KDE-HMMs

New, Nonparametric Acoustic Models for Speech Synthesis

Gustav Eje Henter
Joint work with W. Bastiaan Kleijn and Arne Leijon at KTH

CSTR internal presentation

Monday 20 January, 2014

Take-Home Message

Current acoustic models in parametric speech synthesis are not a good fit

We present a new acoustic model for speech that

1. Converges asymptotically on the true data-generating process
2. Can be interpreted as probabilistic hybrid speech synthesis
3. Models nonlinear time series better

The advantages come thanks to nonparametric speech synthesis

Outline

1. Introduction
2. Kernel density estimation
3. KDE Markov models
   - Experiments
4. KDE-HMMs
   - Parameter estimation
   - Experiments
5. Summary and outlook

Outline

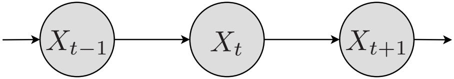
1. Introduction
2. Kernel density estimation
3. KDE Markov models
   - Experiments
4. KDE-HMMs
   - Parameter estimation
   - Experiments
5. Summary and outlook

Standard Sequence Models

Markovian paradigm

- Finite-length memory
- Examples:
  - Discrete Markov chain $p_{X_t \mid X_{t-1}}(x_t \mid x_{t-1})$
  - Linear autoregressive (AR) models

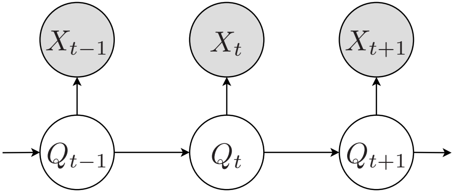
$$X_t = \mu + \sum_{l=1}^{p} \alpha_l (x_{t-l} - \mu) + \mathcal{E}_t$$



Standard Sequence Models

Hidden-state paradigm

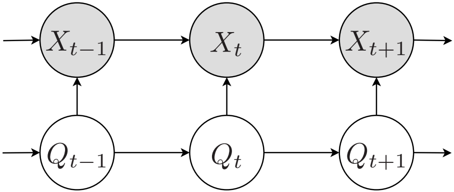
- Unbounded memory
- Admits a control signal
- Examples:
  - Hidden Markov model (discrete state Q_t)
  - Kalman filter (continuous state)



Standard HMM Acoustic Model

Standard models for parametric speech synthesis are HMMs or HSMMs

- States Q_t represent (sub)phone, context, and prosodic information
- Observables $\boldsymbol{X}_t \in \mathbb{R}^D$ are vocoder parameters
- State-conditional output distributions $f_{\boldsymbol{X}_t|Q_t}(\boldsymbol{x}_t \mid q_t)$ are Gaussian
- Dynamic features (Δs and $\Delta\Delta$s) tie adjacent observations together
  - Autoregressive HMMs (AR-HMMs) less mathematically objectionable



Problems

Even using ground-truth durations, generated features are poor

- Sampled output is warbly (Shannon, Zen, & Byrne, 2011)
- Most probable output sequence (ML parameter generation, MLPG) sounds muffled and buzzy

Note: Unit selection does not have these problems

Problem Analysis

What is wrong with our parametric models?

- The model is inadequate
  - State-conditional outputs are overly simplistic—essentially just linear AR processes
  - Results on full-covariance models from Shannon, Zen, & Byrne (2011) suggest that trajectory time dependence is not well modelled
- Nonlinear AR models are a closer match
  - Product of experts increase held-out data likelihood substantially, but not synthesis quality (Shannon, 2012)

New Idea

What to do?

- No one knows what the "true" distribution f of speech is
- It is not obvious how to improve current models
- This calls for a generally applicable technique!
- Proposal: Kernel Conditional Density Estimation + Markov processes
  - Can describe any Markov model
  - Then add hidden state to control process output

Outline

1. Introduction
2. Kernel density estimation
3. KDE Markov models
   - Experiments
4. KDE-HMMs
   - Parameter estimation
   - Experiments
5. Summary and outlook

Kernel Density Estimation

Kernel Density Estimation (KDE) is a nonparametric density estimation technique

- Training data $\mathcal{D} = \{\boldsymbol{y}_1, \ldots, \mathbf{y}_N\}$ in $\mathbb{R}^D$ sampled from reference $f_{\boldsymbol{X}}$
- Test points $\{\boldsymbol{x}_1, \ldots, \mathbf{x}_T\}$
- KDE can be seen as a smoothing or blurring (convolution) of the empirical density function

$$\dot{f}_{\boldsymbol{X}}(\boldsymbol{x} \mid \mathcal{D}) = \frac{1}{N} \sum_{n=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{y}_n)$$

with a nonnegative kernel function $k(\boldsymbol{r})$

- Intuition: KDE is to squint while looking at the datapoints

Kernel Density Estimation

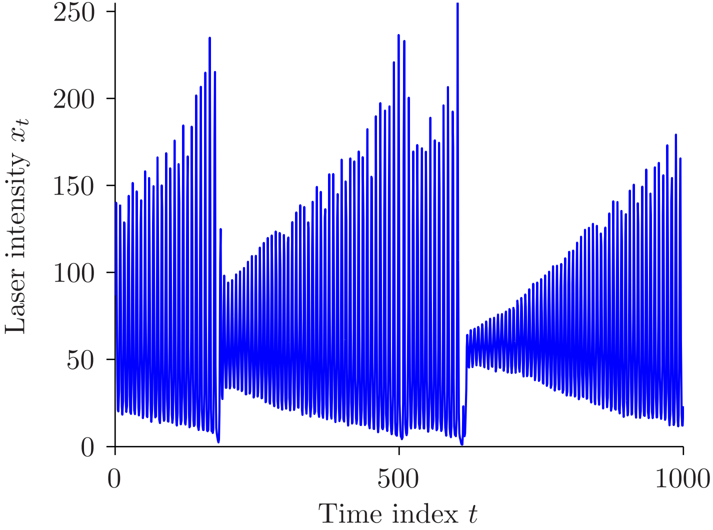
- The estimated PDF can be written

$$\widehat{f}_{\boldsymbol{X}}\left(\boldsymbol{x} \mid \mathcal{D},\, h\right) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^{D}}k\left(\frac{\boldsymbol{x}-\boldsymbol{y}_n}{h}\right)$$

  where h is a bandwidth parameter controlling the degree of smoothing
- We require $\int_{\boldsymbol{r}} k\left(\boldsymbol{r}\right)\mathrm{d}\boldsymbol{r} = 1$ and $\int_{\boldsymbol{r}} \boldsymbol{r}k\left(\boldsymbol{r}\right)\mathrm{d}\boldsymbol{r} = \mathbf{0}$
- Probabilistic interpretation:
  - Mixture distribution with $k\left(\boldsymbol{r}\right)$-shaped zero-mean components
  - One component centered on each training-data point
- We use Gaussian kernels throughout
  - Bandwidth h matters more than kernel shape $k\left(\boldsymbol{r}\right)$

