Density estimation through flows in feature space

Esteban G. Tabak
(NYU, Courant Institute)
with
Peter M. Laurence
Ricardo Pignol
Cristina V. Turner
Eric Vanden-Eijnden

Paris, April 2011

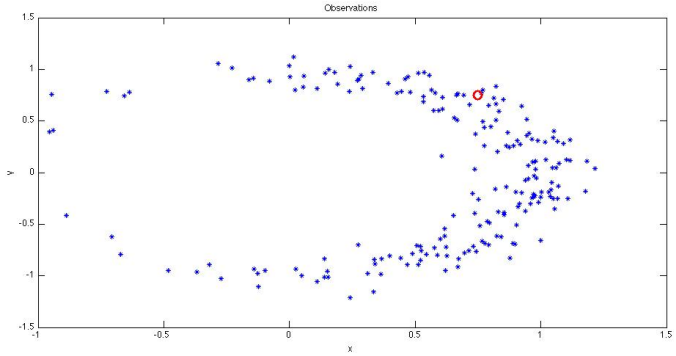
Main topics of this talk

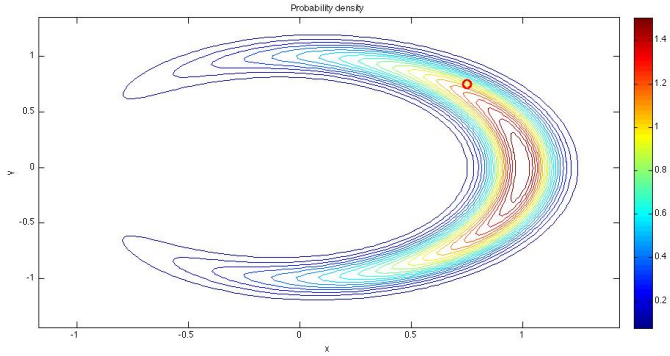
- Probability density estimation
- Conditional density estimation
- Time series analysis
- Density estimation with constraints

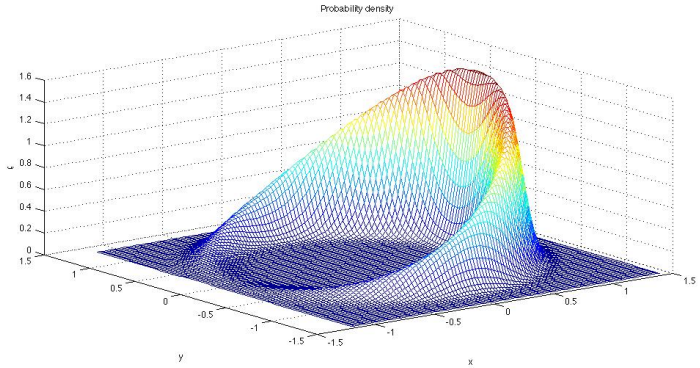
Density estimation

Consider first a set of independent observations x_j of a continuous vectorial variable $x \in R^n$. We would like to infer from the data a probability distribution $\rho(x)$.

Density estimation







Blue dots: observations; red circle: point where density is sought.

Likelihood of the observations:

$$L = \prod_i \rho(x_i).$$

Log-likelihood:

$$\log(L) = \sum_i \log\left(\rho(x_i)\right).$$

Maximize log-likelihood. Parametric approach:

$$\beta = \arg\max_{\beta \in A} L = \sum_{j=1}^{m} \log\left(\rho(x_j; \beta)\right).$$

Normalizing approach

Seek a transformation

$$x \to y(x;\beta)$$

such that $\mu(y)$ is an isotropic Gaussian:

$$\mu(y) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}|y|^2}$$

Then

$$\rho(x;\beta) = J^{y_\beta}(x)\, \mu(y(x;\beta))$$

$$\beta = \arg\max_{\beta \in A} L = \sum_{j=1}^{m} \left[\log\left(J^{y_\beta}(x_j)\right) - \frac{\|y_\beta(x)\|^2}{2} \right]$$

Example: linear maps $\rightarrow$ Gaussian estimation

$$y_\beta(x) = A(x-b) \quad \left(\beta = \left\{A \in R^{n\times n}, b \in R^n\right\}\right),$$

$$\max L \quad \rightarrow \quad b = \bar{x}, \quad A = \Sigma^{-\frac{1}{2}}$$

$$\rho(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\bar{x})^t\Sigma^{-1}(x-\bar{x})}.$$

Map composition

$$y_N(x;\beta) = \phi_{\beta_N} \circ \phi_{\beta_{N-1}} \circ \ldots \circ \phi_{\beta_1}(x),$$

with a family of "building blocks" ϕ_β that includes the identity:

$$\phi_0(z) = z$$

$$J^{y_i}(x_j) = J^{\phi_i}\left(y_{i-1}(x_j)\right) J^{y_{i-1}}(x_j)$$

A gradual flow

$$x \to y(x;t) = \phi_t(x)$$

$$u(z) = \frac{\partial z}{\partial t}$$

Observations x_i: active Lagrangian markers.
Points x_j where density is sought: passive Lagrangian markers.

