Advances in Models for Acoustic Processing

David Barber and Taylan Cemgil

Signal Processing and Communications Lab.

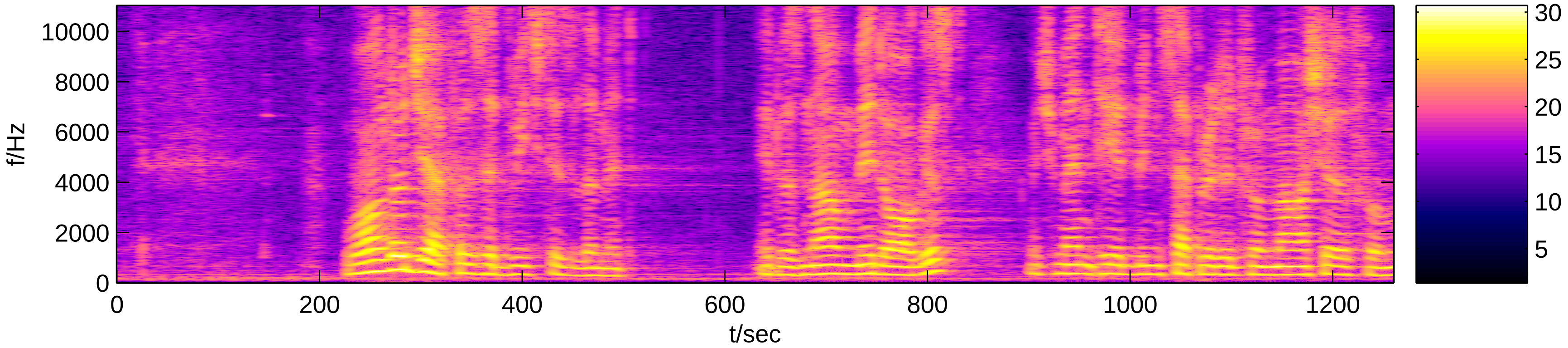
UNIVERSITY OF CAMBRIDGE
Department of Engineering

9 Dec 2006

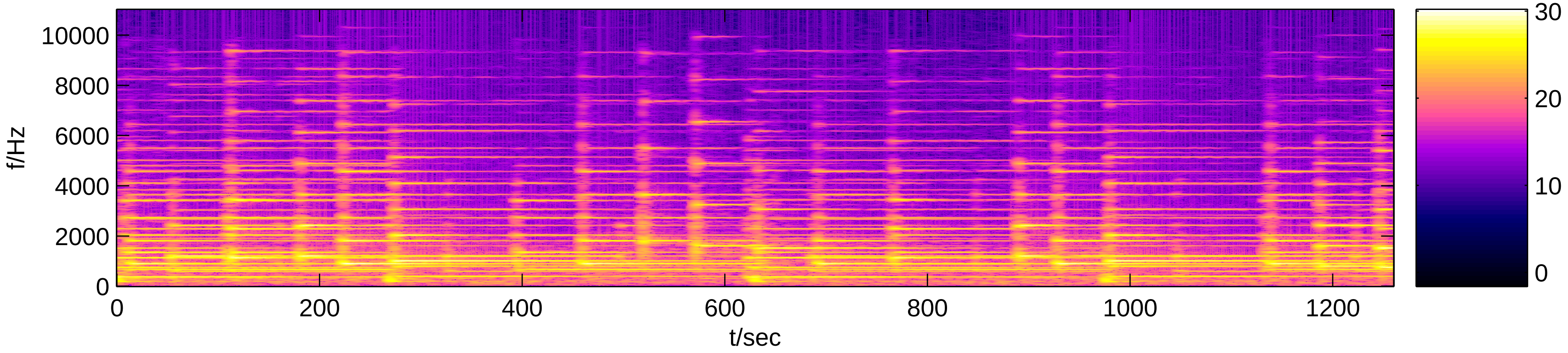
Outline

- Acoustic Modeling and applications
- Parameter estimation and Inference
  - Subspace methods, Variational, Monte Carlo
- Issues

Acoustic Modeling



(Speech)



(Piano)

