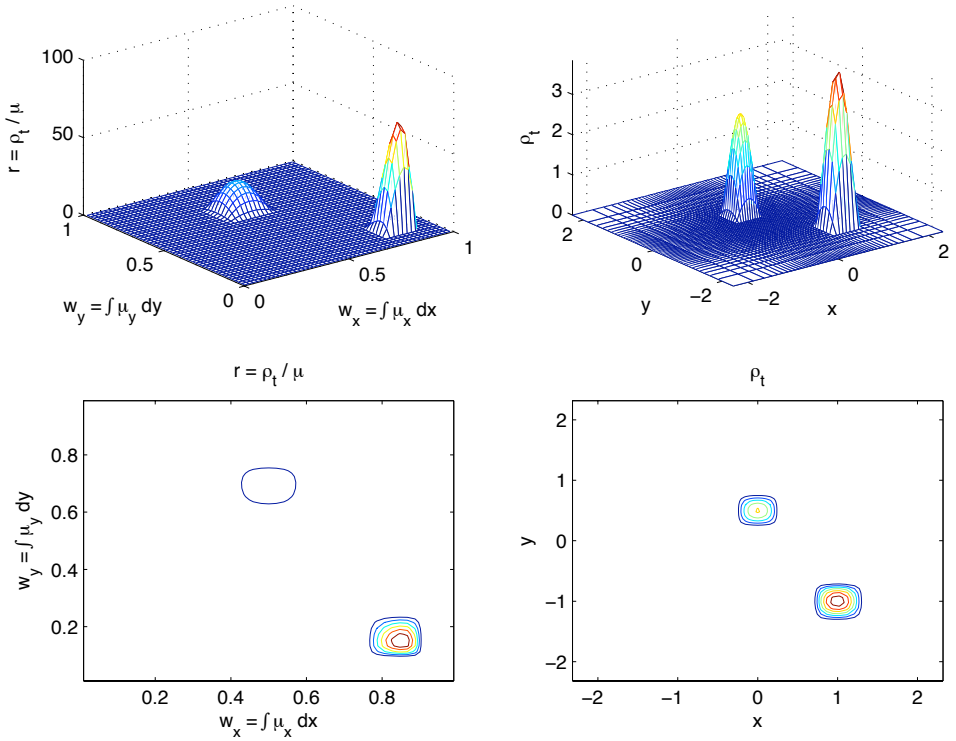

FIG. 3.2. *Numerical solution of (3.6) for the two dimensional example of the density made of the superposition of two bumps described in text. The left panel shows r_t as a function of (w_x, w_y) at time $t=0$; the right panel shows the corresponding ρ_t as a function of (x,y).*

densities,

$$\mu(z) = \prod_{i=1}^{n} \mu_i(z_i), \tag{3.5}$$

as is the case for an isotropic Gaussian, one can introduce variables $w_i = \int_{-\infty}^{z_i} \mu_i(s)ds$ (the cumulative distributions associated with the individual μ_i's), and rewrite (3.4) as

$$\frac{\partial r_t}{\partial t} = \sum_{i=1}^{n} \frac{\partial}{\partial w_i} \left(\mu {\mu_i}^2 \frac{\partial}{\partial w_i} \left(\frac{1}{2} r_t^2 \right) \right), \tag{3.6}$$

where $r_t = \rho_t/\mu$. A numerical example is shown in figures 3.2, 3.3, and 3.4, where a multimodal density, consisting of the superposition of two bumps of the form

$$b_i(x,y) = \alpha_i \left[a_i^2 - (x - x_i^0)^2 \right]_+ \left[a_i^2 - (y - y_i^0)^2 \right]_+, \qquad i = 1,2 \tag{3.7}$$

where $[\cdot]_+ = \max(\cdot, 0)$, evolves into the target isotropic Gaussian $\mu(x,y)$.

4. Evolution of the Kullback-Leibler divergence

In order to prove that the solution ρ_t of the PDE (2.9) always converges to the target density μ, one can consider the Kullback-Leibler (KL) divergence [8, 1] of μ