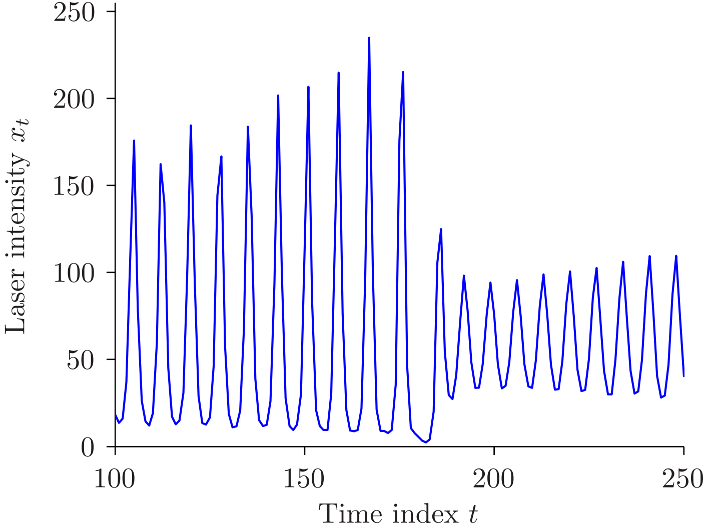
Example Data

Running example: Santa Fe chaotic FIR laser series (1D, $N = 1000$ plotted)



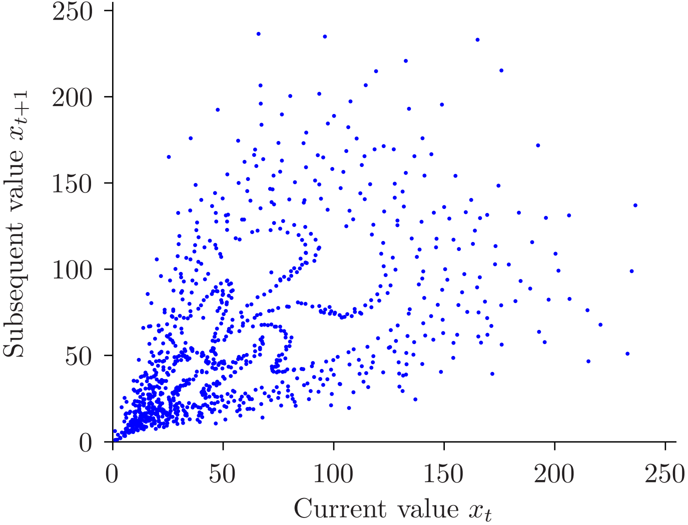
Example Data

Running example: Santa Fe chaotic FIR laser series (detail)



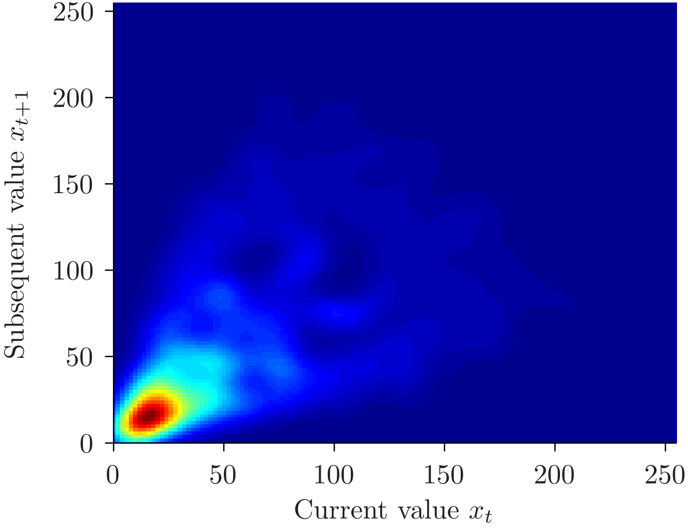
Example Data

Scatter plot of consecutive values $\{(x_t, x_{t+1})\}_t$ reveals attractor structure



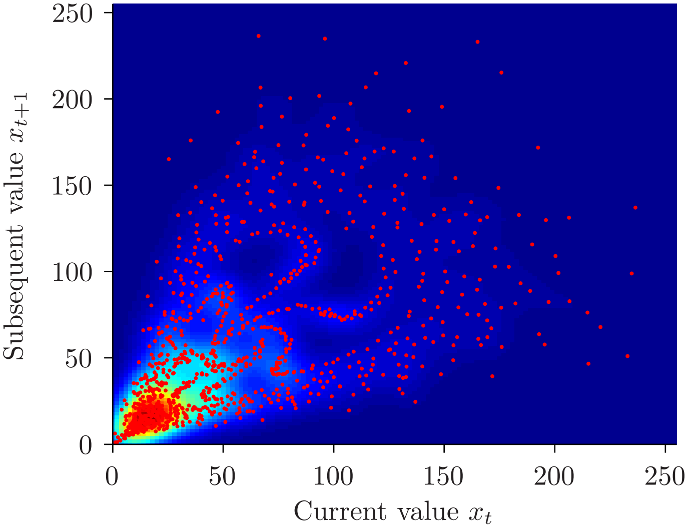
Example KDE

Gaussian blur of points = 2D KDE (bandwidth $\widehat{h}$ optimised for log-prob)



Example KDE

Scatter plot superimposed on 2D KDE fit



KDE Properties

Strengths:

- Asymptotically consistent: $\lim_{N\to\infty} \widehat{f}_{\boldsymbol{X}} = f_{\boldsymbol{X}}$ under appropriate bandwidth selection ($h \to 0$, $Nh \to \infty$), *regardless of* $f_{\boldsymbol{X}}$
- Built from data points (nonparametric)
- Single free parameter

Weaknesses:

- Data demanding
- Computationally demanding
  - Substantial speedups are possible (e.g., Holmes, Gray, & Isbell, 2007)

Outline

1. Introduction
2. Kernel density estimation
3. KDE Markov models
   - Experiments
4. KDE-HMMs
   - Parameter estimation
   - Experiments
5. Summary and outlook