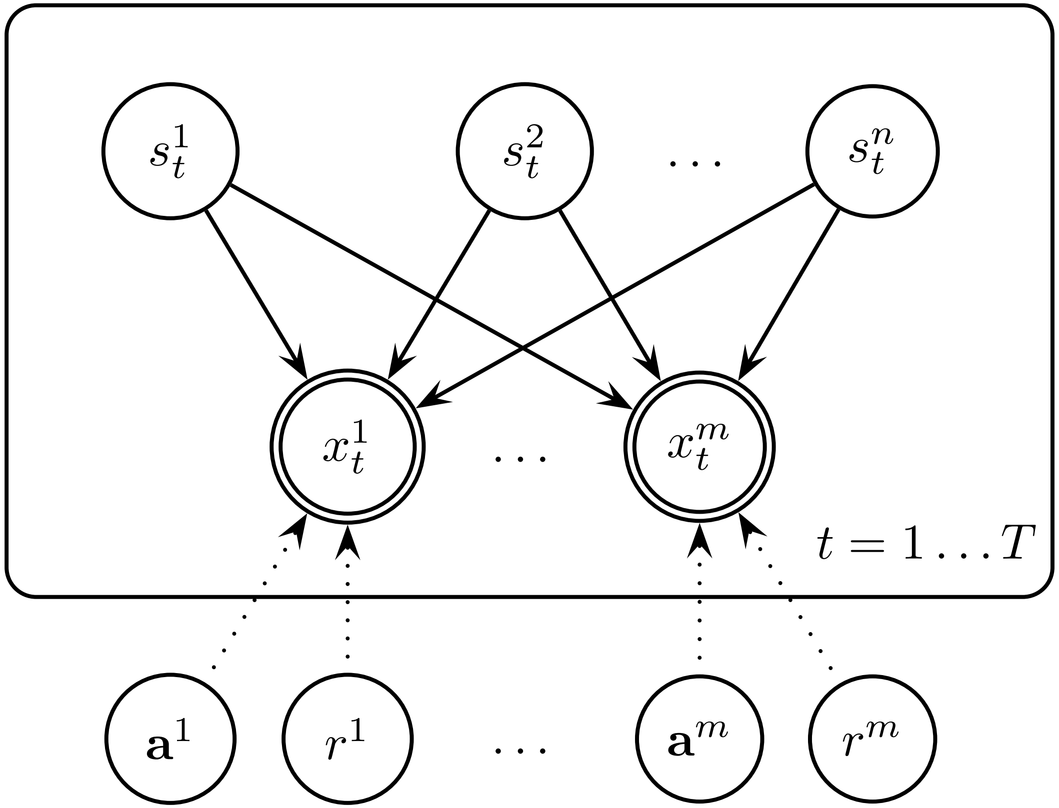
Probabilistic Models

- Once a realistic model is constructed many related task can be cast to posterior inference problems

$$p(\text{Structure}|\text{Observations}) \propto p(\text{Observations}|\text{Structure})p(\text{Structure})$$

  - analysis,
  - localisation,
  - restoration,
  - transcription,
  - source separation,
  - identification,
  - coding,
  - resynthesis, cross synthesis

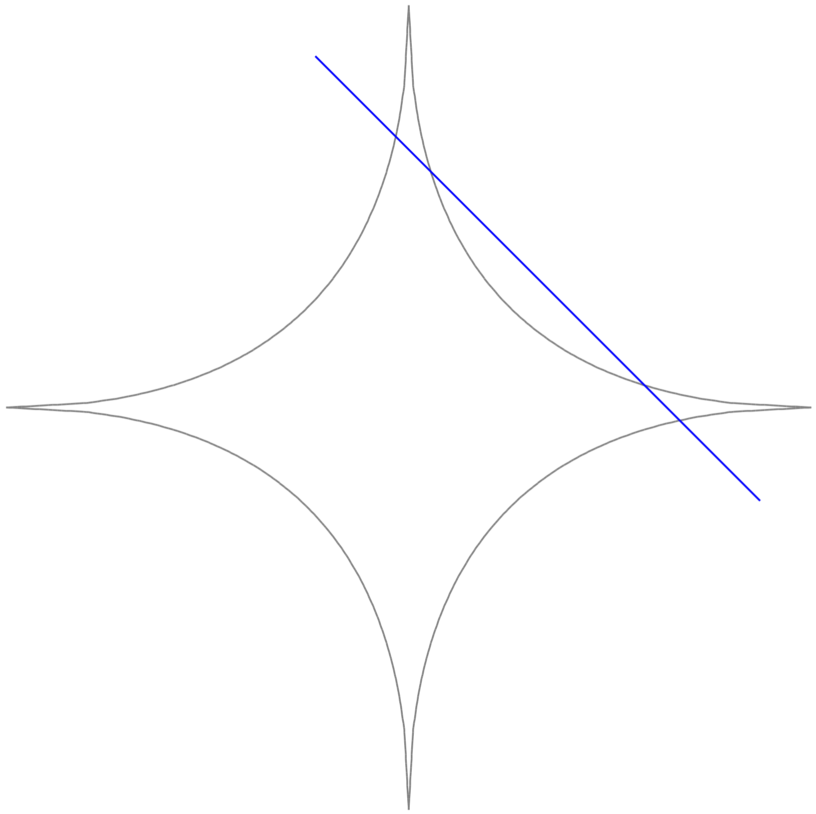
Source Separation



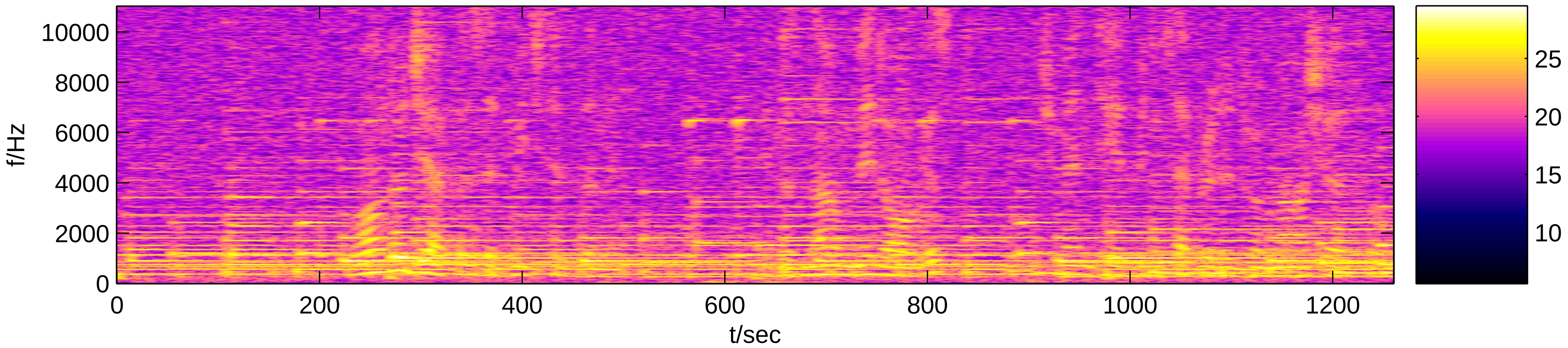
- Joint estimation Sources, Channel noise and mixing system