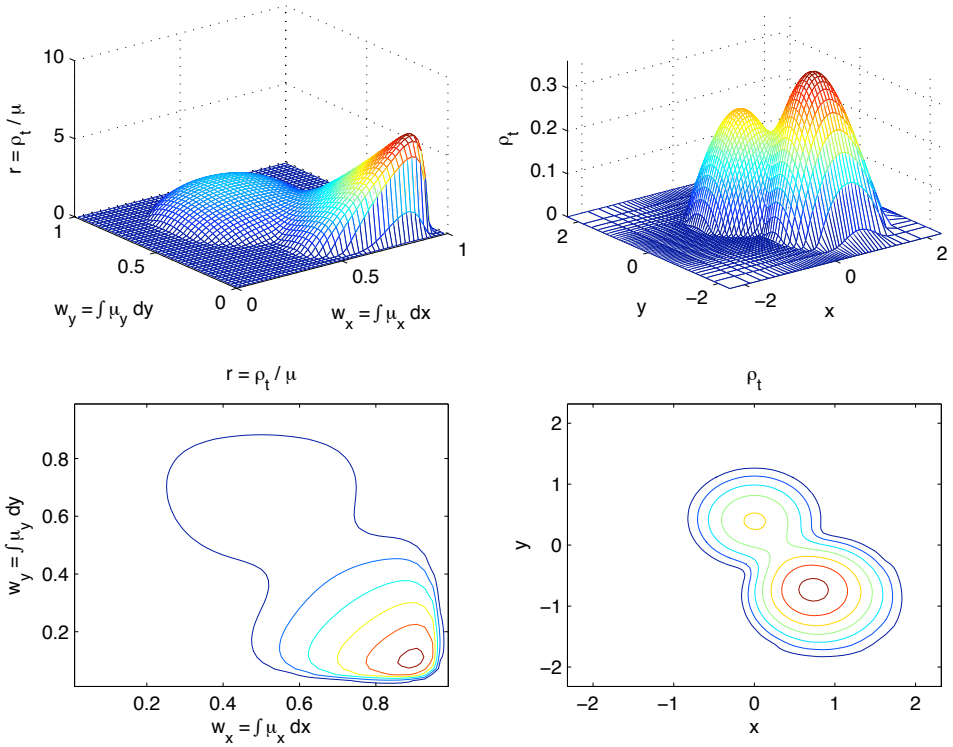

FIG. 3.3. *Same as in figure 3.2 at time* $t=0.5$.

and ρ_t,

$$D_{KL}(\mu,\rho_t)=\int \log\left(\frac{\mu}{\rho_t}\right)\mu\,dz, \tag{4.1}$$

a non-negative, convex function of ρ_t, which achieves its minimum value of zero when $\rho_t=\mu$. Its evolution under (2.9) is given by

$$\frac{d}{dt}D_{KL}(\mu,\rho_t)=-\int \frac{\mu}{\rho_t}\frac{\partial\rho_t}{\partial t}\,dz=-\int\left(\frac{\mu^3}{\rho_t}\right)\left|\nabla\left(\frac{\rho_t}{\mu}\right)\right|^2 dz\le 0, \tag{4.2}$$

with equality if and only if $\rho_t=\mu$. Hence the Kullback-Leibler divergence of μ and ρ_t does not stop decreasing until $\rho_t(z)$ reaches its target $\mu(z)$.

Let us give a more general argument to prove convergence. This argument will be useful later when we constrain the family of flows ϕ_t, and sheds light on the nature of the proposed dual ascent. Consider the Kullback-Leibler divergence of ρ and $\tilde\rho_t$ instead of that of μ and ρ_t:

$$D_{KL}(\rho,\tilde\rho_t)=\int \log\left(\frac{\rho}{\tilde\rho_t}\right)\rho\,dx=S(\rho)-L_\rho(\phi_t), \tag{4.3}$$

where $S(\rho)$ is the time-independent Shannon's entropy of the actual probability density [7],

$$S(\rho)=\int \log(\rho)\,\rho\,dx, \tag{4.4}$$

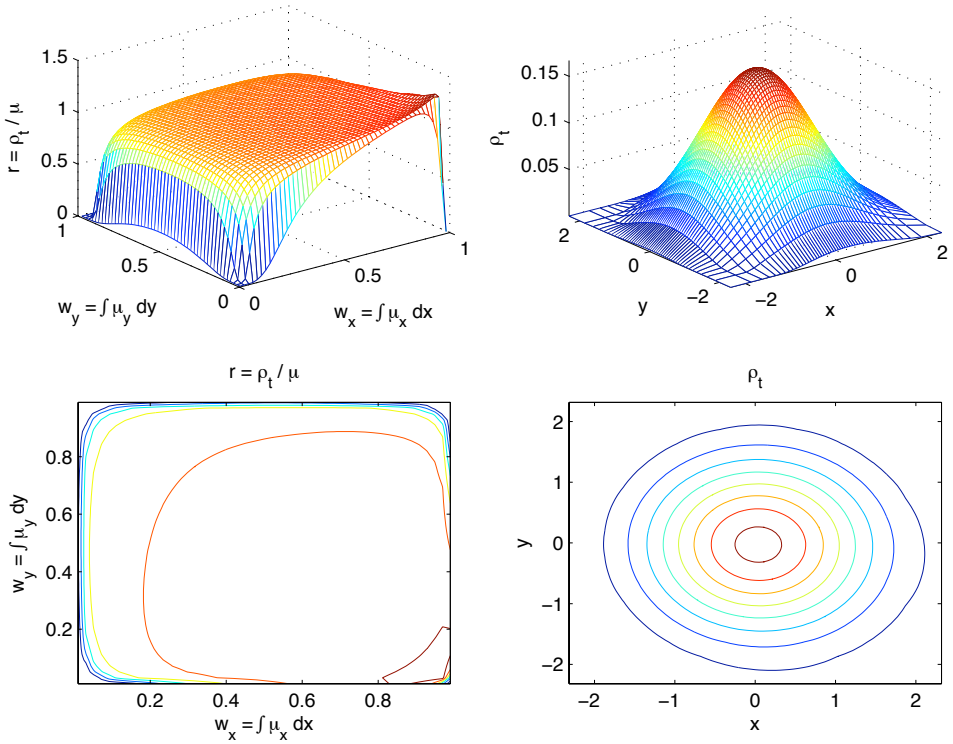

FIG. 3.4. *Same as in figure 3.2 at time* $t = 20$.

