Diffeomorphic Warping

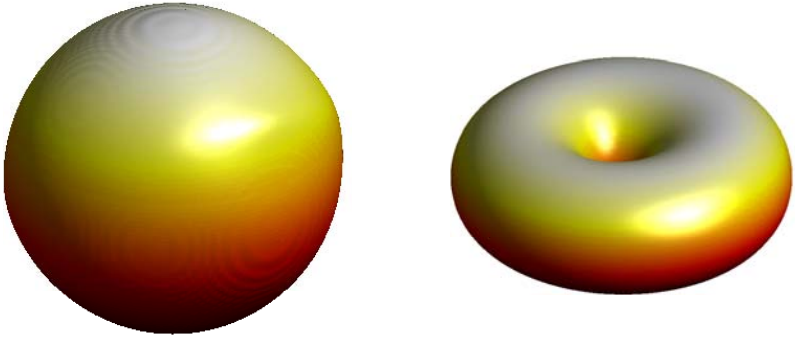
Ben Recht

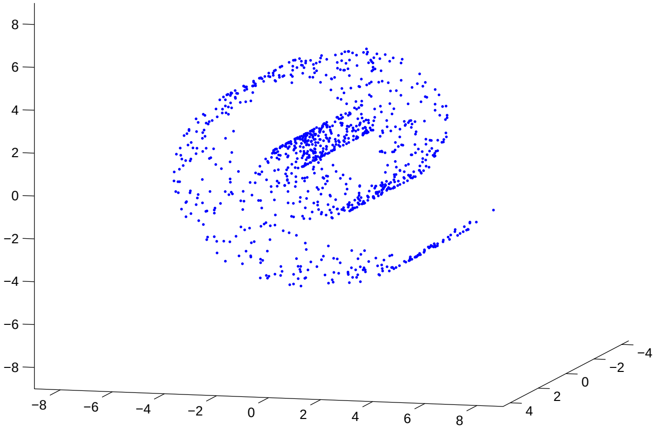
August 17, 2006

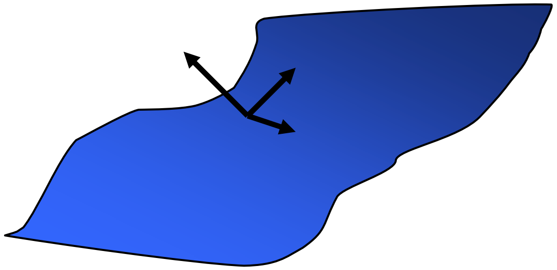
Joint work with Ali Rahimi (Intel)

What “Manifold Learning” Isn’t

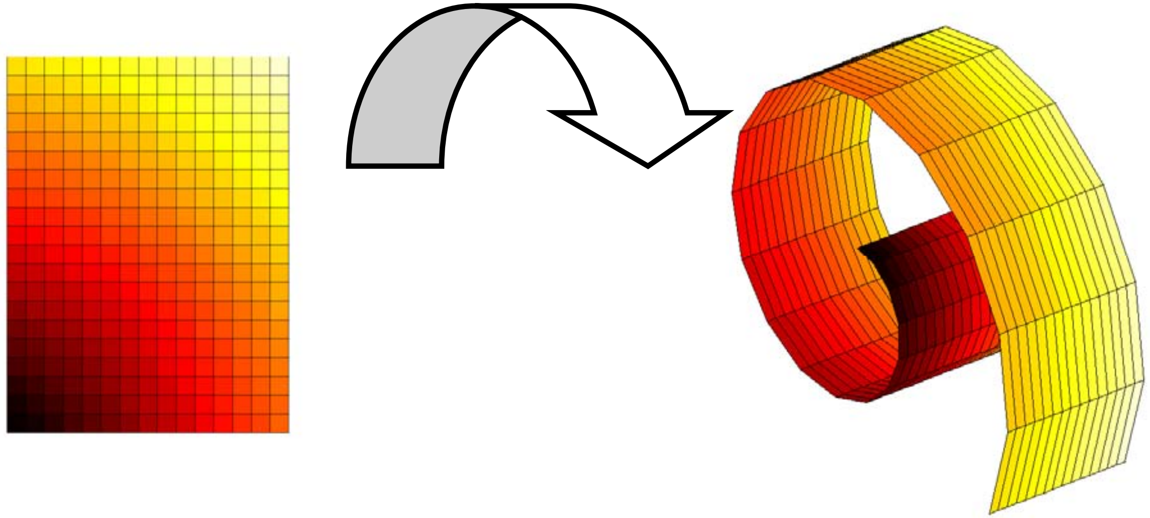
- Common features of Manifold Learning Algorithms:
  - 1-1 charting
  - Dense sampling
  - Geometric Assumptions







What Manifold Learning might be…



- Sample data in a low dimensional space
- Pass each data point through the same nonlinearity
- How to recover the data?

Probabilistic Model

- Data $x_1,\ldots,x_n$ in $\mathbb{R}^d$ sampled from a joint distribution p(X)
- Each x is passed through a nonlinear function

  $f\colon \mathbb{R}^d \rightarrow \mathbb{R}^D$. $y_i=f(x_i)$

- The distribution for Y is given by

$$p_{\mathbf{Y}}(\mathbf{Y};f) = p_{\mathbf{X}}(f^{-1}(y_1),\ldots,f^{-1}(y_N)) \\ \times \prod_{i=1}^{N} \mathsf{det}\left(\nabla f(f^{-1}(y_i))\nabla f(f^{-1}(y_i))'\right)^{-1/2}$$

Diffeomorphic Warping

- If we assume that f is a diffeomorphism, there exists an inverse function in the neighborhood of the image such that g(f(x))=x and ∇g∇f = I for all x.

$$
\begin{aligned}
p_{\mathbf{Y}}(\mathbf{Y}; f) &= p_{\mathbf{X}}(f^{-1}(y_1), \ldots, f^{-1}(y_N)) \\
&\qquad \times \prod_{i=1}^{N} \operatorname{det}\left(\nabla f(f^{-1}(y_i))\nabla f(f^{-1}(y_i))'\right)^{-1/2} \\
&= p_{\mathbf{X}}(g(y_1), \ldots, g(y_N)) \prod_{i=1}^{N} \operatorname{det}\left(\nabla g(y_i)'\nabla g(y_i)\right)^{-1/2}
\end{aligned}
$$

Diffeomorphic Warping

- If we assume that f is a diffeomorphism, there exists an inverse function in the neighborhood of the image such that g(f(x))=x and ∇g∇f = I for all x.
- Taking a logarithm, we may search for the maximum likelihood g

$$\max_g \log p_{\mathbf{X}}(g(y_1), \ldots, g(y_N)) + \frac{1}{2} \sum_{i=1}^{N} \log \det \left(Dg(y_i)' Dg(y_i) \right)$$

Benefits of this Perspective

- Asymptotic Convergence
- Out of Sample Extension
- No neighborhood estimates
- Incorporates Prior Knowledge
- Easy to make “semi-supervised”

Asymptotic Convergence

- If y_i is sampled iid,

$$\frac{1}{N}\sum_{i=1}^{N} \log p_y(y_i; g) \rightarrow \int_y p_y(y) \log p_y(y; g)$$

- Which is minimized when $p_y(y;g)=p_y$
- Similarly, if joint distribution is stationary and ergodic sequence and k-th order Markov, log p_Y converges to the cross entropy (Shannon-McMillian-Breiman Theorem)