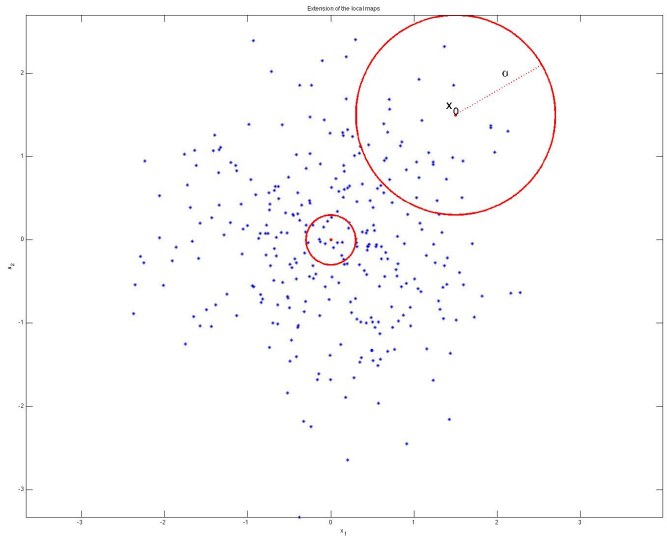
Dynamics: memory-less ascent of the log-likelihood.

Duality: as $t \to \infty$,

- $\rho_t(y) \to \mu(y)$ Actual density of $y(x;t)$
- $\tilde{\rho}_t(x) \to \rho(x)$ Estimated density of x

Elementary maps

$$y = x + \phi\left(\frac{x - x_0}{\alpha}\right), \quad \alpha = (2\pi)^{\frac{1}{2}} \left(\Omega_n^{-1} \frac{n_p}{m}\right)^{\frac{1}{n}} e^{\frac{\|x0\|^2}{2n}}$$



For maps centered in relatively unpopulated areas, the radius α must be larger than in areas with high probability density, so as to encompass a similar number of points.

A simple building block,

robust in high dimensions:

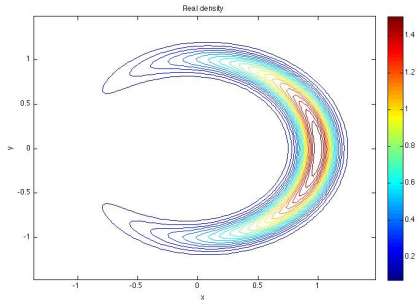
$$y = x + \beta f(\|x - x_0\|)(x - x_0),$$

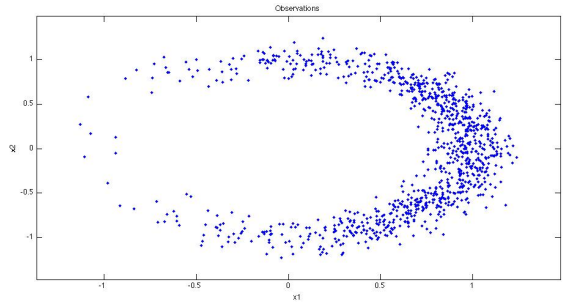
$$f(r) = \frac{1}{\alpha} \frac{\operatorname{erf}\left(\frac{r}{\alpha}\right)}{\frac{r}{\alpha}},$$

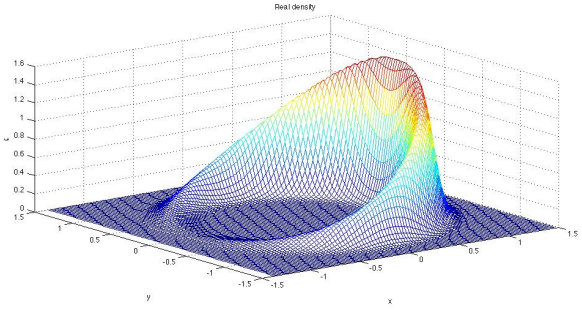
x_0 random, $\quad \alpha = \alpha(x_0)$