Handling Time Dependence

So far we have said nothing about time dependence

- Key idea: A joint KDE PDF $\widehat{f}_{\underline{\boldsymbol{X}}_{t-p}^{t}}\left(\underline{\boldsymbol{x}}_{t-p}^{t}\right)$ for sequence segments

$$\underline{\boldsymbol{x}}_{t-p}^{t} = \left[\boldsymbol{x}_{t-p}^{\intercal},\, \ldots,\, \boldsymbol{x}_{t-1}^{\intercal},\, \boldsymbol{x}_{t}^{\intercal}\right]^{\intercal}$$

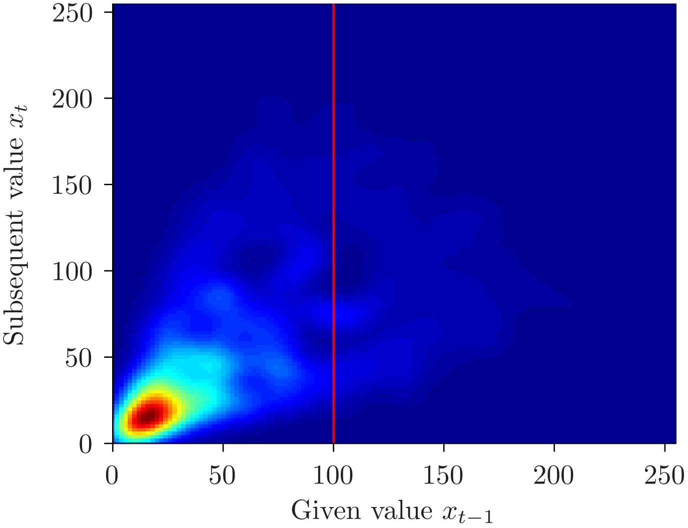
  induces a conditional distribution $\widehat{f}_{\boldsymbol{X}_t \mid \underline{\boldsymbol{X}}_{t-p}^{t-1}}\left(\boldsymbol{x}_t \mid \underline{\boldsymbol{x}}_{t-p}^{t-1}\right)$

  - Hyndman, Bashtannyk, & Grunwald (1996)

- These next-step distributions are sufficient to define a p-order Markov process
  - KDE Markov model (KDE-MM)
  - Nonlinear and nonparametric
  - Many independent proposals, e.g., Rajarshi (1990)

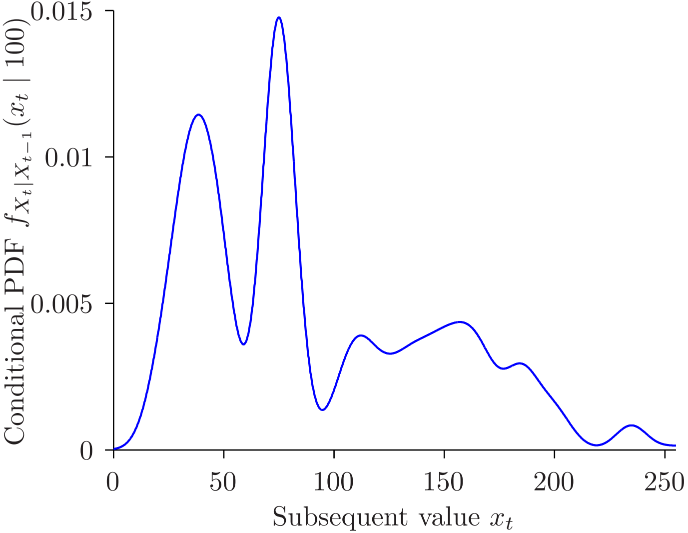
Graphical Illustration

A conditional distribution is a cut through the KDE



Graphical Illustration

Resulting normalised next-step PDF $\widehat{f}_{X_t|X_{t-1}}(x \mid x_{t-1} = 100)$



KCDE Definition

Kernel Conditional Density Estimation (KCDE) is a normalisation of the KDE, with resulting PDF

$$\widehat{f}_{\boldsymbol{X}_t \mid \underline{\boldsymbol{X}}_{t-p}^{t-1}}\left(\boldsymbol{x}_t \mid \underline{\boldsymbol{x}}_{t-p}^{t-1}, \mathcal{D}\right) = \frac{1}{h^D} \frac{\sum_n \prod_{l=0}^{p} k\left(\frac{\boldsymbol{x}_{t-l} - \boldsymbol{y}_{n-l}}{h}\right)}{\sum_n \prod_{l=1}^{p} k\left(\frac{\boldsymbol{x}_{t-l} - \boldsymbol{y}_{n-l}}{h}\right)},$$

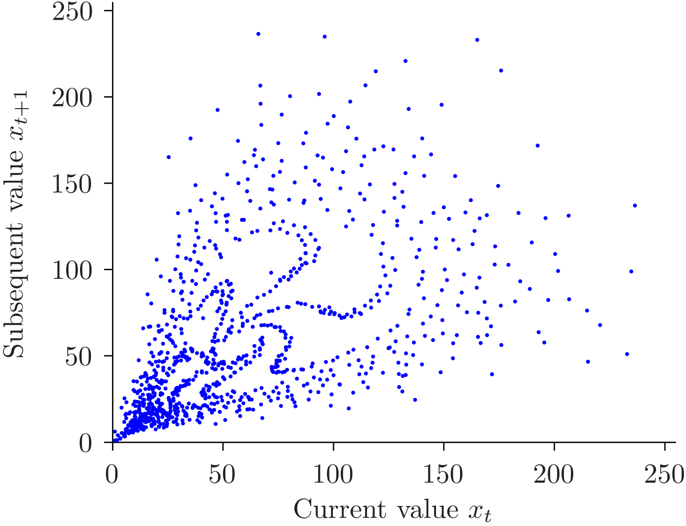
assuming the kernel factors as $k\left(\underline{\boldsymbol{r}}\right) = \prod_{l=0}^{p} k\left(\boldsymbol{r}_l\right)$

KDE-MM Remarks

- KDE-MM converges on the true process as $N \to \infty$
  - Subject to some technical criteria
  - Ergodicity, stationarity, appropriate bandwidth selection
- Maximum likelihood estimation for h is inappropriate
  - Training set likelihood is degenerate as $h \to 0$
  - One component centered on each data point

Degeneracy Illustrated

As $h \to 0$, kernels become spikes at the points in $\mathcal{D}$; no generalisation