Diffeomorphic Warping

$$\max_{g} \log p_{\mathbf{X}}(g(y_1),\ldots,g(y_N)) + \frac{1}{2}\sum_{i=1}^{N} \log\det\left(\nabla g(y_i)'\nabla g(y_i)\right)$$

Ingredients	
• Set of functions	• **RKHS**
• Prior on X	• **Dynamics**
• Optimization Tools	• **Duality**

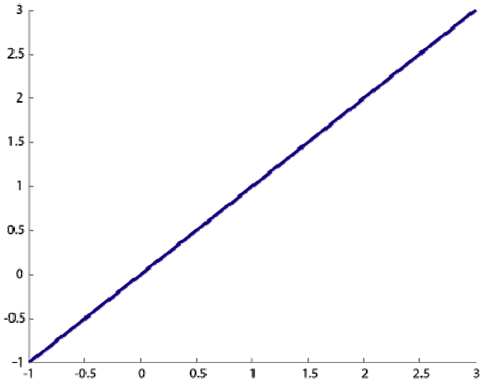
Kernels

- **k** be a function of two variables which is *positive definite*

$$\sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0$$

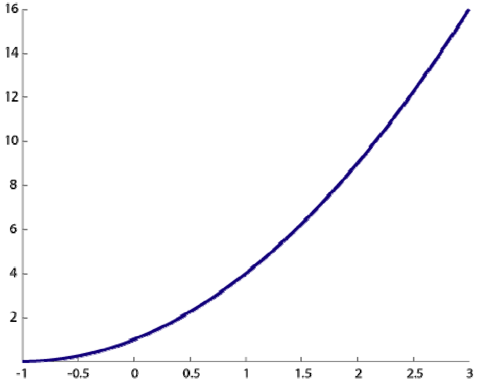
for all $\mathbf{x}_i$ and c_i. Such a function is called a *positive definite kernel.*

Examples of Kernels



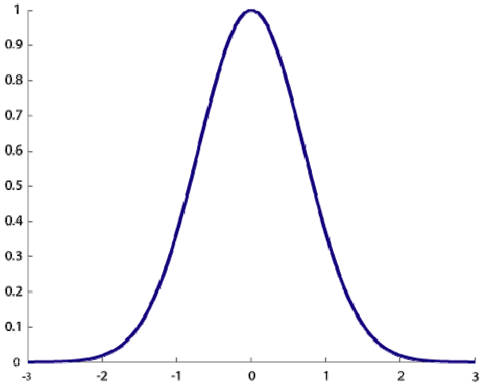
Linear

$$k(x_1, x_2) = x_1^\top x_2$$



Polynomial

$$k(x_1, x_2) = (1 + x_1^\top x_2)^d$$



Gaussian

$$k(x_1, x_2) = \exp(-C\|x_1 - x_2\|^2)$$

Aside: Checking Positivity

- When is a symmetric function a kernel?
- Polynomials – trivial to check (positive definite form on monomials)
- Gaussians – Fourier transform of positive function
- Sum and Mixtures
- Pointwise Products (Schur Product theorem)
- What else? And what algorithmic tools can we develop to check whether a kernel is positive?

Reproducing Kernel Hilbert Spaces

- **Theorem:** If X is a compact set and $\mathcal{H}$ is a Hilbert space of functions from X to $\mathbb{R}$. Then all functionals

$$\delta_{\mathbf{x}}(f) = f(\mathbf{x})$$

are bounded iff there is a unique positive definite kernel k($\mathbf{x}_1$, $\mathbf{x}_2$) such that for all f∈ $\mathcal{H}$ and **x**∈ X

$$\langle k(\mathbf{x}, \cdot), f \rangle = f(\mathbf{x})$$

k is called the *reproducing kernel* of $\mathcal{H}$.

RKHS (converse)

- If $k(x_1,x_2)$ is a positive definite kernel on X, consider the set of functions

$$\mathcal{F} = \{\phi_{\mathbf{x}}(\mathbf{y}) := k(\mathbf{x},\mathbf{y})\}$$

- And define the inner product $\langle \phi_{\mathbf{x}}, \phi_{\mathbf{y}} \rangle = k(\mathbf{x},\mathbf{y})$

- Then the span of $\mathcal{F}$ is an inner product space and its completion is an RKHS

Properties of RKHS

- Let $\alpha = \sum_i c_i \mathbf{k}(\mathbf{x}_i, \cdot)$

$$\|\alpha\|_K^2 = \langle \alpha, \alpha \rangle = \sum_{i,j} c_i c_j \langle \mathbf{k}(\mathbf{x}_i, \cdot), \mathbf{k}(\mathbf{x}_j, \cdot) \rangle = \mathbf{c}^\top \mathbf{K} \mathbf{c}$$

where $\mathbf{K}_{ij} = \mathbf{k}(\mathbf{x}_i, \mathbf{x}_j)$

- If $\mathbf{x} \in X$ $\quad \langle \alpha, \mathbf{k}(\mathbf{x}, \cdot) \rangle = \sum_i c_i \mathbf{k}(\mathbf{x}_i, \mathbf{x}) = \alpha(\mathbf{x})$

Duality and RKHS

$$\begin{array}{ll} \text{min}_{f \in \mathcal{H}} & \|f\|^2 \\ \text{s.t.} & \sum_{j=1}^{N} a_{ij} f(\mathbf{x}_j) \leq b_i \end{array}$$

$$f^*(\mathbf{x}) = \sum_{i,j=1}^{N} \lambda_i a_{ij} k(\mathbf{x}_j, \mathbf{x})$$

- λ_i is the Largrange multiplier associated with constraint i
- No duality gap

Duality and RKHS

$$\begin{aligned}\mathcal{L} &= \|f\|^2 - 2\sum_i \lambda_i \left(\sum_{j=1}^{N} a_{ij} f(\mathbf{x}_j) - b_i\right)\\ &= \langle f, f\rangle - 2\sum_i \lambda_i \left(\sum_{j=1}^{N} a_{ij} \langle f, \phi_{\mathbf{x}_j}\rangle - b_i\right)\\ &= \langle f, f\rangle - 2\langle \sum_{i,j} f, \lambda_i a_{ij} \phi_{\mathbf{x}_j}\rangle - \sum_i \lambda_i b_i\end{aligned}$$

Dual: $\max_{\lambda} -\lambda' \mathbf{Q} \lambda + \mathbf{b}^\top \lambda$

Duality and RKHS

$$\begin{array}{ll} \mathsf{min}_{f\in\mathcal{H}} & \|f\|^2 \\ \mathsf{s.t.} & \sum_{j=1}^{N} a_{ij} f(\mathbf{x}_j) \leq b_i \end{array}$$