and $L_\rho(\phi_t)$ is the log-likelihood in (2.1). Then

$$\frac{d}{dt} D_{KL}(\rho, \tilde{\rho}_t) = -\frac{d}{dt} L_\rho(\phi_t). \tag{4.5}$$

Now consider the map $\phi_t(x)$ as the composition of two maps, $\phi_t(x) = \phi_{t_1+t_2}(x) = (\phi_{t_2} \circ \phi_{t_1})(x)$. Replacing this in the log-likelihood (2.1), and changing variables to $y = \phi_{t_1}(x)$, yields

$$L_\rho[\phi_{t_1+t_2}] = L_{\rho_{t_1}}[\phi_{t_2}] + \tilde{L}_\rho[\phi_{t_1}], \tag{4.6}$$

where

$$L_{\rho_{t_1}}[\phi_{t_2}] = \int \left(\log(J_{\phi_{t_2}}(y)) + \log(\mu(\phi_{t_2}(y)))\right) \rho_{t_1}(y)\, dy \tag{4.7}$$

is the log-likelihood associated with $\phi_{t_2}(y)$ under a distribution $\rho_{t_1}(y)$, and

$$\tilde{L}_\rho[\phi_{t_1}] = \int \log(J_{\phi_{t_1}}(x))\, \rho(x)\, dx \tag{4.8}$$

is a quantity that does not depend on t_2. Hence, at fixed t_1,

$$\frac{d}{dt} L_\rho(\phi_t) = \frac{d}{dt_2} L_{\rho_{t_1}}(\phi_{t_2}). \tag{4.9}$$

On the other hand, modifying the intermediate time t_1 does not affect the value of D_{KL}, just the relative weight of its two components under the partition (4.6). If now we take the limit in which $t_1 \uparrow t$ and $t_2 \downarrow 0$ with $t_1 + t_2 = t$, we obtain precisely the likelihood L_{ρ_t} that determines the flow ϕ_t in (2.4), since

$$\frac{d}{dt_2} L_{\rho_{t_1}}(\phi_{t_2}) \to u_t(\phi_t) \quad \text{as } t_1 \uparrow t \text{ and } t_2 \downarrow 0 \text{ with } t_1 + t_2 = t. \tag{4.10}$$

As a result

$$\frac{d}{dt} D_{KL}(\rho, \tilde{\rho}_t) = -\frac{d}{dt} L_{\rho_t} = -\int \frac{\delta L_{\rho_t}}{\delta \phi_t} \dot{\phi}_t(y)\, dy = -\int |u_t(\phi_t(y))|^2\, dy \le 0, \tag{4.11}$$

with equality only when $\tilde{\rho}_t = \rho$. This shows convergence of the solution of (2.9) towards μ since $\tilde{\rho}_t = \rho$ if and only if $\rho_t = \mu$.

This proof of convergence of the solution of (2.9) extends straightforwardly to more general scenarios where the flows ϕ_t have further constraints. For $\tilde{\rho}_t$ to converge to ρ, it just requires that, for the allowed flows ϕ_t, the implication

$$\rho_t \neq \mu \quad \Rightarrow \quad \frac{\delta L_{\rho_t}[\phi_t]}{\delta \phi_t} \neq 0 \tag{4.12}$$

holds, where the variation is taken at fixed ρ_t.

Next, we discuss a class of restricted maps satisfying this property, that will be instrumental in the development of a flow-based algorithm for density estimation.

5. One-dimensional maps and marginal densities

We consider a family of restricted flows, in which the particles are only allowed to follow a one-dimensional motion, i.e., move only in one particular direction, and with a speed that depends only on their coordinate in that direction.