- Typically underdetermined (Channels $<$ Sources) $\Rightarrow$ Multimodal posterior

Source Separation



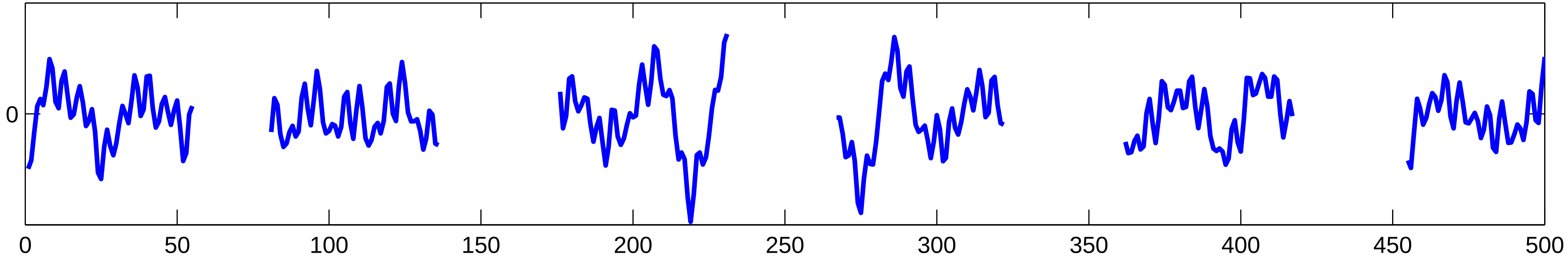
Source Separation



(Speech + Piano + Guitar)

Audio Interpolation

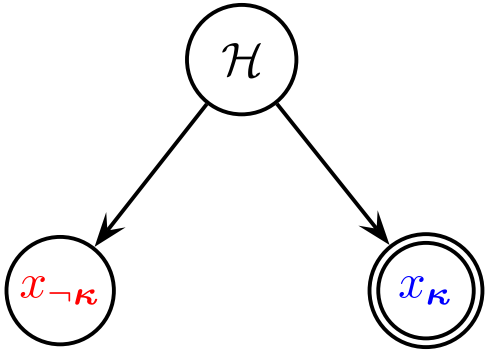
- Estimate missing samples given observed ones
- Restoration, concatenative expressive speech synthesis, ...



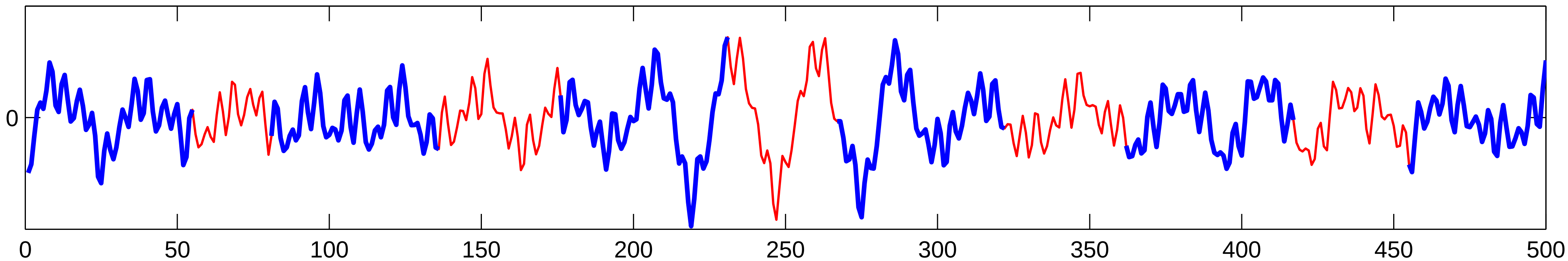
Audio Interpolation

$$p(x_{\neg\kappa}|x_{\kappa}) \quad \propto \quad \int d\mathcal{H} p(x_{\neg\kappa}|\mathcal{H}) p(x_{\kappa}|\mathcal{H}) p(\mathcal{H})$$