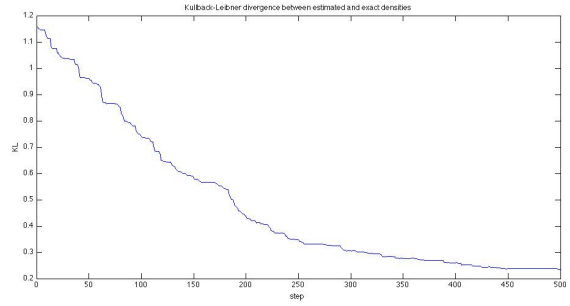
$$\beta = -\frac{L_\beta}{L_{\beta\beta}}.$$

A synthetic two-dimensional example

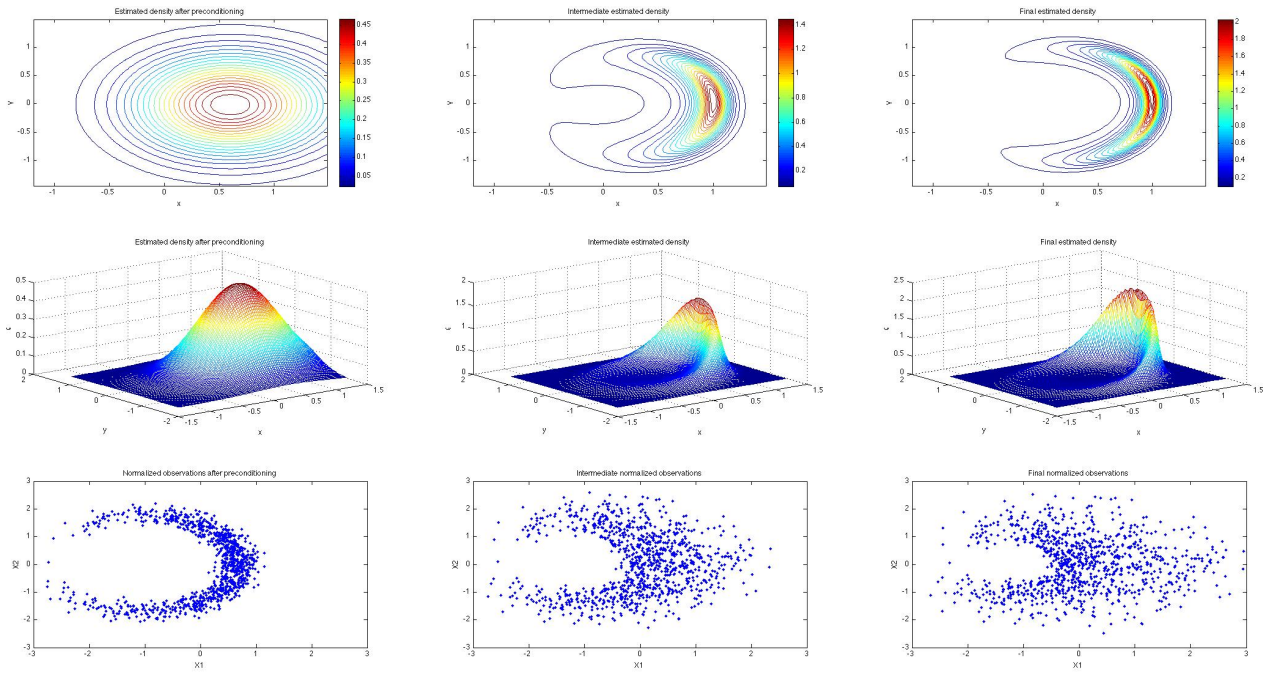






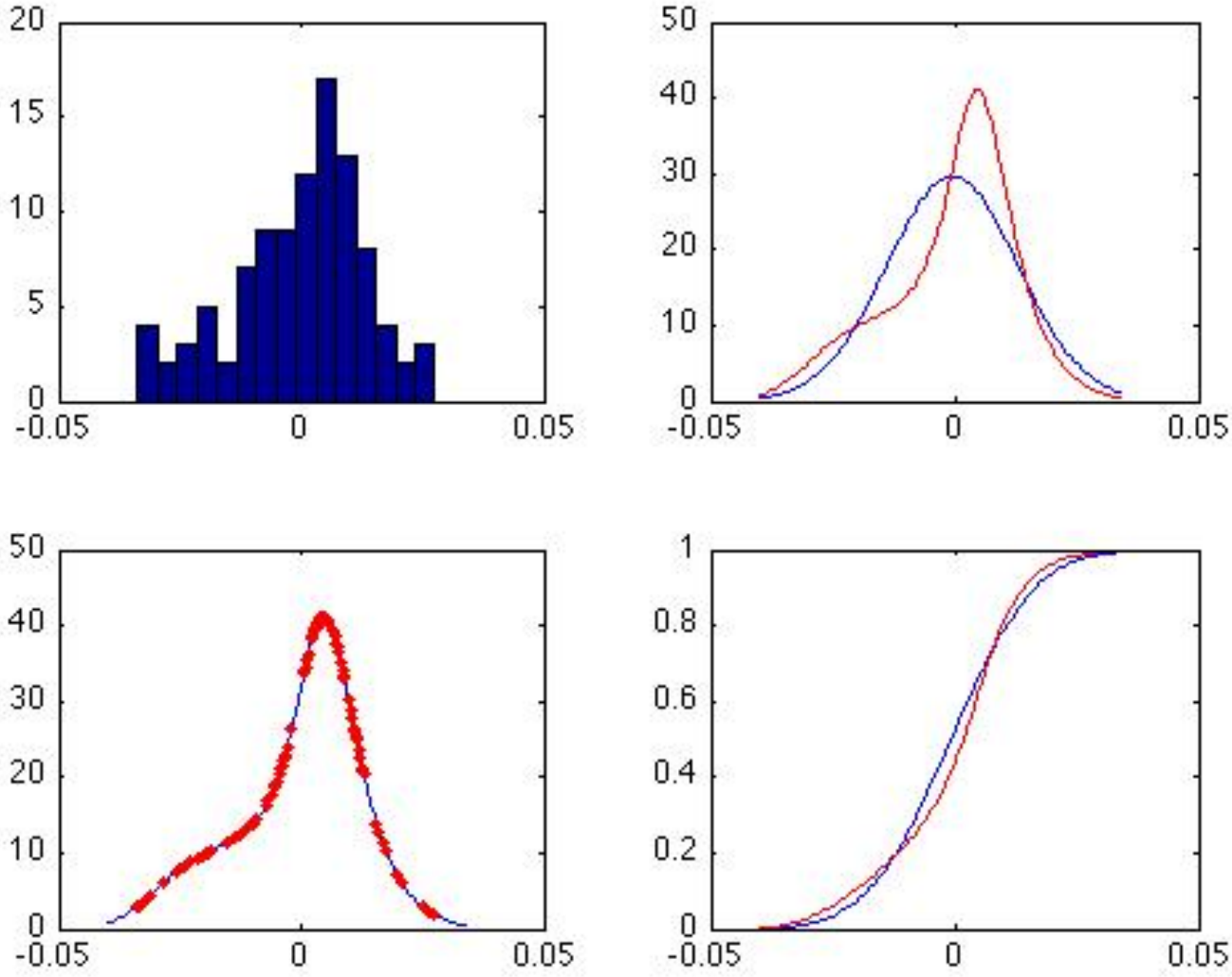


On the left, the proposed probability density, displayed through contours and in perspective. On the right, the 1000 point sample used to test the procedure, and the evolution of Kullback-Leibler divergence between the analytical density and the one discovered by the algorithm.