Given an arbitrary direction θ, one can introduce an associated coordinate system that decomposes the particle position z and flow $\phi_t(z)$ into their components in that direction and its orthogonal complement,

$$x = \begin{pmatrix} x_\theta \\ x_\perp \end{pmatrix}, \qquad \phi_t = \begin{pmatrix} \phi_\theta \\ \phi_\perp \end{pmatrix}. \tag{5.1}$$

If one considers one-dimensional flows of the form

$$\phi_t = \begin{pmatrix} \phi_\theta(x_\theta) \\ x_\perp \end{pmatrix}, \tag{5.2}$$

the log-likelihood in (2.1) becomes

$$L_{\rho_t}[\phi_t] = \int \left(\log\left(\frac{d\phi_\theta}{dx_\theta} \right) + \log(\mu(\phi_t(x))) \right) \rho(x)\, dx. \tag{5.3}$$

If, moreover, for every θ the target density μ admits the factorization

$$\mu(x) = \mu_\theta(x_\theta)\, \mu_\perp(x_\perp), \tag{5.4}$$

then

$$L_{\rho_t}[\phi_t] = L_{\tilde{\rho}}[\phi_\theta] + \tilde{L}, \tag{5.5}$$

where

$$\bar{\rho}(x_\theta) = \int \rho(x)\, dx_\perp \tag{5.6}$$

is the marginal density associated with the direction θ,

$$L_{\bar{\rho}}[\phi_\theta] = \int \left(\log\left(\frac{d\phi_\theta}{dx_\theta} \right) + \log(\mu_\theta(\phi_\theta(x_\theta))) \right) \bar{\rho}(x_\theta)\, dx_\theta \tag{5.7}$$

is a one-dimensional log-likelihood functional, and

$$\tilde{L} = \int \log(\mu_\perp(x_\perp))\, \rho(x)\, dx \tag{5.8}$$

does not depend on the flow ϕ_t.

Then, within this restricted class of flows, the flow that descends the global KL-divergence between ρ_t and μ also descends the divergence between their marginals, $\bar{\rho}$ and μ_θ. Since the one-dimensional maps are unrestricted, we know from the arguments in section 4 that this flow will only stop once $\bar{\rho} = \mu_\theta$. If the direction θ is fixed throughout the flow, clearly this is not equivalent to the global statement that ρ_t equals μ: only one marginal has been properly adjusted.

Consider, however, the following procedure: at each time, all directions θ are considered and, at each point x, the corresponding velocity u_t is computed as the angular average of all the resulting one-dimensional fields. Since this is a superposition of infinitesimal flows, linearity applies, and we conclude as in section 4 that, while not all the one-dimensional flows are stagnant, the KL-divergence between $\tilde{\rho}_t$ and ρ will continue to decrease. But, for each direction θ, the flow only stops when the corresponding marginal $\bar{\rho}$ equals μ_θ. Since all the marginals of two distributions agree only when the distributions are equal, we conclude that the flow will make ρ_t converge to μ, and hence $\tilde{\rho}_t$ to ρ.

The only family of distributions satisfying the factorization requirement (5.4) for all directions θ is the isotropic Gaussian

$$\mu(x) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{|x|^2}{2\sigma^2} \right). \tag{5.9}$$

This is also a natural choice for a target, since Gaussian distributions are ubiquitous, as a consequence of the central limit theorem (CLT). Furthermore, while evolving the particles toward Gaussianity, one is aided by the attractive nature of Gaussian distributions, also a consequence of the CLT. Hence we expect robustness of the convergence under observational and numerical noise.

6. Back to density estimation

Clearly, we do not really know the probability distribution $\rho(x)$ — else there would be no problem to solve —, but just the finite sample consisting of the m observations x^j. Yet the procedure described above extends very naturally to this discrete scenario. The points x^j are natural Lagrangian markers for the flow ϕ_t. Additional points x where one seeks to evaluate the density $\rho(x)$ can be carried along passively by the flow $\phi_t(x)$. The only new requirement is to define, at each time, a finite-dimensional class of maps so that one can compute the gradient of the log-likelihood L in (1.6) with respect to the corresponding parameters, which is the discrete version of the variation with respect to ϕ_t in the continuous case.

6.1. Random directions. From the discussion in section 5, it is enough to concern ourselves with one-dimensional maps. Every row of the matrix $X = \{x_i^j\}$ is a sample of the marginal with respect to all the other variables. In order to obtain marginals in other directions, is is enough to rotate the matrix X though an orthogonal matrix U. One simple algorithmic choice is the following:

At each time step, given the matrix $Z = \{z_i^j\}$ of current position of the particles, rotate it through a randomly chosen orthogonal matrix U:

$$Z \to UZ.$$

(We do not need to think of this as an actual particle movement; it is more natural to view it as a change of coordinates. Notice that orthogonal transformations have unitary Jacobians, and henceforth no effect on the estimated density $\tilde{\rho}_t$.)