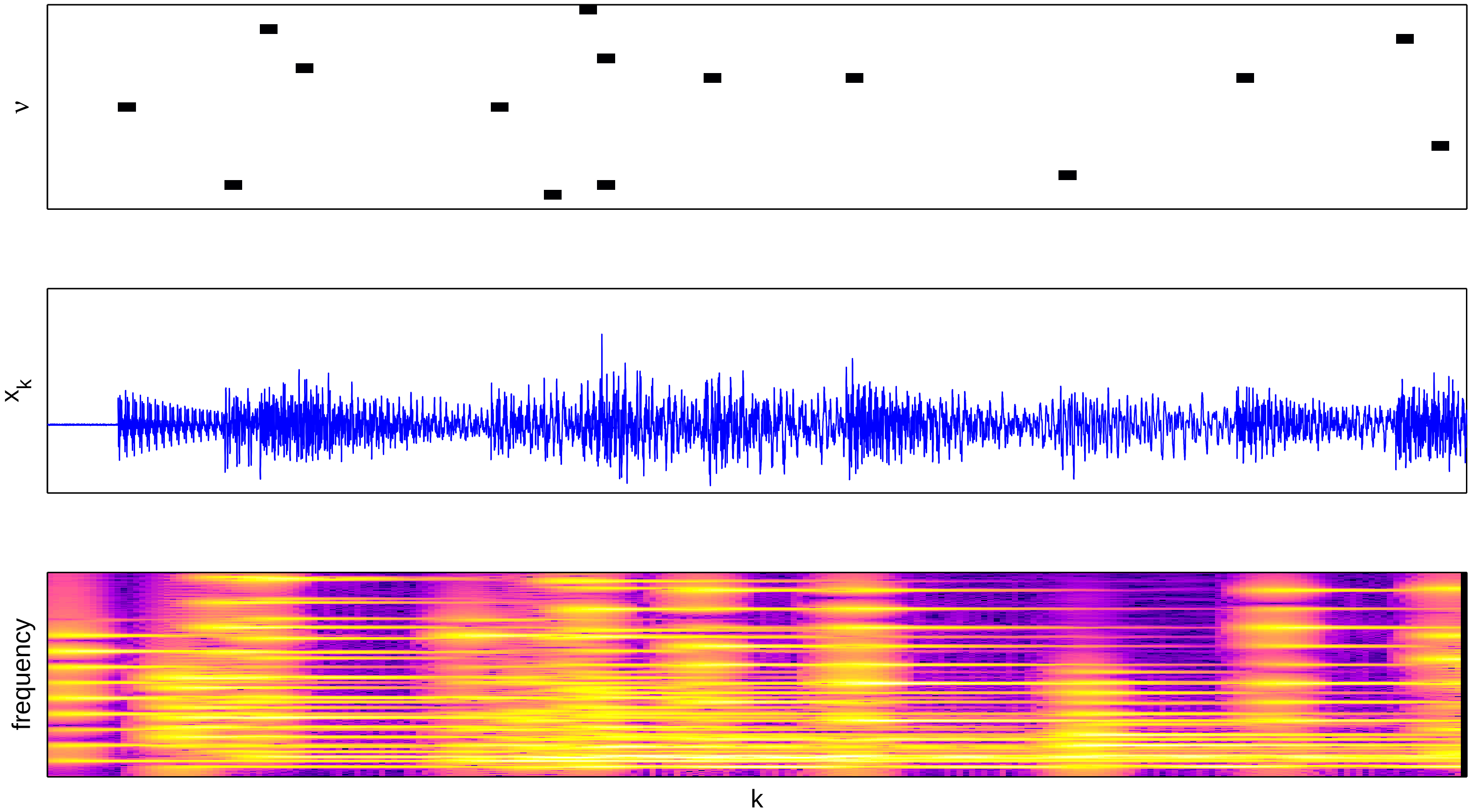
$$\mathcal{H} \quad \equiv \quad (\text{parameters, hidden states})$$



Missing $\qquad y$

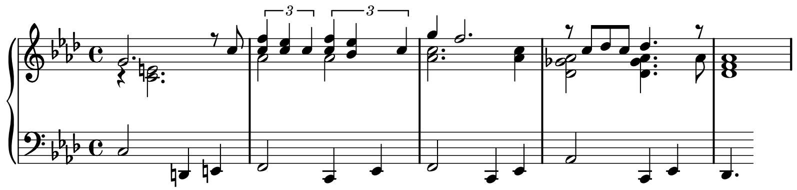


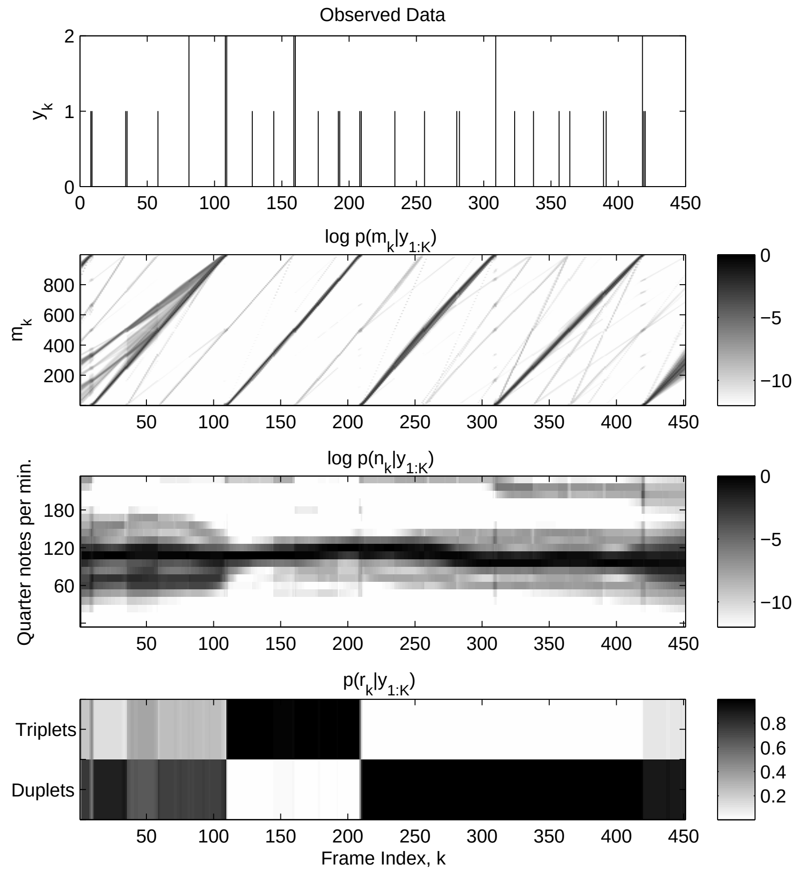
Application: Analysis of Polyphonic Audio