Degeneracy Circumvented

Maximising the pseudo-likelihood (a kind of cross-validation)

$$\widetilde{f}_{\underline{\boldsymbol{X}}}\left(\underline{\boldsymbol{y}}_1^T \mid \mathcal{D},\, h\right) = \prod_n \frac{1}{h^D} \frac{\sum_{n' \neq n} \prod_{l=0}^{p} k\left(\frac{\boldsymbol{y}_{n-l} - \boldsymbol{y}_{n'-l}}{h}\right)}{\sum_{n' \neq n} \prod_{l=1}^{p} k\left(\frac{\boldsymbol{y}_{n-l} - \boldsymbol{y}_{n'-l}}{h}\right)}$$

prevents points from "explaining themselves"

A Mixture Model

Rewrite the KDE-MM PDF as

$$\widehat{f}_{\boldsymbol{X}_t \mid \underline{\boldsymbol{X}}_{t-p}^{t-1}}\left(\boldsymbol{x}_t \mid \underline{\boldsymbol{x}}_{t-p}^{t-1}\right) = \sum_n \frac{\prod_{l=1}^{p} k\left(\frac{\boldsymbol{x}_{t-l}-\boldsymbol{y}_{n-l}}{h}\right)}{\sum_{n'} \prod_{l=1}^{p} k\left(\frac{\boldsymbol{x}_{t-l}-\boldsymbol{y}_{n'-l}}{h}\right)} \frac{1}{h^D} k\left(\frac{\boldsymbol{x}_t - \boldsymbol{y}_n}{h}\right)$$

$$= \sum_n w_n\left(\underline{\boldsymbol{x}}_{t-p}^{t-1}\right) \frac{1}{h^D} k\left(\frac{\boldsymbol{x}_t - \boldsymbol{y}_n}{h}\right)$$

This is a mixture distribution with context-dependent weights

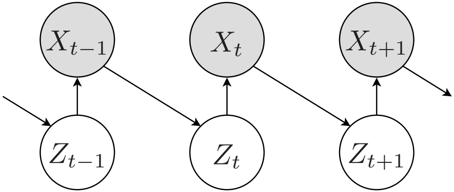
KDE-MM Output

KDE-MM data generation algorithm:

1. Given $\underline{\boldsymbol{x}}_{t-p}^{t-1}$, one selects a mixture component $z_t \leq N$ according to

$$p_{Z_t|\underline{\boldsymbol{X}}_{t-p}^{t-1}}\left(z_t \mid \underline{\boldsymbol{x}}_{t-p}^{t-1}\right) = w_{z_t}\left(\underline{\boldsymbol{x}}_{t-p}^{t-1}\right) = \frac{\prod_{l=1}^{p} k\left(\frac{\boldsymbol{x}_{t-l}-\boldsymbol{y}_{z-l}}{h}\right)}{\sum_n \prod_{l=1}^{p} k\left(\frac{\boldsymbol{x}_{t-l}-\boldsymbol{y}_{n-l}}{h}\right)}$$

2. $x_t = y_{z_t} + \eta_t$, where η_t is kernel-shaped IID noise
3. Increment t and start over



Connection to Unit Selection

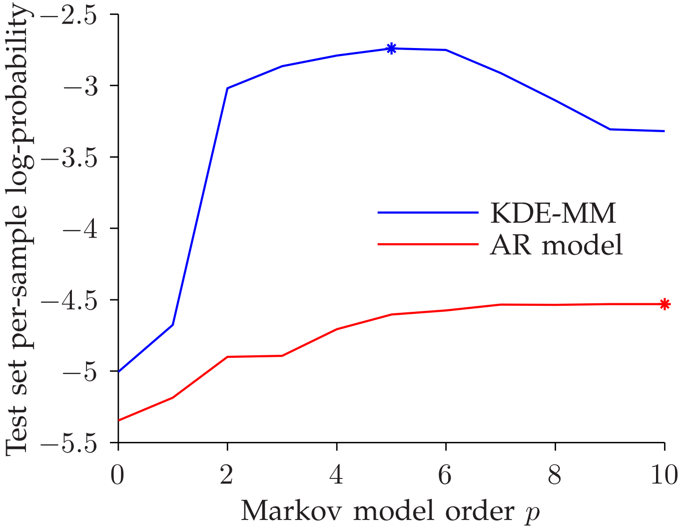
- Data-driven output generation
- Concatenate well-matching data frames (plus some noise)
  - Follow single trajectories in isolated regions
  - May switch to another trajectory where the context is ambiguous
  - The bandwidth h controls context sensitivity
- Reminiscent of unit selection synthesis
  - $h \rightarrow 0$ approaches unit selection, but fully probabilistic!
  - Also similar to the time-series bootstrap from statistics

Outline

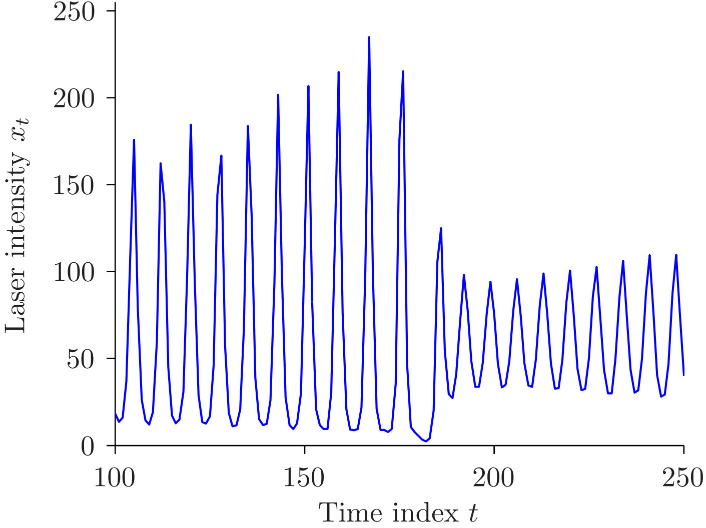
1. Introduction
2. Kernel density estimation
3. KDE Markov models
   - Experiments
4. KDE-HMMs
   - Parameter estimation
   - Experiments
5. Summary and outlook

Evaluation

p-order KDE-MMs vs. linear AR models on held-out laser data ($N = 3000$)