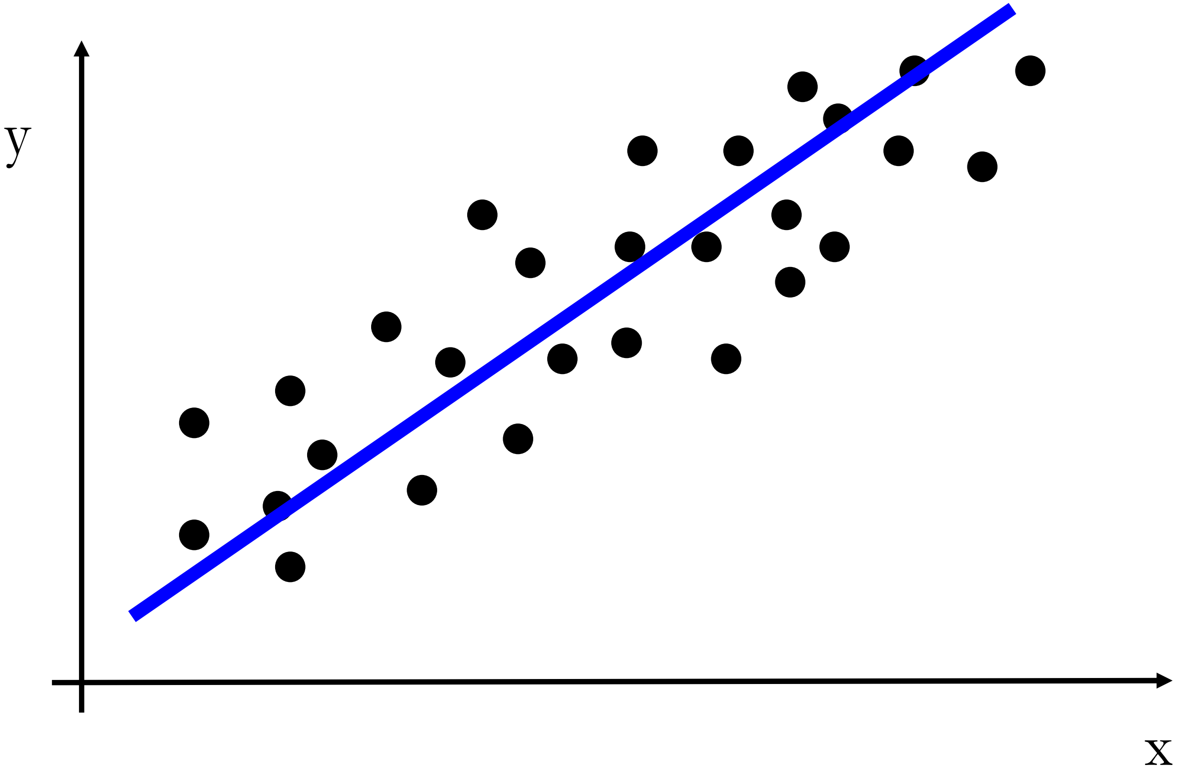
$$f^*(\mathbf{x}) = \sum_{i,j=1}^{N} \lambda_i a_{ij} k(\mathbf{x}_j, \mathbf{x})$$

$$\max_{\lambda} -\lambda' \mathbf{Q} \lambda + \mathbf{b}^\top \lambda$$

- Optimizations involving norm of f and f on data admit finite representation
- Nonlinearity of kernel gives nonlinear functions
- Extends to gradients of f
- Hugely successful in approximation and learning