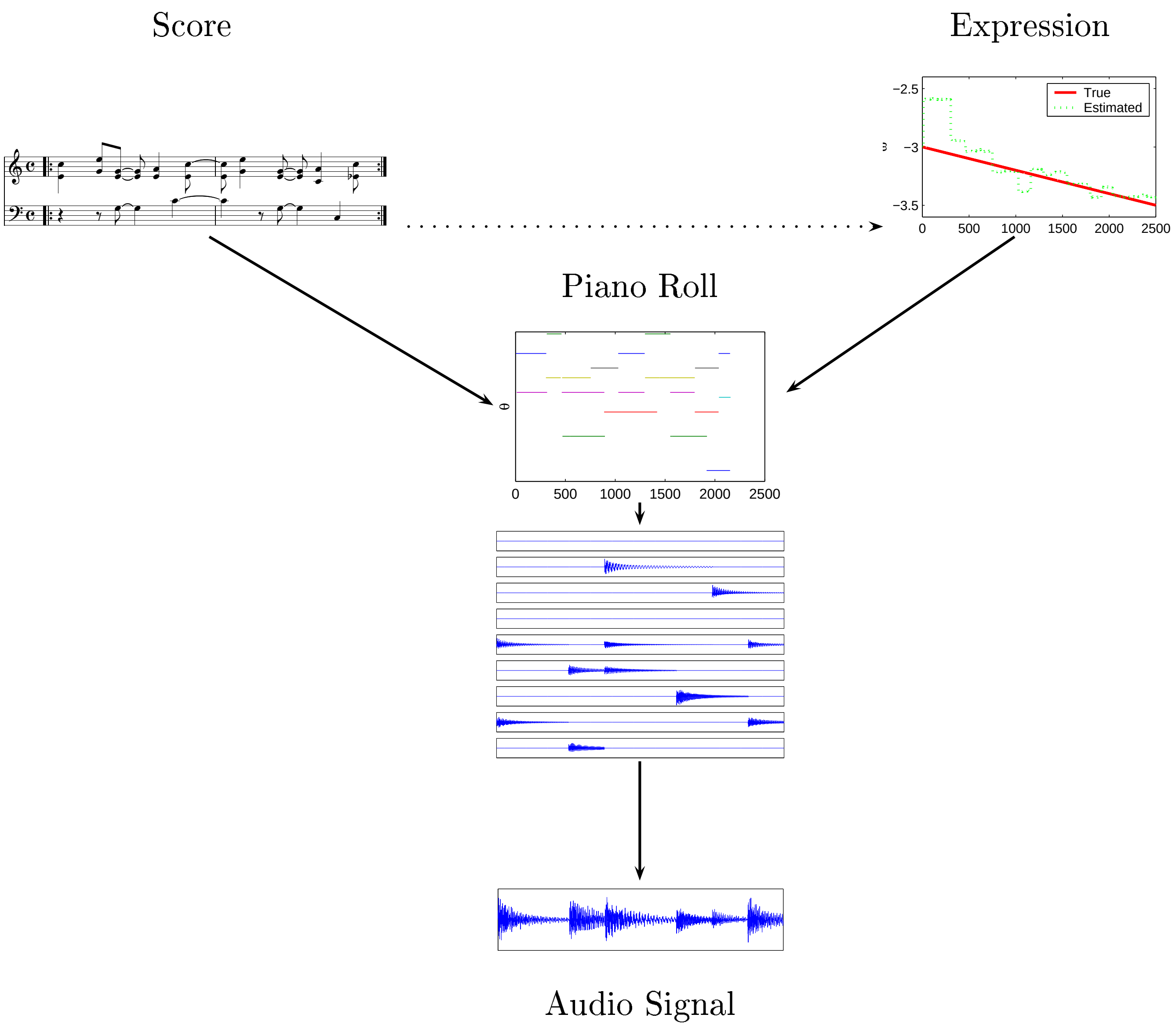
- Each latent process $\nu = 1 \ldots W$ corresponds to a "voice". Indicators $r_{1:W,1:K}$ encode a latent "piano roll"