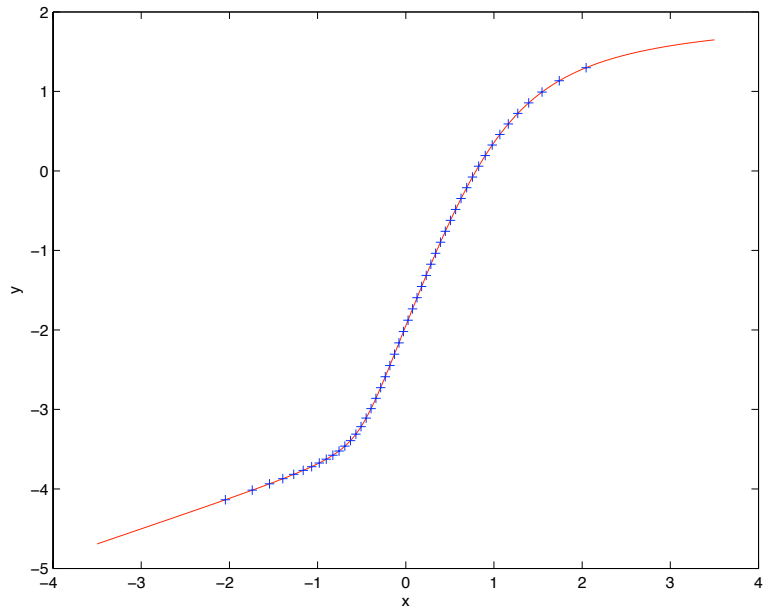

FIG. 6.1. *The solid curve shows the result of the successive application of two maps (6.1) on the identity function. The crosses are points on a equiprobability grid of a Gaussian distribution. Notice that the choice of mollifier ϵ in (6.2) guarantees that the smaller the density of crosses, the more aggressive the mollification of the map.*

6.2. A family of maps. After rotation, one seeks, for each row i of Z, a near-identity map $\phi(z_i)$ that moves it toward Gaussianity. These maps need to satisfy a few requirements.

1. The maps must be smooth enough so that their lengthscale at all points is larger that the typical distance between flow-markers nearby. This is required to not over-resolve the density and make it converge to a set of approximate delta-functions centered at each observation (it amounts to a regularization of the log-likelihood function (1.4) when the number of observations is finite).

2. The maps must be flexible enough to accommodate for quite arbitrary distributions $\rho(x)$, while remaining simple and computationally manageable. This is not an impossibly challenging requirement: the full map between x and y results from the composition of the many near-identity maps that discretize the time evolution of the continuous flow $\phi_t(x)$. Hence it is enough to have among these elementary maps robust building blocks for general transformations.
3. The maps need to be explicit and to have explicitly computable derivatives with respect to z (for the Jacobian) and to their parameters (for the variation), so that the flow ascending the log-likelihood can be determined easily.

In this paper, we selected the following simple five-parameter $(\gamma,\sigma,x_0,\varphi_0,\epsilon)$ family which satisfies these requirements:

$$\varphi(x) = (1-\sigma)x + \varphi_0 + \gamma\sqrt{\epsilon^2 + [(1-\sigma)x - x_0]^2}. \tag{6.1}$$

When γ, σ and φ_0 are zero, the map reduces to the identity. The parameter σ quantifies the amount of stretching; φ_0, displacement; and γ, the slope change at x_0, where it switches between $d\varphi/dx \approx 1-\sigma-\gamma$ and $d\varphi/dx \approx 1-\sigma+\gamma$. The parameter ϵ mollifies the transition between the two slopes of the map to the left and right of x_0. Its value is x_0-dependent:

$$\epsilon = \sqrt{2\pi} n_p \exp\left(\frac{x_0^2}{2}\right), \tag{6.2}$$

where n_p is the desired average number of data points within the transition area (the length of the transition needs to be larger in sparsely populated areas, not to over-resolve the map where there are few points).

In each step of the descent algorithm, the parameters γ, σ, and φ_0 are chosen close to zero, yielding near-identity transformations, in the ascent direction. The other parameters are externally provided, not selected by ascent: x_0, the location of the slope switch, is picked at random from a standard normal distribution, so that, near convergence, the number of opportunities for local distortions is proportional to the density of observations, and ϵ is determined by (6.2).

This family constitutes a simple building block for general maps (see figure 6.1): it consists of a mollified, continuous piecewise linear function, which changes slope at x_0. Without mollification, which implies a smoothness condition, it is clear that any transformation $f(x)$ can be built as a superposition of such elementary maps.

6.3. Ascent. At each step, the parameters $\alpha = (\gamma,\sigma,\varphi_0)$ in (6.1) are picked by ascent, e.g.,

$$\alpha \propto \nabla_\alpha L, \tag{6.3}$$

where L is the one-dimensional version of the log-likelihood in (1.6), with the x^j replaced by z_i^j. The gradient is evaluated at $\alpha = 0$, corresponding to the identity map.

A procedure that we found effective is to pick

$$\alpha = \Delta t \frac{\nabla_\alpha L}{\sqrt{1+\delta^2|\nabla_\alpha L|^2}}, \tag{6.4}$$

where Δt and δ are adjustable parameters which control the size of the ascending steps.

6.4. Computational effort. The amount of computational effort required by the algorithm depends on the number of observations (m), the number of variables (n), and the accuracy desired, which influences the time step Δt, the mollification parameter n_p, and the number of steps n_s.