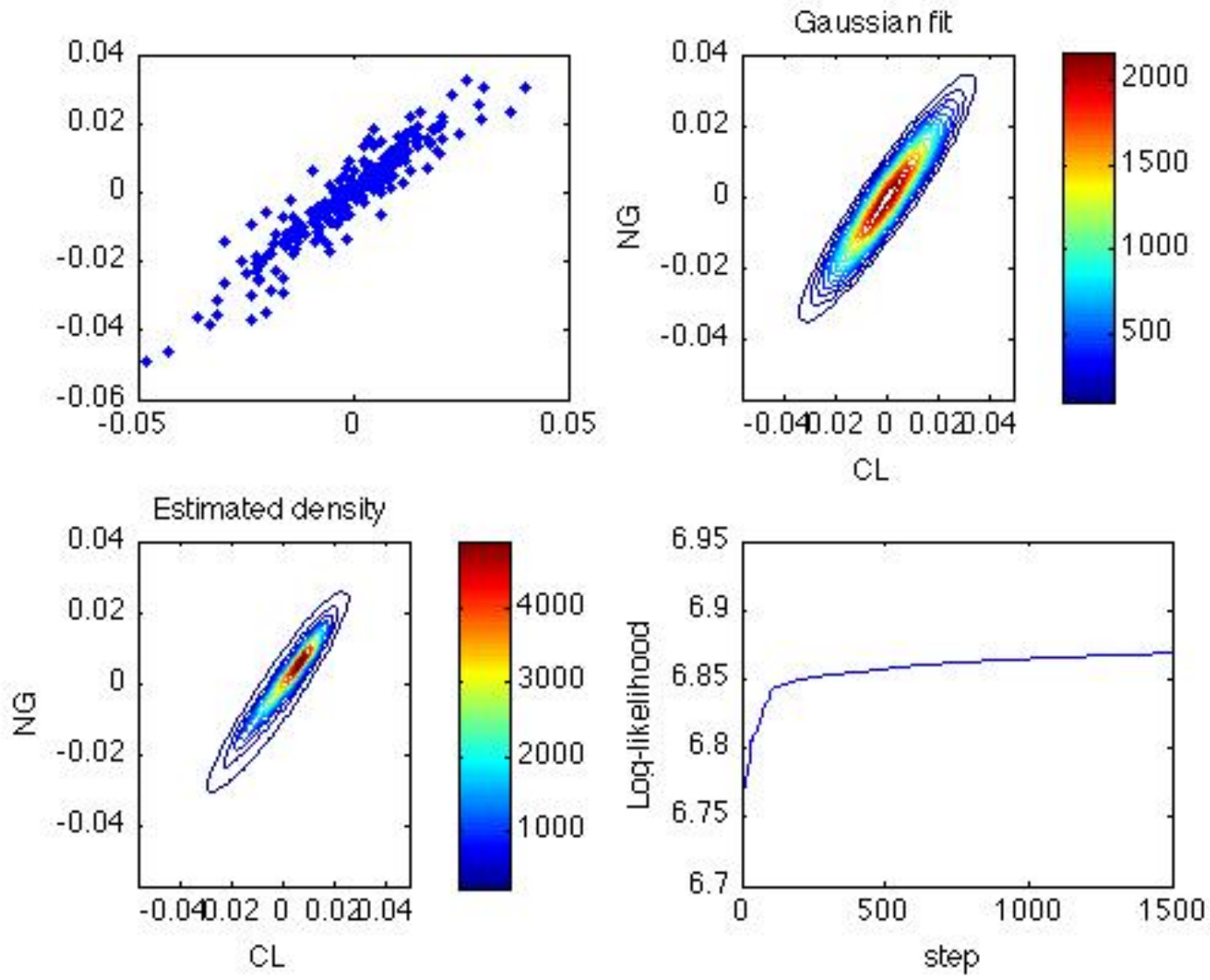
Evolution of the estimated density and normalized observations, through three snap-shots: on the left, the onset of the algorithm, after a pre-conditioning step that re-centers the observations and rescales them isotropically; in the center, the situation after 200 steps and, on the right, a final estimation after 600 steps.

100 days of crude oil future returns



Blue: Gaussian fit. Red: the algorithm.

Joint density of futures for crude oil and natural gas



Conditional density estimation

$$\rho(x|s) = \rho(x; s),$$

as opposed to

$$\rho(x|s) = \frac{\rho(x, s)}{\rho(s)}.$$

Same idea, but with the flow in x space depending on x *and* s:

$$y = y(x; s), \quad \rho(x|s) = \frac{\delta y}{\delta x}(x; s)\, \mu(y(x; s)).$$

A building block:

$$y = x + \beta g(\|s - s_0\|) f(\|x - x_0\|)(x - x_0),$$

$$g = e^{-\frac{\|s - s_0\|^2}{\gamma}}.$$

A synthetic example

A Gaussian mixture, with weights depending on s:

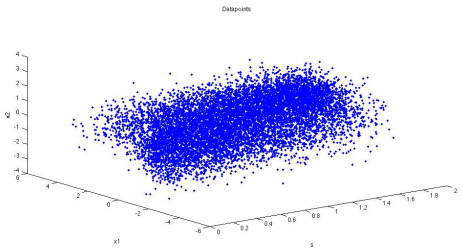
$$\rho(x|s) = p(s)\mathcal{N}(x|\mu_1, \Sigma_1) + (1 - p(s))\mathcal{N}(x|\mu_2, \Sigma_2),$$