Tempo, Rhythm, Meter analysis

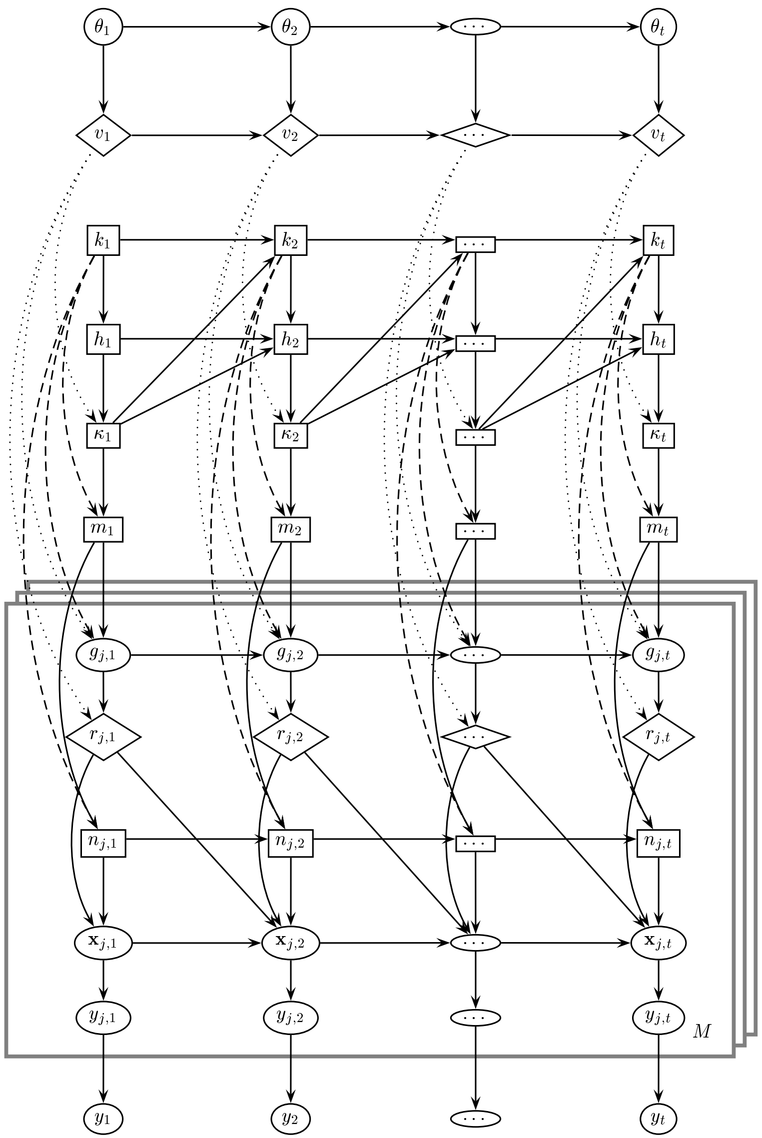




Hierarchical Modeling



Hierarchical Modeling



Time Series Modeling

- Sound is primarily about oscillations and resonance
- Cascade of second order sytems
- Audio signals can often be compactly represented by sinusoidals

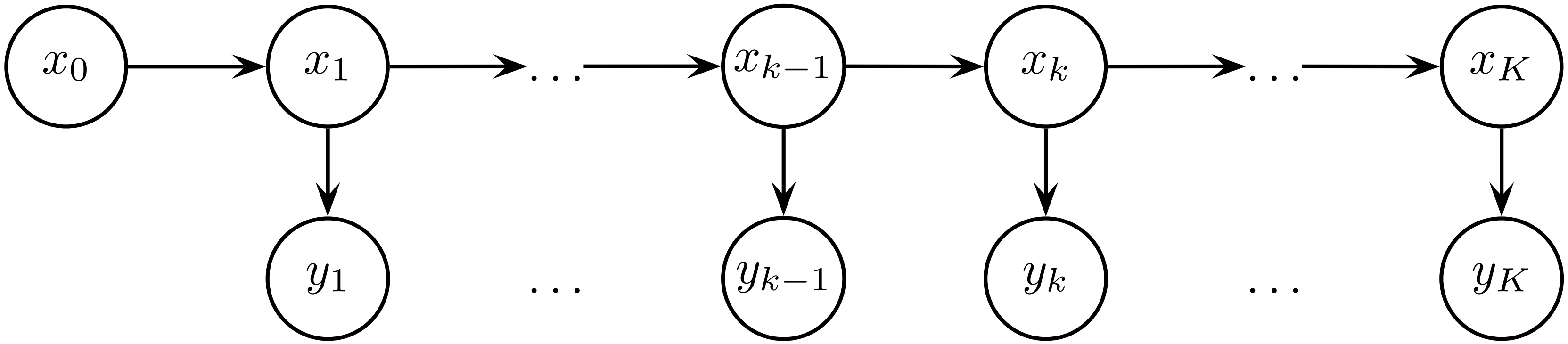
$$
\begin{aligned}
\text{(real)} \quad y_n &= \sum_{k=1}^{p} \alpha_k e^{-\gamma_k n} \cos(\omega_k n + \phi_k) \\
\text{(complex)} \quad y_n &= \sum_{k=1}^{p} c_k (e^{-\gamma_k + j\omega_k})^n
\end{aligned}
$$