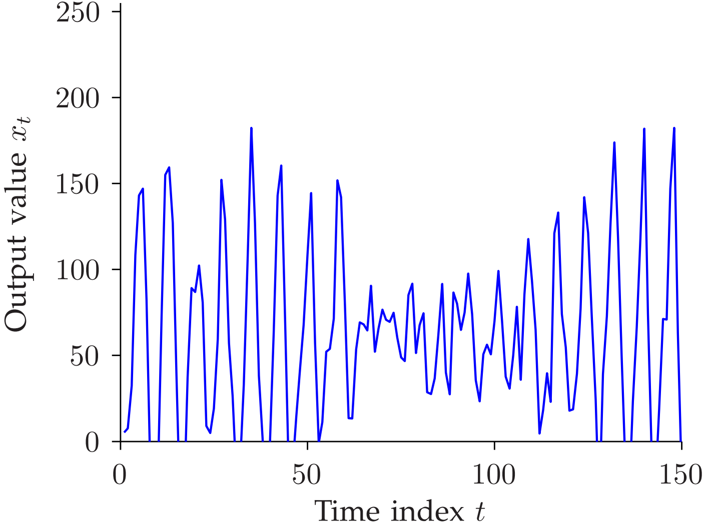


Excerpt from original laser data-series



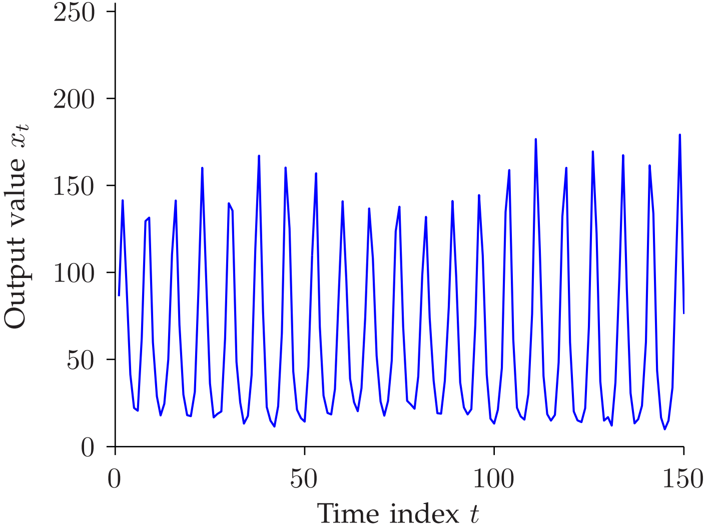
Sample Output

Sample from best linear AR model (order $p = 10$)



Sample Output

Sample from best KDE-MM ($p = 6$)



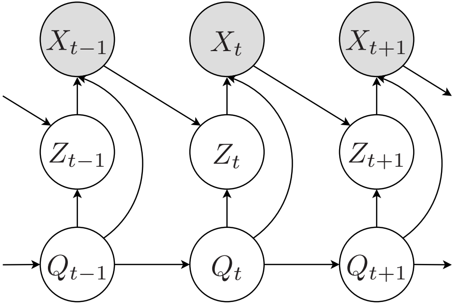
Outline

1. Introduction
2. Kernel density estimation
3. KDE Markov models
   - Experiments
4. KDE-HMMs
   - Parameter estimation
   - Experiments
5. Summary and outlook

KDE in Synthesis

To use KDE/KCDE in synthesis, we need a hidden state to control the output

- Novel proposal: KDE-HMM (a nonlinear autoregressive HMM)
- Nonlinear autoregressive HMM
  - States follow a Markov chain $p_{Q_t \mid Q_{t-1}}(q_t \mid q_{t-1})$
  - State-conditional next-step distribution $\widehat{f}_{\boldsymbol{X}_t \mid Q_t, \underline{\boldsymbol{X}}_{t-p}^{t-1}}(\boldsymbol{x}_t \mid q_t, \underline{\boldsymbol{x}}_{t-p}^{t-1})$ switches between KDE-MMs



KDE-HMM Details

- Data points n are assigned to states using weights w_{qn}
  - $w_{qn} \geq 0$, with $\sum_{n=1}^{N} w_{qn} = 1$ for normalisation
- It is compelling to relax parts of the model
  - State and lag-dependent bandwidths h_{ql}
- Assuming a scalar series, the resulting PDF is

$$\widehat{f}_{X_t \mid Q_t, \underline{X}_{t-p}^{t-1}} \left(x_t \mid q, \underline{x}_{t-p}^{t-1} \right) = \frac{\sum_n \kappa_{qn} \left(\underline{x}_{t-p}^{t-1} \mid \boldsymbol{h}_q \right) k \left(\frac{x_t - y_n}{h_{q0}} \right)}{h_{q0} \sum_n \kappa_{qn} \left(\underline{x}_{t-p}^{t-1} \mid \boldsymbol{h}_q \right)}$$

$$\kappa_{qn} \left(\underline{x}_{t-p}^{t-1} \mid \boldsymbol{h}_q \right) = w_{qn} \prod_{l=1}^{p} k \left(\frac{x_{t-l} - y_{n-l}}{h_{ql}} \right)$$

KDE-HMM Properties

Advantages:

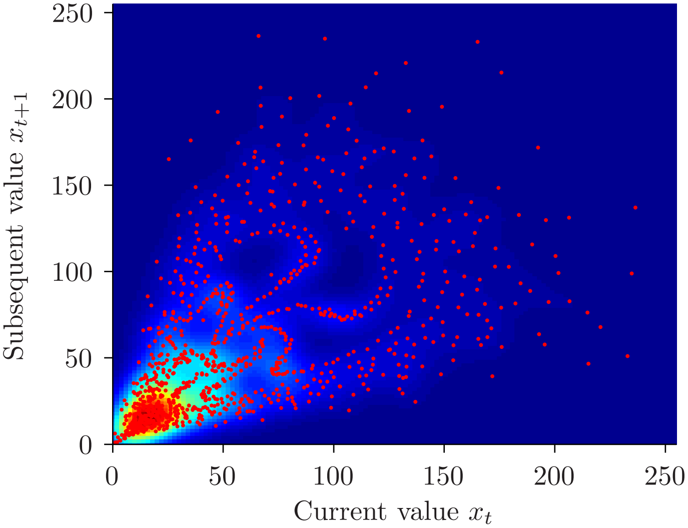
- Flexible short-range correlation modelling
- Hidden state allows output control
- Context-dependent bandwidths

Disadvantages:

- Data requirements
- Computational cost

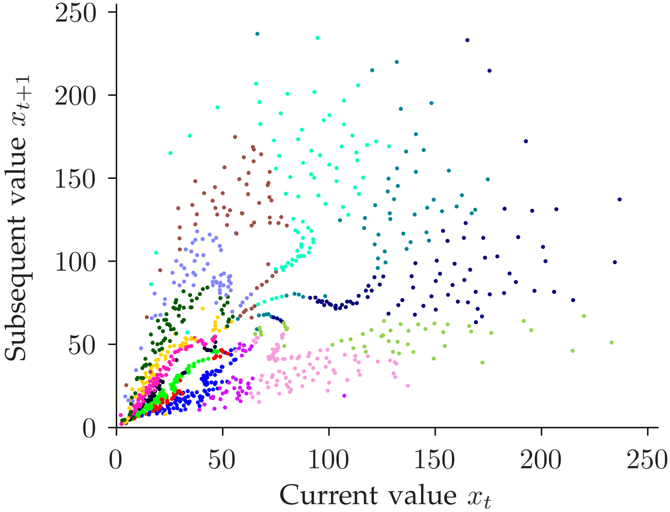
Context-Dependent Bandwidths

Single bandwidth is too coarse in the center, because of the sparse edges



Context-Dependent Bandwidths

Data points coloured according to estimated instantaneous phase



Outline

1. Introduction
2. Kernel density estimation
3. KDE Markov models
   - Experiments
4. KDE-HMMs
   - Parameter estimation
   - Experiments
5. Summary and outlook

Parameter Estimation

Standard techniques apply to derive expectation maximisation (EM) update equations for bandwidths and weights

- Auxiliary function

$$\mathcal{Q}\left(\boldsymbol{\theta}' \mid \widehat{\boldsymbol{\theta}}\right) = \ldots + \frac{1}{2} \sum_{q,\, t,\, n \neq t} \gamma_{qt} \varrho_{qnt}^{\mathrm{num}} \left(\ln \frac{1}{h'_{q0}} - \frac{1}{h'^{2}_{q0}} \left(x_t - y_n\right)^2 \right)$$
$$+ \sum_{q,\, t,\, n \neq t} \gamma_{qt} \varrho_{qnt}^{\mathrm{num}} \left(\ln w'_{qn} - \frac{1}{2} \sum_{l=1}^{p} \frac{1}{h'^{2}_{ql}} \left(x_{t-l} - y_{n-l}\right)^2 \right)$$
$$- \sum_{q,\, t} \gamma_{qt} \ln \left(\sum_{n \neq t} w'_{qn} \exp \left(-\frac{1}{2} \sum_{l=1}^{p} \frac{1}{h'^{2}_{ql}} \left(x_{t-l} - y_{n-l}\right)^2 \right) \right)$$

- Negative log-sum-exp term due to conditioning is an issue

Handling Log-Sum-Exp

1. Extended Baum-Welch (EBW) heuristic from discriminative training
   - Guaranteed ascent for small step lengths (nonconstructive proof)
2. Minorise-maximisation
   - Optimise a locally tight lower bound $\widetilde{\mathcal{Q}}\left(\boldsymbol{\theta}' \mid \widehat{\boldsymbol{\theta}}\right) \leq \mathcal{Q}\left(\boldsymbol{\theta}' \mid \widehat{\boldsymbol{\theta}}\right)$
   - Such bounds can have the same form as other terms in $\mathcal{Q}$ using reverse-Jensen inequalities (Jebara, 2002)

$$
\begin{aligned}
-\ln\left(\sum_{n \neq t} w_{qn} \exp\left(\sum_{l=1}^{p} T_{nl}\left(x_{t-l}\right) \frac{1}{h_{ql}'^{2}} - \mathcal{K}\left(\boldsymbol{h}_q'\right)\right)\right) \\
\geq \sum_{n \neq t} \omega_{qtn} \left(\sum_{l=1}^{p} U_{tnl}\left(x_{t-l}\right) \frac{1}{h_{ql}'^{2}} - \mathcal{K}\left(\boldsymbol{h}_q'\right)\right) - k_{qt}
\end{aligned}
$$

   - Modified sufficient statistics U_{tnl} and weights ω_{qtn} depend on current parameter values $\boldsymbol{h}_q$

Minorise-Maximisation Updates

One obtains a regularisation of the h_{q0} update formula:

$$\widehat{h}_{ql}^{2(\mathrm{new})} = \frac{W_q \widehat{h}_{ql}^{2} + \sum_{t,\, n \neq t} \gamma_{qt} \left(\varrho_{qnt}^{\mathrm{num}} - \varrho_{qnt}^{\mathrm{den}} \right) \left(x_{t-l} - y_{n-l} \right)^2}{W_q + \sum_{t,\, n \neq t} \gamma_{qt} \left(\varrho_{qnt}^{\mathrm{num}} - \varrho_{qnt}^{\mathrm{den}} \right)}$$

- Dependence on previous estimate $\widehat{h}_{ql}^{2}$ through local bound
- Similar formula for updated weights $\widehat{w}_{qn}^{(\mathrm{new})}$
- "Brake weights" W_q restrict update step length
    - Large weights slow convergence

Releasing the Brakes

1. Best reverse-Jensen bounds
   - Guaranteed ascent, but impossibly conservative, e.g.,

$$W_q \gg 10^3 \cdot \left| \sum_{t,\, n \neq t} \gamma_{qt} \left(\varrho_{qnt}^{\mathrm{num}} - \varrho_{qnt}^{\mathrm{den}} \right) \right|$$

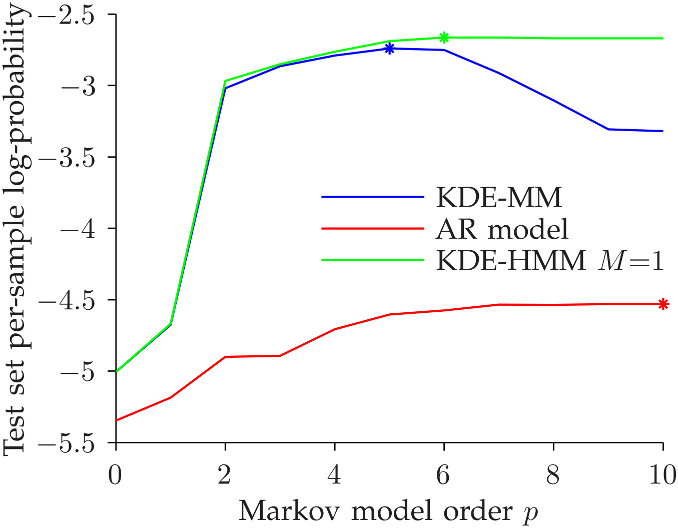
2. Less conservative weights are possible
   - Use approximations related to EBW heuristics (Afify, 2005)
   - Fix w_{qn}, only update bandwidths
   - Reduced total weight, e.g., $\widetilde{W}_q \approx 4 \cdot \left| \sum_{t,\, n \neq t} \gamma_{qt} \left(\varrho_{qnt}^{\mathrm{num}} - \varrho_{qnt}^{\mathrm{den}} \right) \right|$
   - Always increase likelihood in experiments