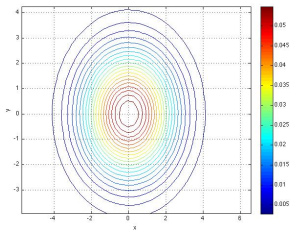
$$p(s) = s/2, \quad 0 \leq s \leq 2,$$

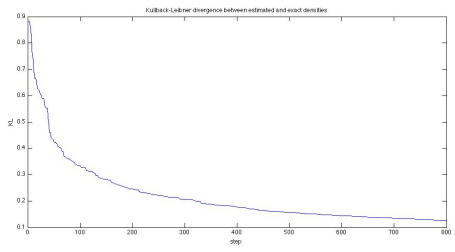
$$\mu_1 = (-2, 0), \quad \mu_2 = (2, 0),$$

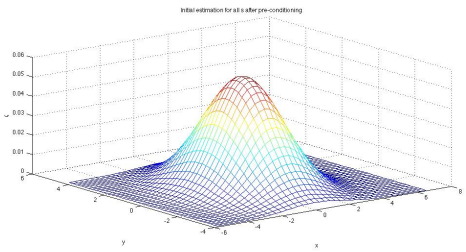
$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0.5 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix}.$$

A synthetic example



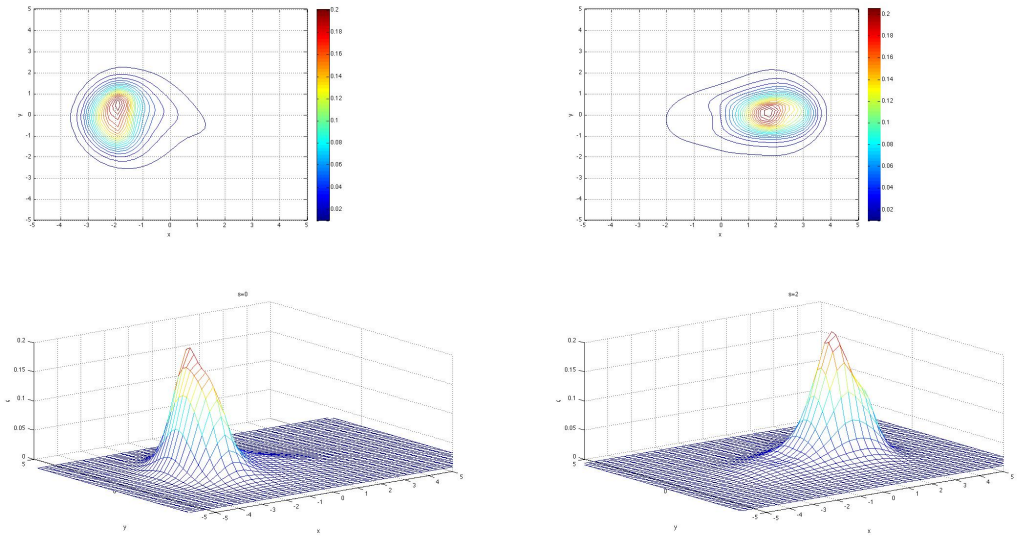






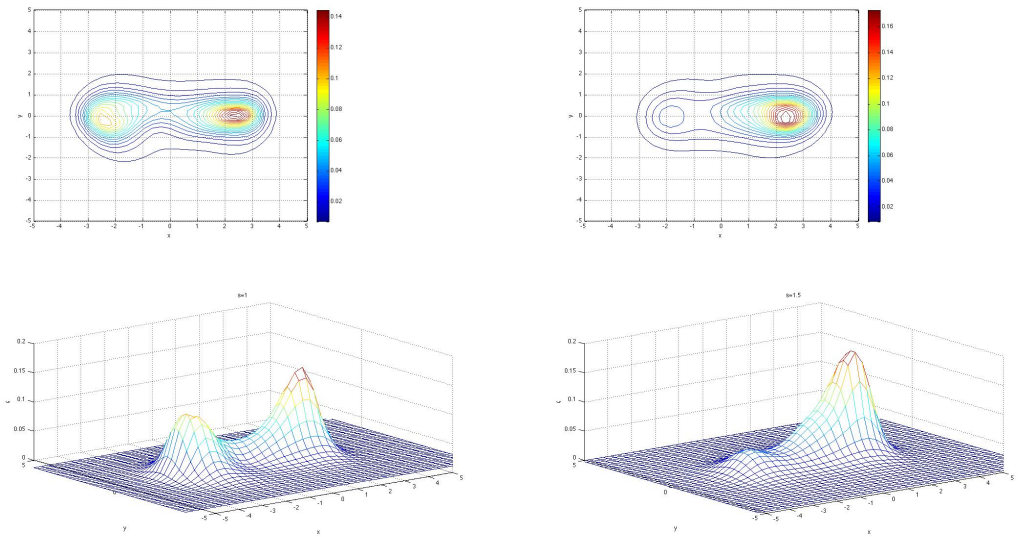
Datapoints drawn from the example above, preconditioning, and evolution of the relative entropy with the analytical answer.