$$
\mathbf{y} = F(\gamma_{1:p}, \omega_{1:p})\mathbf{c}
$$

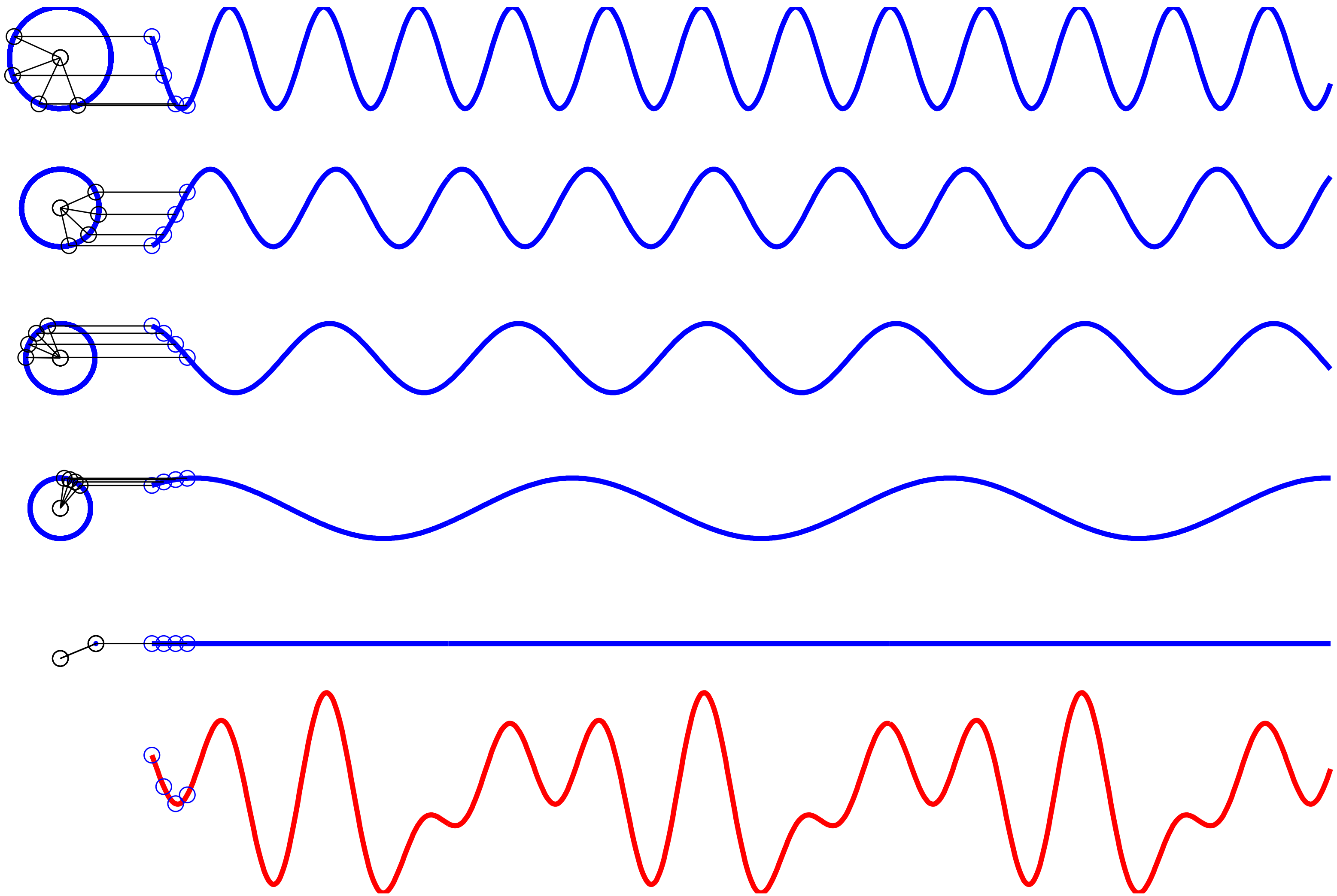
State space Parametrisation

$$x_{n+1} = \underbrace{\begin{pmatrix} e^{-\gamma_1 + j\omega_1} & & \\ & \ddots & \\ & & e^{-\gamma_p + j\omega_p} \end{pmatrix}}_{A} x_n \qquad x_0 = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{pmatrix}$$

$$y_n = \underbrace{\begin{pmatrix} 1 & 1 & \dots & 1 & 1 \end{pmatrix}}_{C} x_n$$



State Space Parametrisation



Classical System identification approach

- The state space representation implies

$$\begin{array}{rcl} x_{n+1} & = & Ax_n \\ y_n & = & Cx_n \end{array} \quad \Rightarrow \quad y_n = CA^n x_0$$

- Therefore we can write for arbitrary L and M the Hankel matrix

$$\underbrace{\begin{pmatrix} y_0 & y_1 & \cdots & y_M \\ y_1 & y_2 & \cdots & y_{M+1} \\ \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & \cdots & y_{L+M} \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} C \\ CA \\ \vdots \\ CA^L \end{pmatrix}}_{\Gamma_{L+1}} \underbrace{\begin{pmatrix} x_0 & Ax_0 & \ldots & A^M x_0 \end{pmatrix}}_{\Omega_{M+1}}$$

Identification via matrix factorisation

1. Given the "impulse response" Hankel matrix Y (Ho and Kalman 1966, Rao and Arun 1992, Viberg 1995), compute a matrix factorisation (typically via SVD)

$$Y = \bar{\Gamma}_{L+1}\bar{\Omega}_{M+1} = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^L \end{pmatrix} \underbrace{\begin{pmatrix} x_0 & Ax_0 & \dots & A^M x_0 \end{pmatrix}}$$

2. Read off C and x_0 from factors $\bar{\Gamma}_{L+1}$ and $\bar{\Omega}_{M+1}$
3. Compute transition matrix by exploiting *shift invariance*

$$\begin{pmatrix} CA \\ CA^2 \\ \vdots \\ CA^L \end{pmatrix} = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{L-1} \end{pmatrix} A \quad \Rightarrow \quad A = \Gamma_{1:n}^{\dagger}\Gamma_{2:n+1}$$

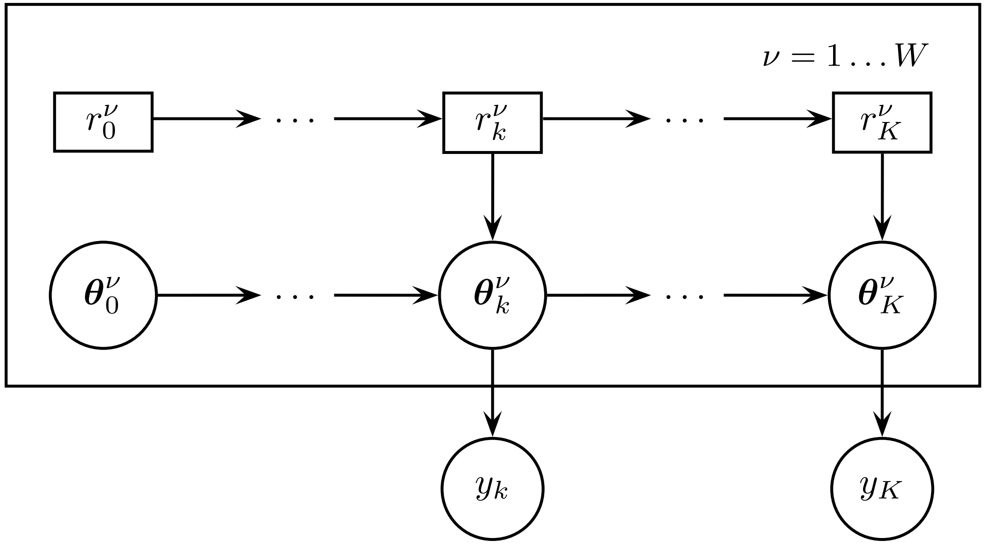
Matrix factorisation ideas have lead to useful methods (N4SID, NMF, MMMF...)