Every time-step, each variable x_i ascends the log-likelihood independently, so the associated effort is linearly proportional to n (and trivially parallelizable). It is also linear in m, since the log-likelihood and its gradient consist of sums over the available observations, and the map is performed observation-wise (The map is also performed on the extra marker points where the density is sought, that the algorithm carries passively; for the purpose of counting, we are including these points in the total m.) Hence the effort of the core of the algorithm is linear in m and n.

Each time step also includes a random rotation. Constructing a general random unitary matrix involves $O(n^3)$ operations; performing the rotation adds $O(n^2 m)$ operations. This is not too expensive when the number of variables is small. When n is large, on the other hand, more effective strategies can be devised:

- The unitary transformations do not need to be random. In particular, they can be read off-line. This has the additional advantage of allowing one to keep track of the full map $y(x)$, not just of the image of the tracer points x^j and the corresponding Jacobian. The full map is useful in a number of applications, such as the calculation of nonlinear principal components, and the addition of new observations half-way through the procedure.
- The transformations may have extra structure. For example, they may consist of the product of rotations along planes spanned by random pairs of coordinate axes. This makes their matrices sparse, reducing the cost of performing a rotation to $O(nm)$ operations, and the amount of matrix entries to read or compute to $O(n)$. With this, the complete algorithm is linear in n and m.

As for the number of steps n_s, it is more difficult to offer precise estimates. We have found empirically that, for a desired accuracy, this number is roughly independent of the number of variables and observations. If n_s is picked small to economize effort, the time-step Δt should be correspondingly large, and the mollification parameter n_p small (with few steps, the risk of over-fitting dissapears, and one should permit the most effective maps).

7. Examples Figure 7.1 shows the results of a one-dimensional simulation, where the algorithm is applied to a sample of 200 observations drawn from the centered exponential density (it is convenient, though not necessary, to start the procedure by removing the mean of the observations)

$$\rho(x) = \begin{cases} e^{-(x+1)} & \text{if } x \geq -1, \\ 0 & \text{otherwise.} \end{cases} \tag{7.1}$$

The top-left plot displays a histogram of the sample. The top row shows the evolution of this histogram, as the sample is deformed by the normalizing flow. The bottom row of plots shows the evolution of the corresponding estimate of the original density, computed on a uniform grid.

The duality is clear: as the distorted sample becomes more Gaussian, the estimate for the original $\rho(x)$ moves from the original Gaussian guess to a sensible exponential. There are more subtle manifestations of duality too: the still finite remainder, after 1000 iterations, of the original discontinuity on the left of the histogram, translates into a smoothed discontinuity on the left of $\rho_t(x)$.

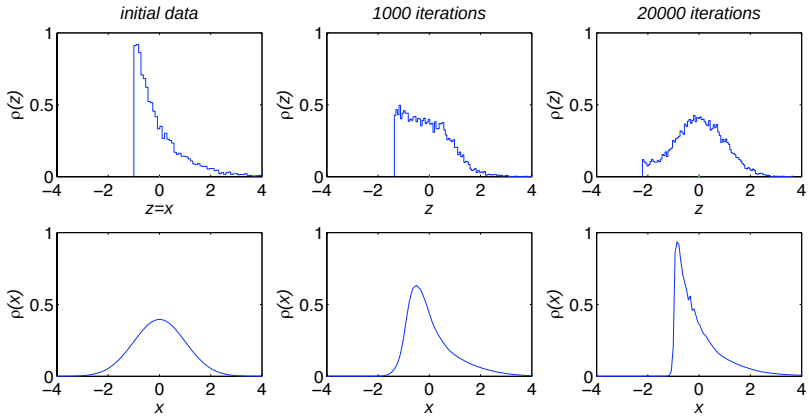

FIG. 7.1. *Illustration of the algorithm on a sample drawn from a one-dimensional centered exponential density. The top panels show the evolution of the empirical distribution toward the target Gaussian. The bottom panels show the dual evolution of the density of the original sample as estimated by the procedure. Note that, as the empirical density evolves from the exponential toward the Gaussian, the estimated density evolves from the Gaussian toward the exponential.*