A synthetic example



Estimated conditional density for two extreme values of s.

A synthetic example



Estimated conditional density for two intermediate values of s.

Time series

Seek a conditional density (the dynamics):

$$\rho(x_{j+1}|x_j, S),$$

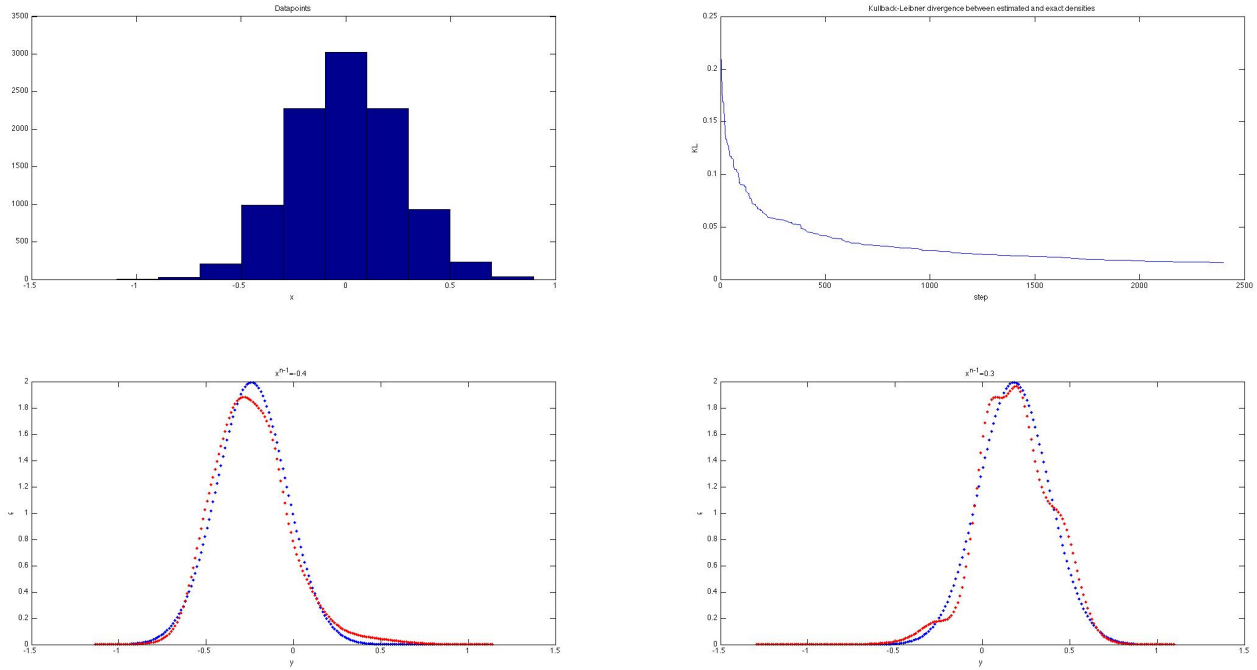
where the vector S may contain time, all exogenous variables and external controls.

Example:

$$x^{n+1} = 0.6x^n + 0.2w^n.$$

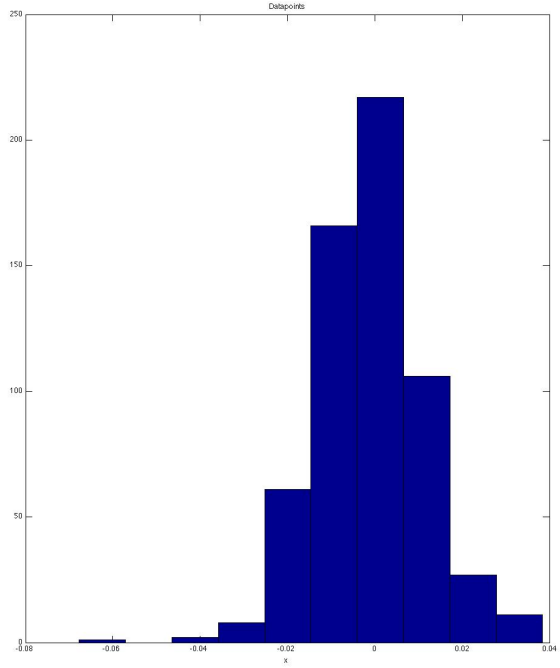
$$\rho\left(x^{n+1}|x^n\right) = \mathrm{N}\left(0.6x^n, 0.04\right).$$