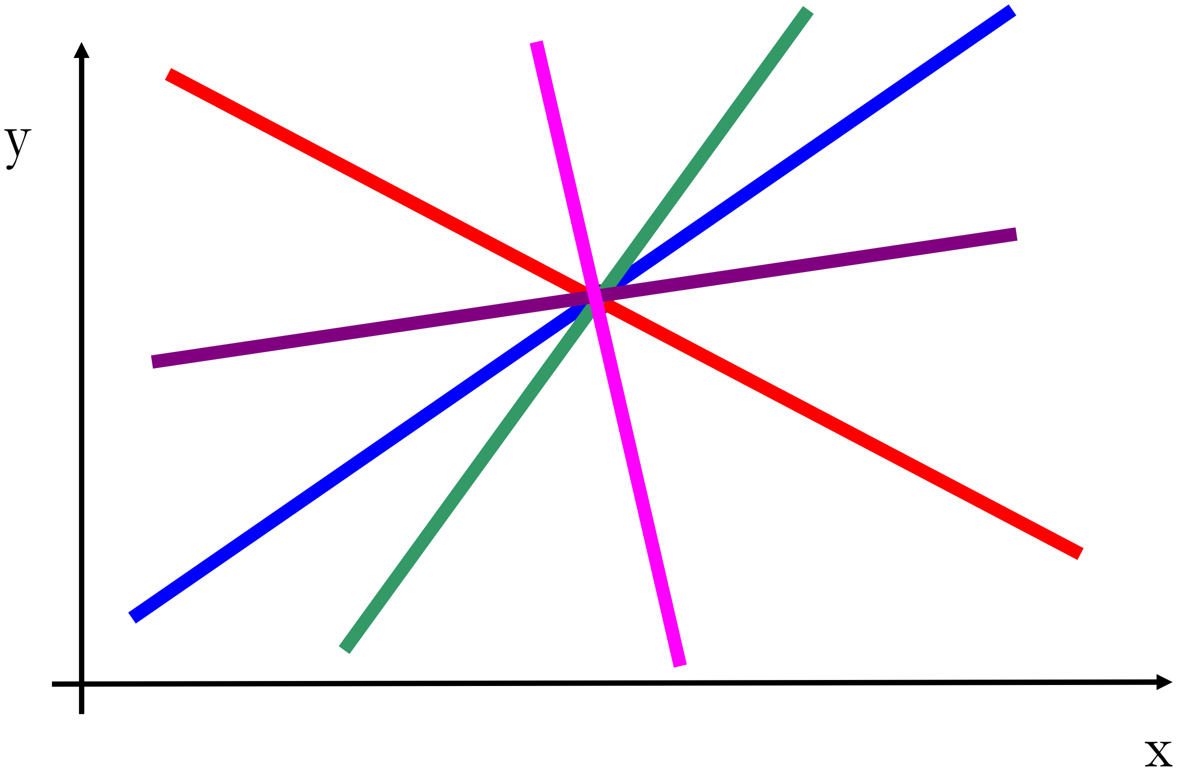
Linear Regression

- Find best linear model agreeing with data



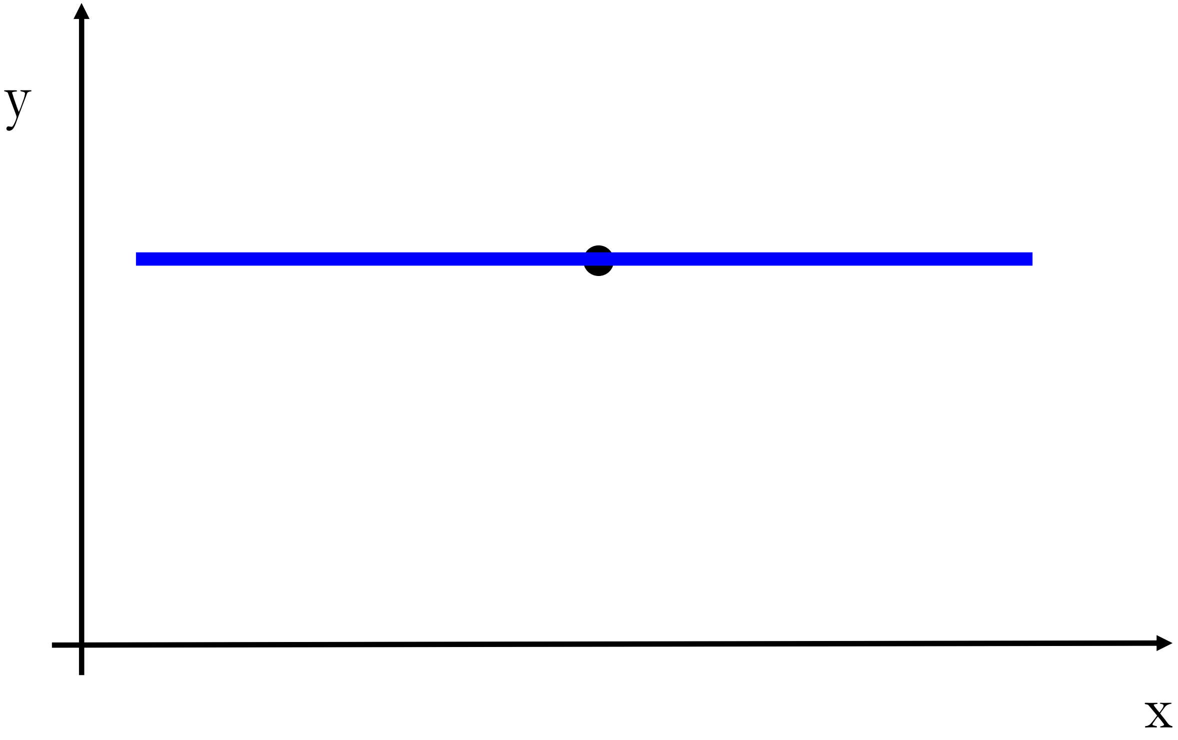
Linear Regression

- Find best linear model agreeing with data



Linear Regression

- Find best linear model agreeing with data



Linear Regression

- Find best linear model agreeing with data

$$\min_{\mathbf{w},d} \sum_{i=1}^{L} (\mathbf{w}^\top \mathbf{x}_i + d - y_i)^2 + \lambda \|\mathbf{w}\|_K^2$$

Agree with data

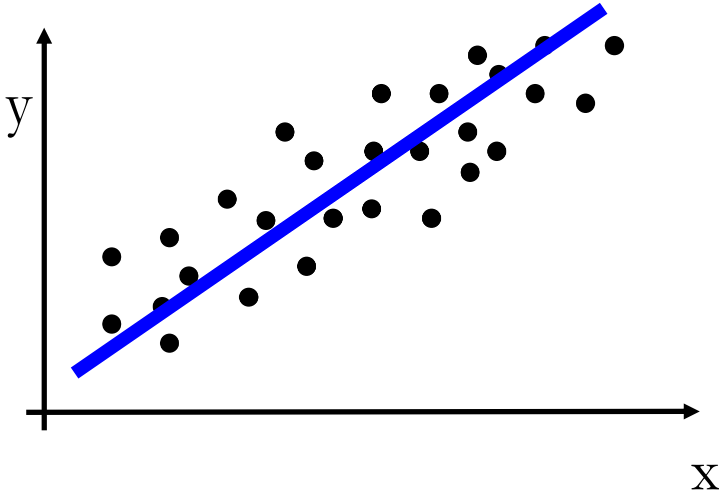
Smoothness/Complexity

- Linear f. Euclidean Norm. Solution:

$$f(\mathbf{x}) = \sum_{i=1}^{L} c_i (\mathbf{x}_i^\top \mathbf{x}) + d$$

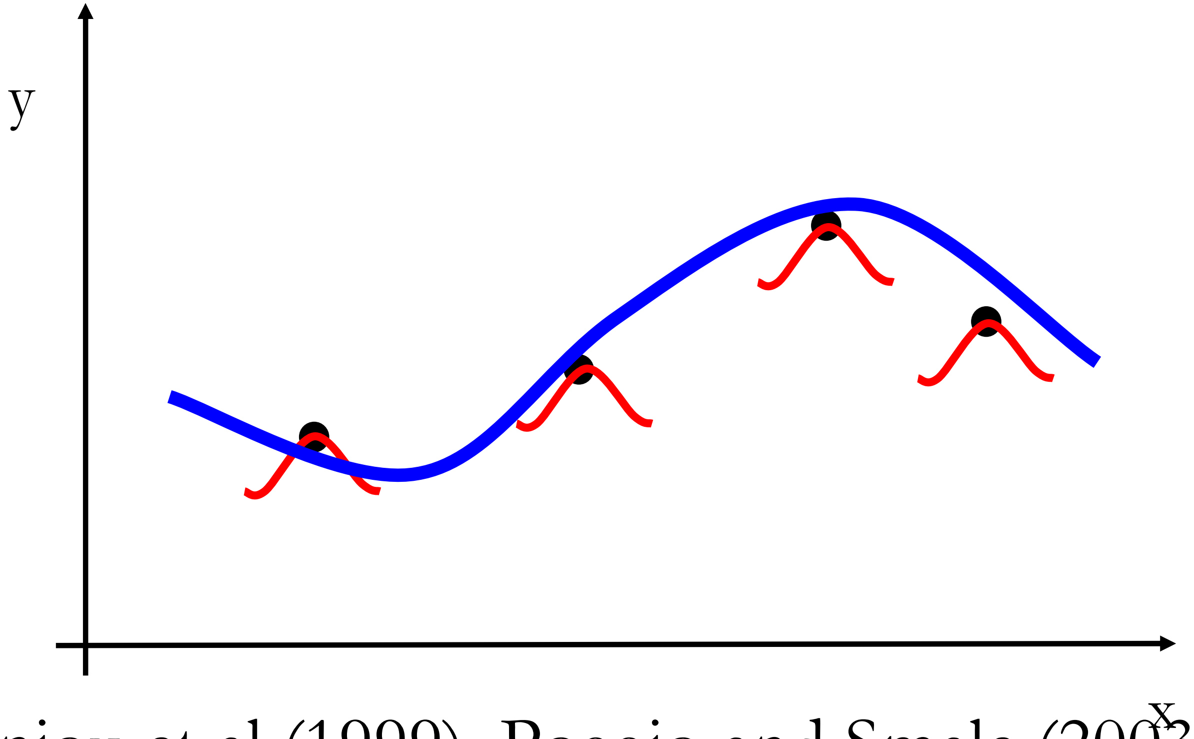
Linear Regression: Morals

- Can be solved with least-squares.
- The solution is a linear combination of the data.
- Computing f(**x**) only involved inner products of the data.
- How about nonlinear models?



Regression (nonlinear)

- Search over an RKHS



- Evgeniou et al (1999), Poggio and Smale (2003)

Nonlinear Regression

- Find best model agreeing with data

$$\min_f \sum_{i=1}^{L} (f(\mathbf{x}_i) - y_i)^2 + \lambda \|f\|_K^2$$

Agree with data

Smoothness/Complexity

- f ∈ RKHS, RKHS norm. Solution:

$$f(\mathbf{x}) = \sum_{i=1}^{L} c_i \mathbf{k}(\mathbf{x}_i, \mathbf{x})$$

Nonlinear Regression

- Find best model agreeing with data.

$$\min_{f} \sum_{i=1}^{L} (f(\mathbf{x}_i) - y_i)^2 + \lambda \|f\|_K^2$$

Agree with data

Smoothness/Complexity

Linear

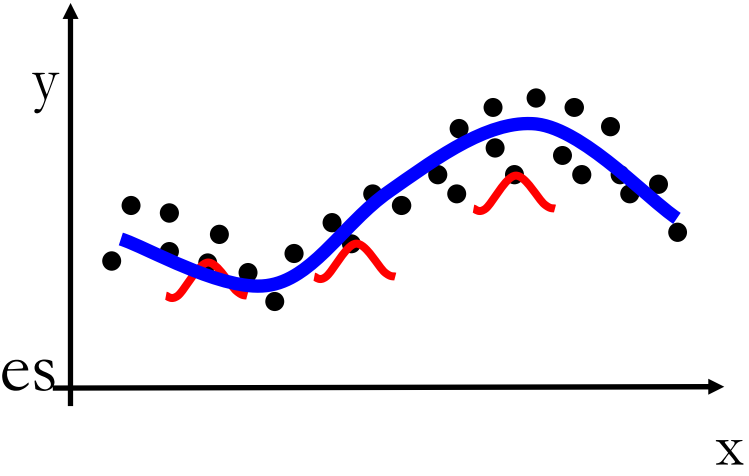
$$f(\mathbf{x}) = \sum_{i=1}^{L} c_i(\mathbf{x}_i^\top \mathbf{x})$$

Nonlinear

$$f(\mathbf{x}) = \sum_{i=1}^{L} c_i \mathbf{k}(\mathbf{x}_i, \mathbf{x})$$

Nonlinear Regression: Morals

- Can be solved with least-squares.
- The solution is a linear combination of kernels centered at the data.
- Computing f(**x**) only involves kernel products of the data.
- RKHS often dense in L_2.