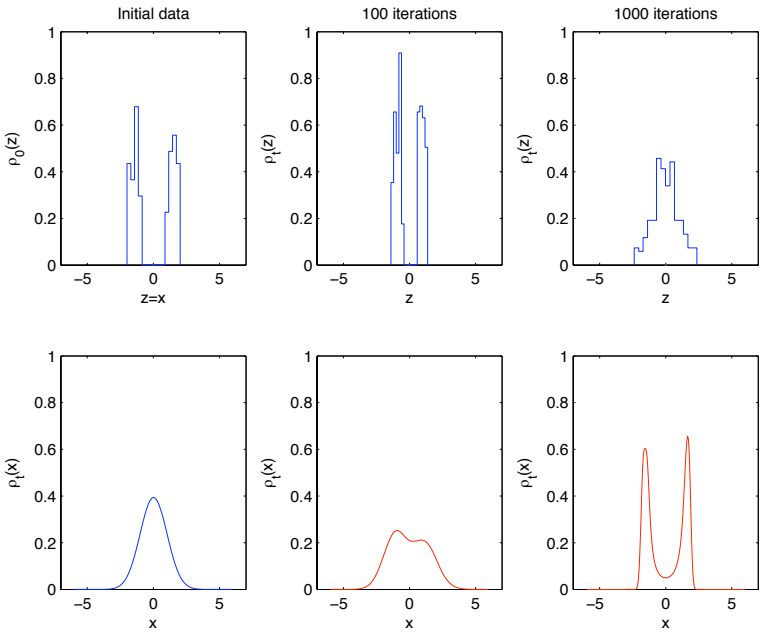

FIG. 7.2. *Same as in figure 7.1 for a sample drawn from a distribution uniform on two intervals. In this figure, the number of data point is* 200 *only. This explains why the density produced by the procedure is not as sharply separated into two as the original density. The quality of the estimated density improves with the number of sampled points, as illustrated in figure 7.3.*

A similar effect of duality can be seen in figure 7.2, which displays a run where the original sample consists of 200 observations of a distribution concentrated in two disjoint segments of the line. The fact that, after 1000 iterations, the two components

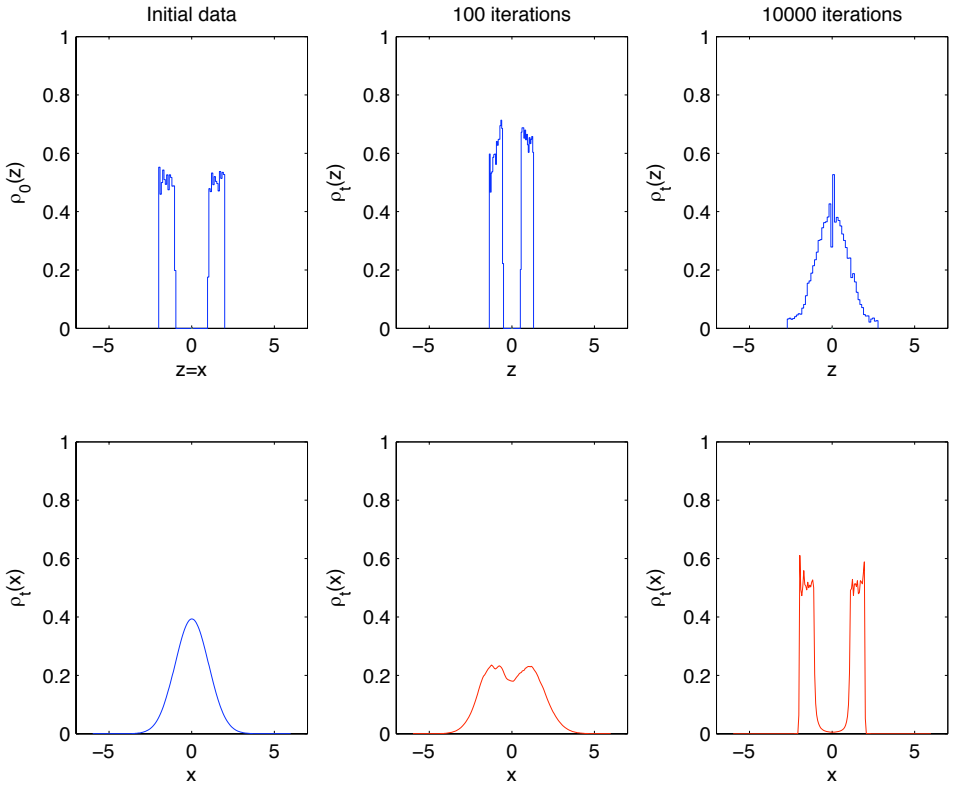

FIG. 7.3. *Same as in figure 7.2 but with a sample containing* 10000 *points. Note that the two modes of the density are much better separated than in figure 7.2.*

of the sample are not yet completely integrated, gives rise to a connected, though tenuous, estimate $\rho_t(x)$.

This lack of a sharp divide (and of a sharp discontinuity in this and the previous case) is a consequence of (i) the finite number of iterations, (ii) the smoothness imposed by the parameter ϵ in (6.2), and (iii) the smallness of the number of observations in the sample. The quality of the estimated density improves as the size of the sample increases, as illustrated in figure 7.3, where the number of observations has risen to 10000 and the parameter n_p to 100 points, yielding discontinuities and separation between populations which are quite distinct.