Outline

1. Introduction
2. Kernel density estimation
3. KDE Markov models
   - Experiments
4. KDE-HMMs
   - Parameter estimation
   - Experiments
5. Summary and outlook

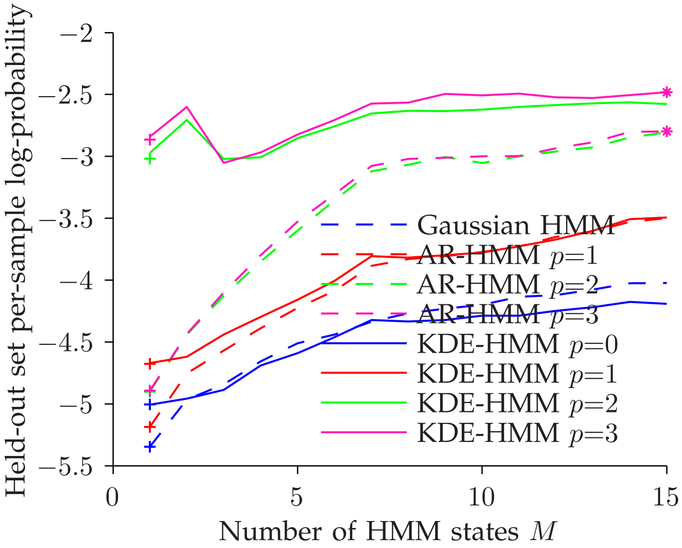
Evaluation

Context-sensitive bandwidth improves on KDE-MMs ($N = 3000$)



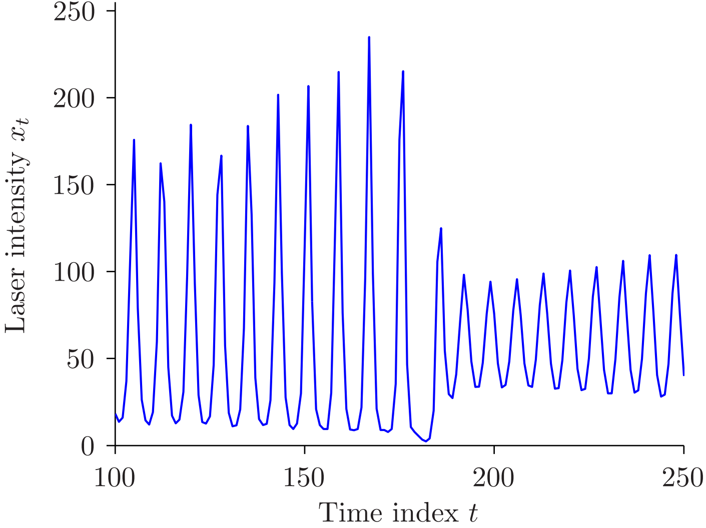
Evaluation

KDE-HMMs yield greater model accuracy than linear AR-HMMs



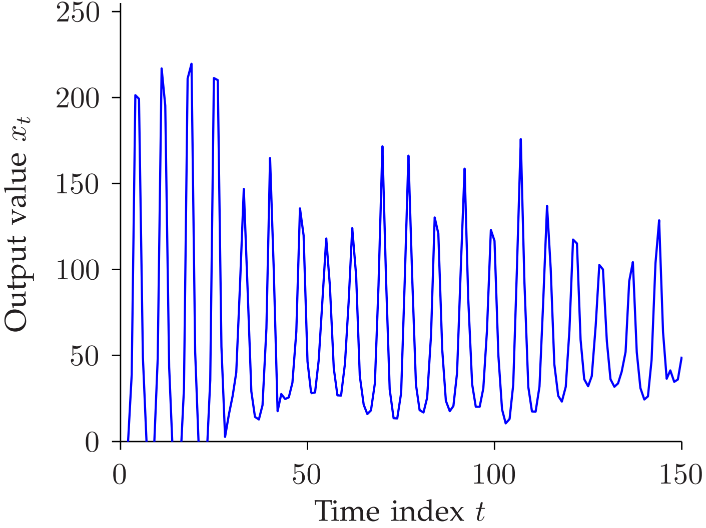
Reference Data

Excerpt from original laser data-series



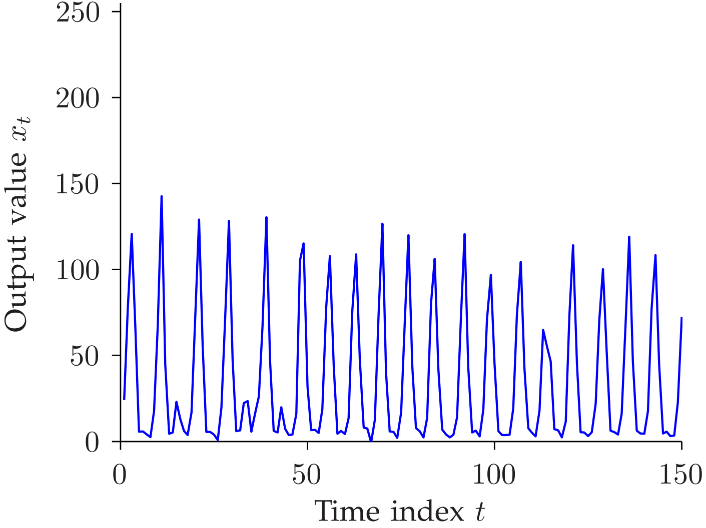
Sample Output

Sample from best linear AR-HMM ($p = 3$, $M = 15$ states)