Generalization and Stability

- Regularizing with the norm makes algorithms ***robust*** to changes in the training data
- Models ***generalize*** if they predict novel data as well as they predict on the training data

- **Theorem:** A model generalizes if and only if it is robust to changes in the data [Poggio et al 2004].
- The RKHS norm is meaningful: penalizes complexity.

Diffeomorphic Warping with RKHS

$$\max_g \log p_{\mathbf{X}}(g(y_1), \ldots, g(y_N)) + \frac{1}{2} \sum_{i=1}^{N} \log\det\left(\nabla g(y_i)' \nabla g(y_i)\right) + \lambda_r \|g\|^2$$

- We know: $g(y) = \sum_{i=1}^{N} c_i k(y_i, y) + a_i^\top \nabla k(y_i, y)$
- But log det is not convex in (a,c)
- Construct a dual problem as approximation. If p_X is a zero-mean gaussian, we get a determinant maximization problem [Vandenberghe et al, 1998]

$$\min_{\mathbf{S} \succeq 0} \mathrm{Tr}(\Omega \mathbf{S}) - \sum_{k=1}^{N} \log(\mathrm{Tr}(\mathbf{J}_k \mathbf{S}))$$

Eigenvalue Approximation

$$\min_{\mathbf{S}\succeq 0} \mathsf{Tr}(\Omega\mathbf{S}) - \sum_{k=1}^{N} \log(\mathsf{Tr}(\mathbf{J}_k\mathbf{S}))$$

- Ω^{-1} is an optimal solution if and only if

$$\Omega - \sum_{k=1}^{N} \frac{1}{\mathsf{Tr}(\mathbf{J}_k\Omega^{-1})}\mathbf{J}_k \succeq 0$$

- This follows from KKT conditions of MAXDET
- The eigenvalues of Ω give coefficients for the expansion of g

Remarks

- Dual can be solved using an interior point method
- Provides a lower bound on the log-likelihood
- We can approximate with a spectral method
- Easy to extend to any log-concave prior on X
- Performs quite well in experiments

Diffeomorphic Warping

$$\max_{g} \log p_{\mathbf{X}}(g(y_1),\ldots,g(y_N)) + \frac{1}{2}\sum_{i=1}^{N} \log\det\left(\nabla g(y_i)'\nabla g(y_i)\right)$$

Ingredients

• Set of functions	• **RKHS**
• Prior on X	• **Dynamics**
• Optimization Tools	• **Duality**

Dynamics

$$\mathbf{s}[t+1] = \mathbf{A}\mathbf{s}[t] + \omega[t]$$
$$\mathbf{x}[t] = \mathbf{C}\mathbf{s}[t] + \nu[t]$$
$$\mathbb{E}[\omega[t]\omega[t]'] = \Lambda_\omega$$
$$\mathbb{E}[\nu[t]\nu[t]'] = \Lambda_\nu$$

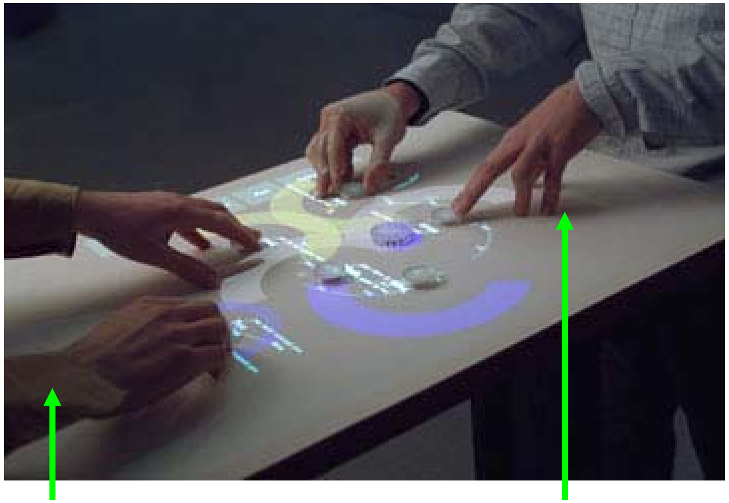
Assume data is generated by an LTIG system

$$\mathbf{C} = \left[\begin{array}{ccc} 1 & 0 & 0 \end{array}\right]$$

$$\mathbf{A} = \left[\begin{array}{ccc} 1 & \delta & 0 \\ 0 & 1 & \delta \\ 0 & 0 & 1 \end{array}\right]$$

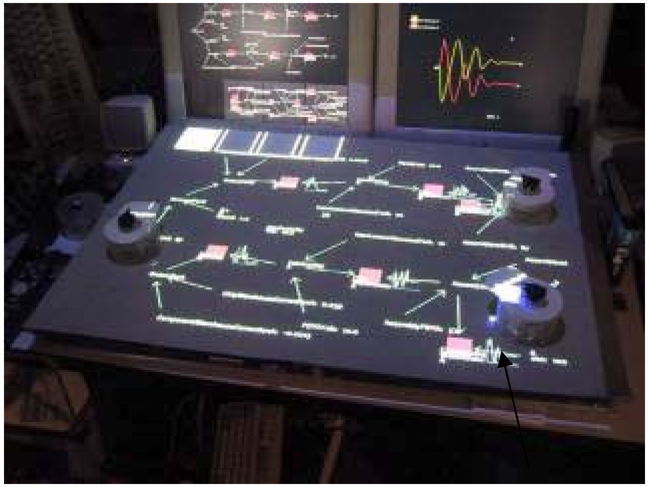
For the experiments, this model can be very dumb!