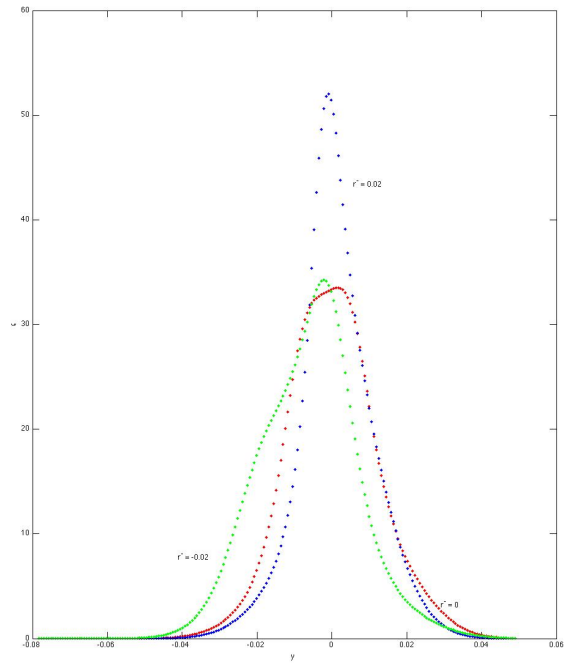
Time series: a synthetic example.



Densities of x^n for two values of x^{n-1}: -0.4 and 0.3. Theoretical conditional distribution in blue, distribution found by the algorithm in red.

Time series for crude oil future returns





Density estimations for r_t conditioned to $r_{t-1} = -0.02$, 0 and 0.02. On the left, histogram of returns for 600 days.

Density estimation with constraints

$$\rho(x) \quad \text{satisfying} \quad \mathrm{E}(f_i(x)) = \bar{f}_i.$$

- Independently observed.
- Better documented.
- Theoretical constraints, such as no arbitrage.

Density estimation with constraints

Find flows such that, at each step,

$$\int f_i(x)\rho^-(x)\ dx = \int f_i(x)\rho^+(x)\ dx.$$

$$z(y) = y + \varphi(y).$$

$$\int f_i(x(y))\nabla \cdot [\mu(y)\varphi(y)]\ dy = 0.$$

$$\varphi(x) = \beta \sum_{k=1}^{q+1} \gamma_k \varphi_k(x)$$

$$\sum_{k=1}^{q} \gamma_k^2 = 1$$

$$I_i^k = \int f_i(x(y))\nabla \cdot [\mu(y)\varphi_k(y)] \; dy$$

evaluated through Montecarlo simulation with importance sampling:

$$I_i^k = \int \psi_i^k(y)\eta(y) \; dy$$

$$\psi_i^k(y) = \frac{f_i(x(y))}{\eta(y)} \nabla \cdot [\mu(y)\varphi_k(y)] \, .$$

$$I_i^k \approx \frac{1}{s} \sum_{i=1}^{s} \psi(y_i^g)$$

Fine tuning

$$\hat{f}_i^{-} = \int f_i(x)\rho^{-}(x)\ dx = \int f_i(x(y))\mu(y)\ dy$$

$$z(y) = y + \varphi(y)\,, \quad \varphi(y) = \sum_{k=1}^{q} \gamma_k \varphi_k(y)$$

$$\hat{f}_i^{+} \approx \hat{f}_i^{-} + \int f_i(x(y))\nabla \cdot [\mu(y)\varphi(y)]\ dy.$$

Setting $\hat{f}_i^{+} = \bar{f}_i$ results in a linear system of equations for the γ_k's.

Summary

- Density estimation through normalizing maps.
- Maps constructed through smooth flows, driven by the gradient of the log-likelihood, with observations playing the role of active Lagrangian markers.
- Simple, effective "building blocks".
- For conditional density estimation, the maps depend also on the variables we are conditioning upon. In time series, on external parameters and prior states.
- Density estimation with constraints via importance sampling.