Figure 7.4 shows the results of a two-dimensional simulation, where the algorithm is applied to a sample of 200 observations drawn from a density made by mixture of three Gaussians:

$$\rho(x,y) = \sum_{j=1}^{3} p_j N_j(x,y),$$

where p_j are positive weights adding to 1 and $N_j(x,y)$ are three Gaussian densities with different means and covariances. Clearly the algorithm yields a very sensible estimation for the density underlying the data. The top row of panels displays not only the evolution of the particles $z^j = \phi_t(x^j)$, but also that of the grid points used to

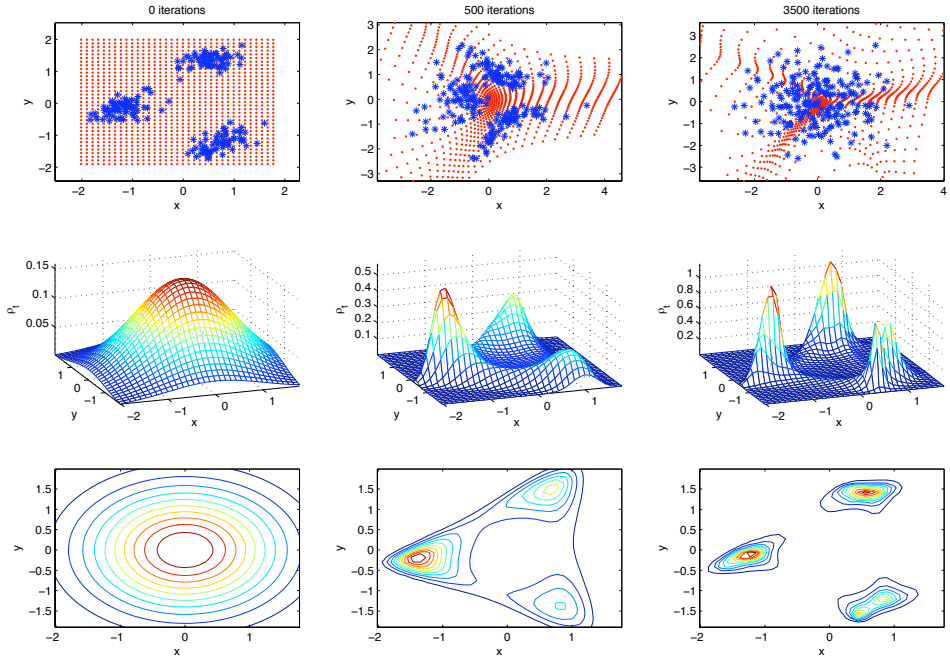

FIG. 7.4. *Illustration of the procedure on a two-dimensional example with a density made by mixture of 3 Gaussians. The top panels show the evolution of the* 200 *sample points as well as that of the grid carried passively by the algorithm. The middle and bottom panels show, respectively, the three-dimensional plot and contourplots of the estimated density as it evolves from a Gaussian toward the estimation of the Gaussian mixture associated with the sample.*

plot the resulting density, which are carried passively by the algorithm. Those grid points which are located in areas with negligible probability are mapped far away, to the tail of the target Gaussian μ.

8. Concluding remarks

A methodology has been developed to compute a robust estimate of the joint probability distribution underlying a multivariate observational sample. The proposed algorithm maps the original variables onto a jointly Gaussian set by ascent of the log-likelihood of the sample. This ascent is performed through near-identity, one-dimensional transformations that push the marginal distribution of each variable toward Gaussianity along a random set of orthogonal directions.

For ease of visualization, the methodology has been exemplified here through the density estimation of synthetic data in one- and two-dimensions. Yet the methodology works in high dimensions too; examples of its specific application to medical diagnosis will be reported elsewhere.

Acknowledgments. This work benefited from discussions with many collaborators and friends. The original motivation arose from a problem posed by the cardiac transplant research group at Columbia University, particularly through Martín Cadeiras and the group's director, Mario Deng. Cristina V. Turner and her numerical analysis team, from the University of Córdoba, Argentina, were a constant source of support. We also thank Paul A. Milewski, from the University of Wisconsin at Madi-

son, and our NYU colleagues Oliver Bühler, Charles Peskin and Marco Avellaneda, for helpful discussions and advice.

REFERENCES

[1] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
[2] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, Springer, 2001.
[3] J.M. Marin, K. Mengersen and C.P. Robert, *Bayesian modelling and inference on mixtures of distributions*, Handbook of Statistics, D. Dey and C.R. Rao (eds.), Elsevier-Sciences, 25, 2005.
[4] D.X. Li, *On default correlation: a copula function approach*, The RiskMetrics Group, working paper, 99–107, 2000.
[5] J.H. Friedman, W. Stuetzle and A. Schroeder, *Projection pursuit density estimation*, J. Amer. Statist. Assoc., 79, 599–608, 1984.
[6] S.S. Chen and R.A. Gopinath, *Gaussianization*, T.K. Leen, T.G. Dietterich, and V. Tresp (eds.), Advances in neural information processing systems, Cambridge, MA: MIT Press, 13, 423–429. 2001.
[7] C.E. Shannon, *A mathematical theory of communication*, Bell System Technical Journal, 27, 379-423, 623–656, 1948.
[8] S. Kullback and R.A. Leibler, *On information and sufficiency*, Annals of Mathematical Statistics, 22, 79–86, 1951.
[9] C. Gu and C. Qiu, *Smoothing spline density estimation: theory*, Annals of Statistics, 21, 217–234, 1993.