The *Sensetable*

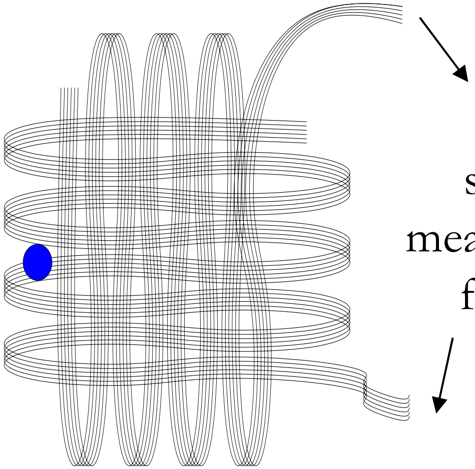


James Patten

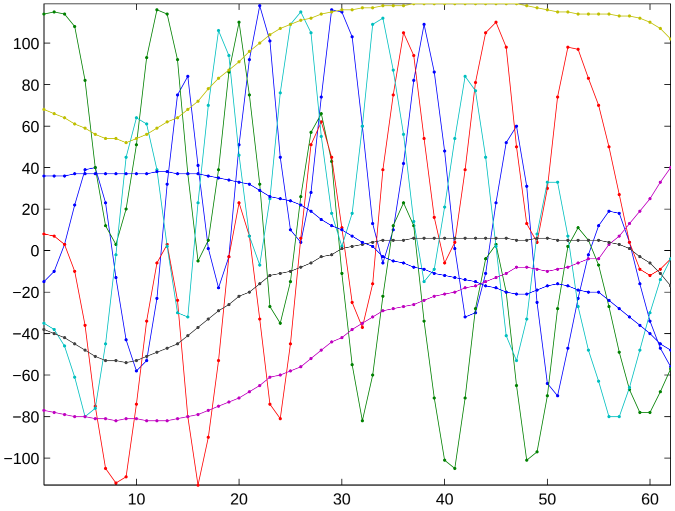
Me



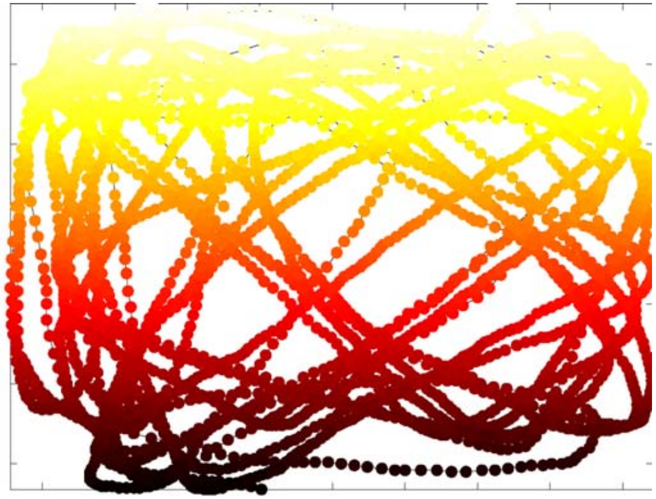
RFID tag



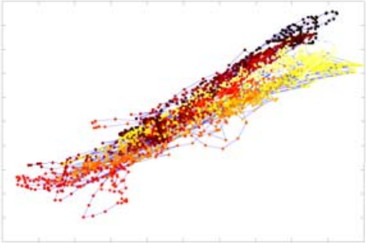
Signal strength measurements from tag



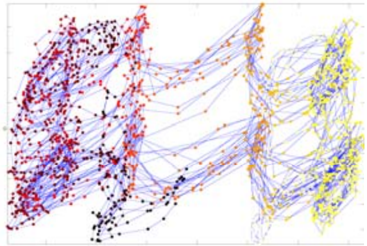
Sensetable: Manifold Learning



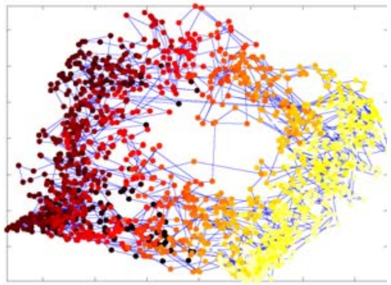
Ground Truth



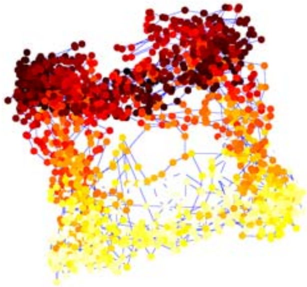
LLE



KPCA

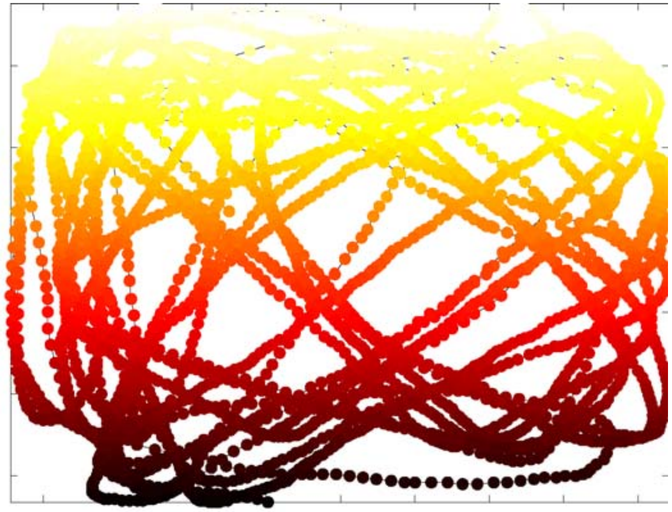


Isomap

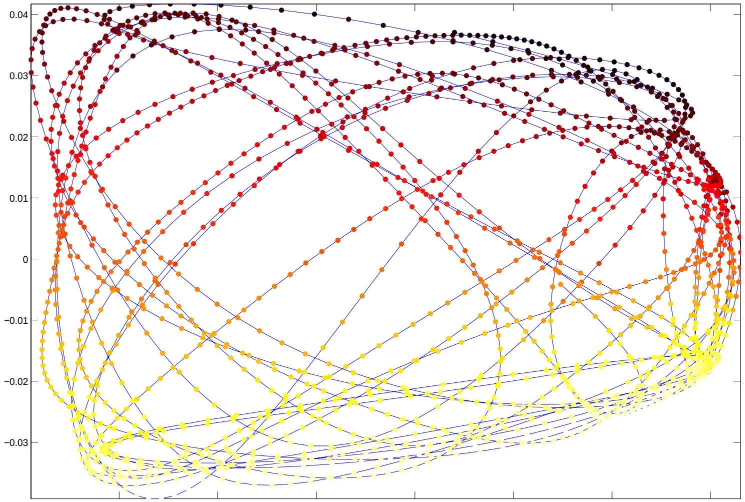


ST-Isomap

Sensetable: DW

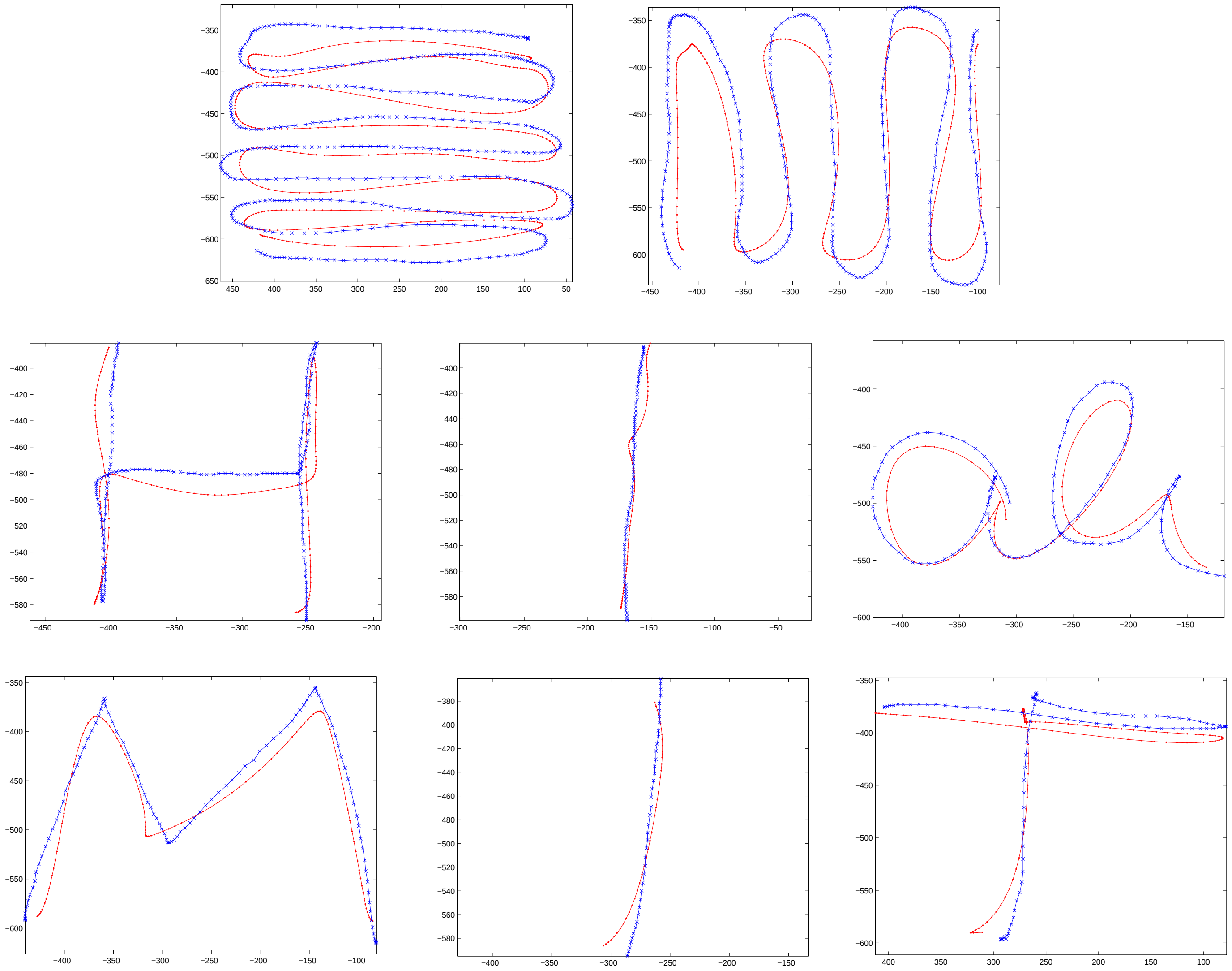


Ground Truth