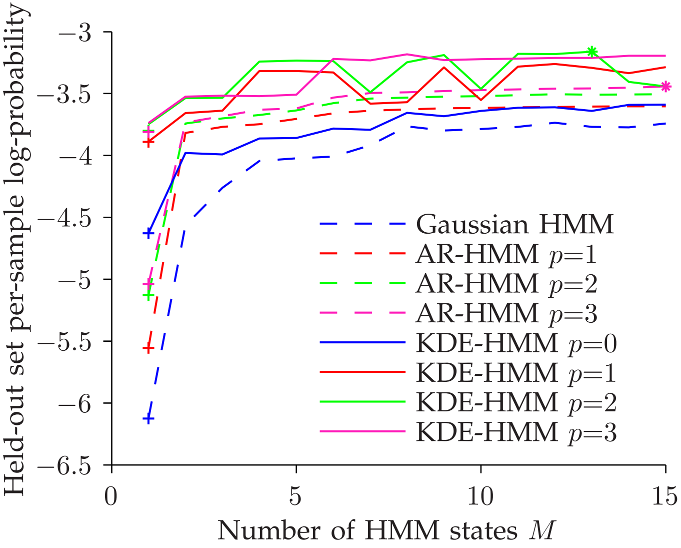
Sample Output

Sample from best KDE-HMM ($p = 3$, $M = 15$)



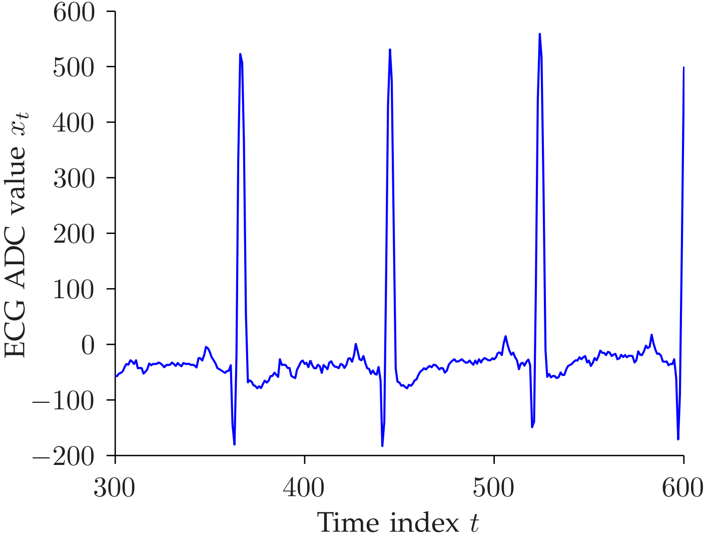
Second Dataset

KDE-HMMs are superior to other models also on ECG data ($N = 3000$)



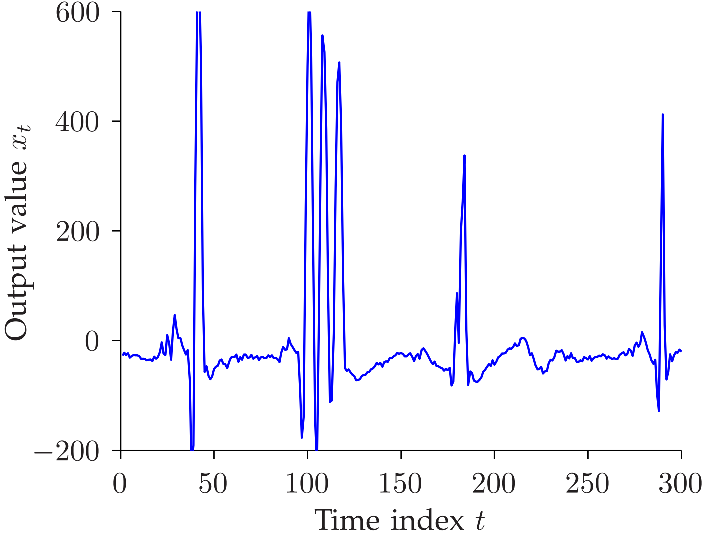
Reference Data

Excerpt from ECG data: empirical standard deviation $\widehat{\sigma}_{\mathrm{ECG}} \approx 109$



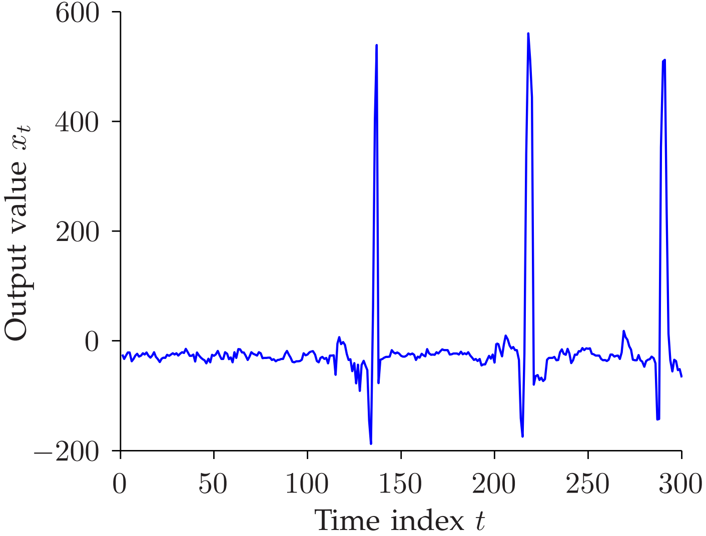
Sample Output

Sample from best linear AR-HMM ($p = 3$, $M = 15$): $\widehat{\sigma}_{\mathrm{AR}} \approx 2490$(!)



Sample Output

Sample from best KDE-HMM ($p = 2$, $M = 13$): $\widehat{\sigma}_{\mathrm{KDE}} \approx 94.3$



Outline

1. Introduction
2. Kernel density estimation
3. KDE Markov models
   - Experiments
4. KDE-HMMs
   - Parameter estimation
   - Experiments
5. Summary and outlook

Successes

1. Theoretically powerful time-series model
   - Nonparametric, asymptotically consistent
2. Parameter update formulas
3. Better modelling of difficult nonlinear series than linear AR-HMMs
4. Compelling for signal synthesis
   - Converges on the true distribution
   - Probabilistic hybrid speech synthesis

Future Possibilities

- Apply to speech
  - Glottal source data
  - Single utterance synthesis
- Also train point-to-state assignments w_{qn} (realignment)
  - Adapt additional EBW heuristics from Woodland & Povey (2002)
- Reduce sample complexity from the infeasible $\mathcal{O}\left(N^2\right)$
  - Approximate kernel evaluations using, e.g., dual trees (Holmes, Gray, & Isbell, 2007)
- Pseudo-likelihood maximisation is unsuitable
  - KDE methods are more developed for integrated square error
  - Unlike recognition, synthesis prioritises peaks rather than tails

The End

The End

Thank you for listening!