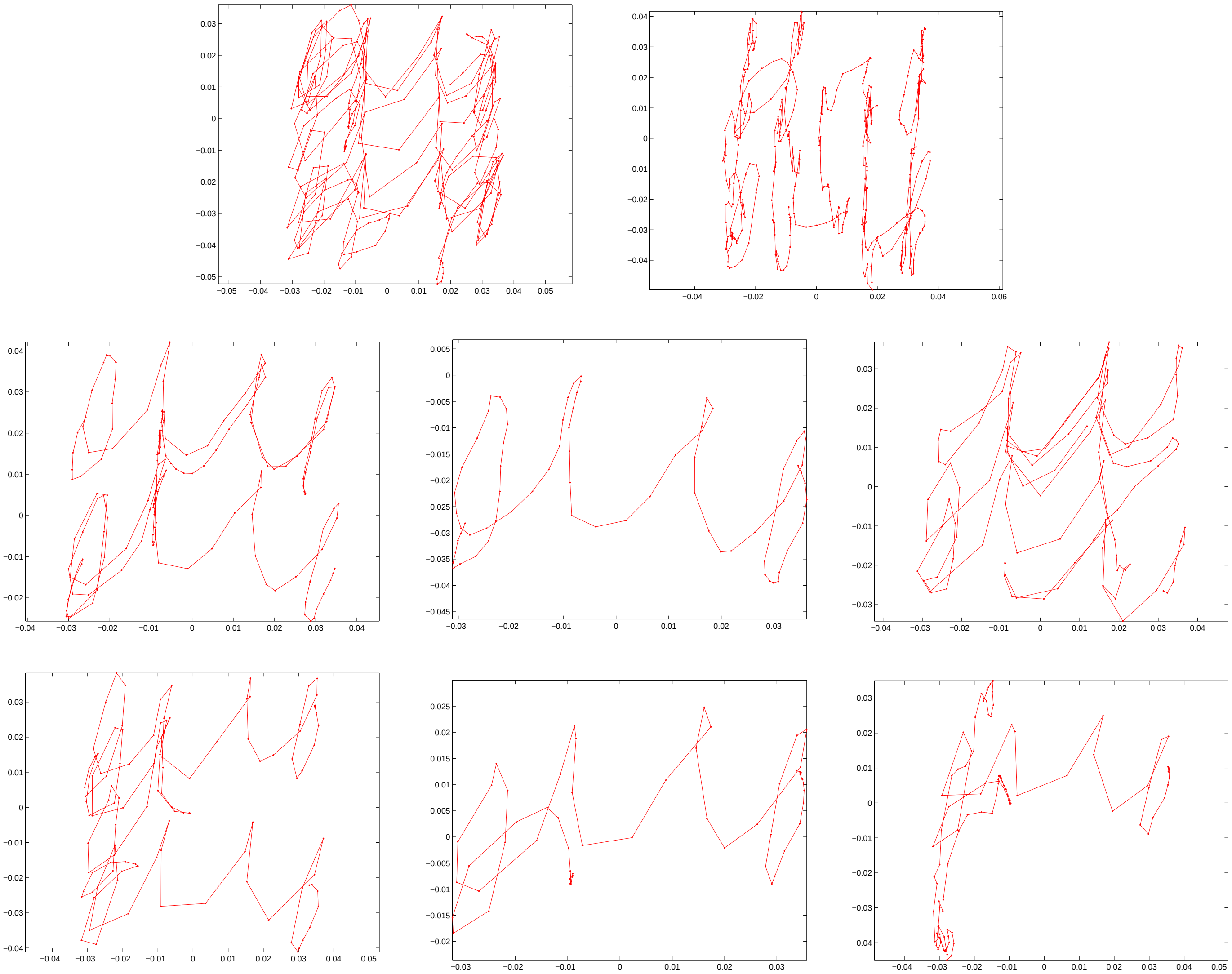
Diffeomorphic Warping

Tracking



-350
-400
-450
-500
-550
-600
-650
-450 -400 -350 -300 -250 -200 -150 -100 -50
-350
-400
-450
-500
-550
-600
-450 -400 -350 -300 -250 -200 -150 -100
-400
-420
-440
-460
-480
-500
-520
-540
-560
-580
-450 -400 -350 -300 -250 -200
-400
-420
-440
-460
-480
-500
-520
-540
-560
-580
-300 -250 -200 -150 -100 -50
-400
-450
-500
-550
-600
-400 -350 -300 -250 -200 -150
-350
-400
-450
-500
-550
-600
-400 -350 -300 -250 -200 -150 -100
-380
-400
-420
-440
-460
-480
-500
-520
-540
-560
-580
-400 -350 -300 -250 -200 -150
-350
-400
-450
-500
-550
-600
-400 -350 -300 -250 -200 -150 -100

Tracking with KPCA



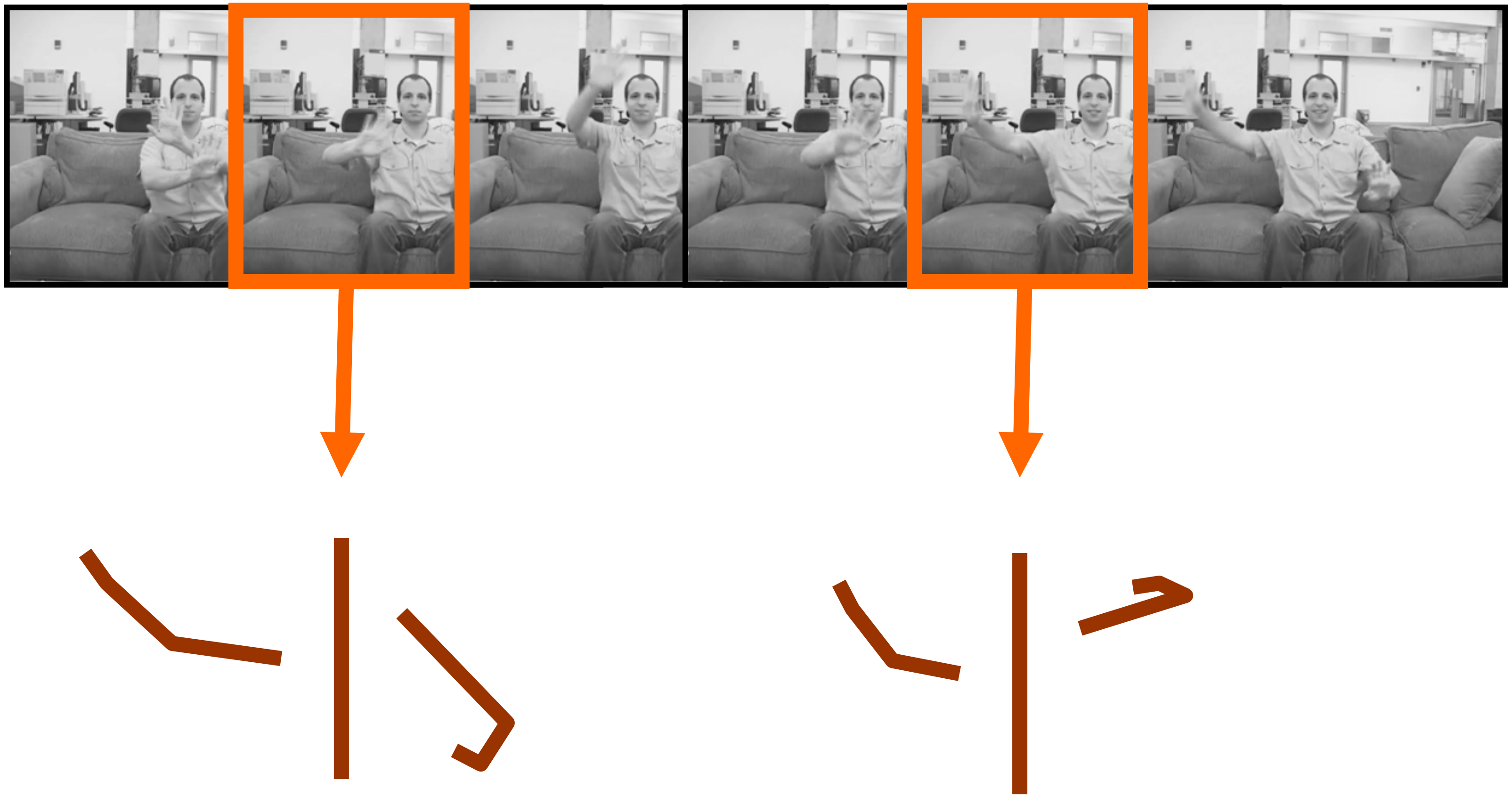
Video

Representation

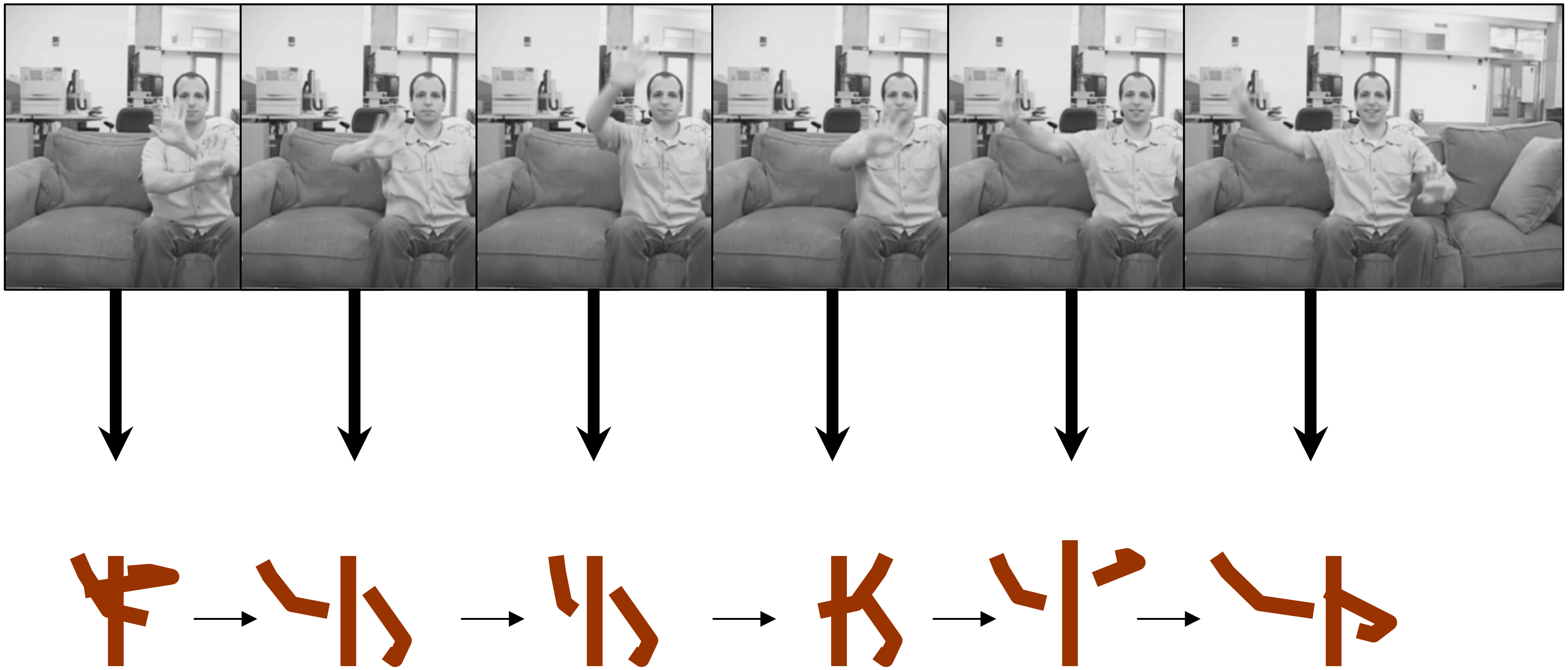
- Big mess of numbers for each frame

$$\longrightarrow \begin{bmatrix} \vdots \\ 43 \\ 76 \\ 121 \\ 147 \\ 158 \\ 170 \\ 172 \\ 168 \\ 169 \\ 176 \\ \vdots \end{bmatrix}$$

- Raw pixels, no image processing



Annotations from user or detection algorithms



Assume that output time series is smooth.

Video

Future Work

- Speeding up the log-det
- Optimizing over families of priors p_X
- Estimating the duality gap

- Learning manifolds that need more than one chart
- Understanding why nonparametric ID is easy while parametric ID is hard