Pros and Cons

- Uses well understood algorithms from numerical linear algebra $\Rightarrow$ often quite fast and numerically stable
- Model selection can be based on numerical rank analysis; inspection of singular values e.t.c.
- Handling of uncertainty and nonstationarity is not very transparent
- Prior knowledge is hard to incorporate

Hierarchical Factorial Models

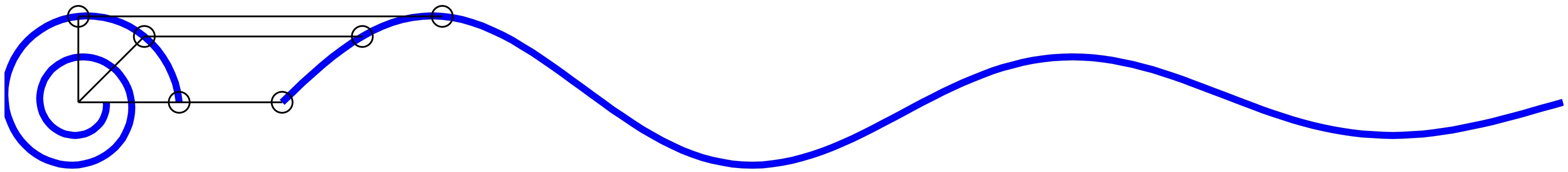
- Each component models a latent process
- The observations are projections



- Generalises Source-filter models

Harmonic model with changepoints

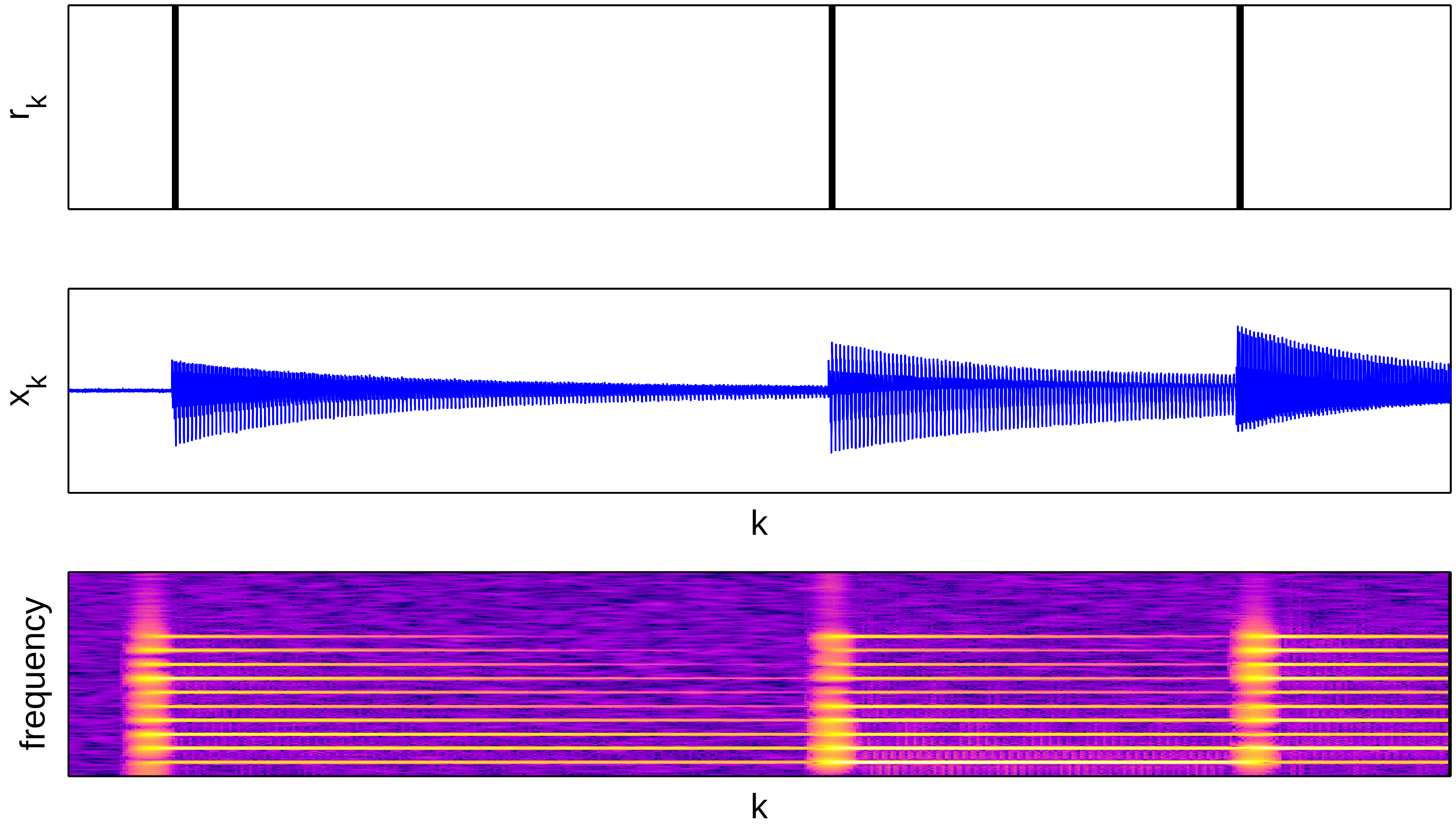
$$
\begin{aligned}
r_k | r_{k-1} &\sim p(r_k | r_{k-1}) \\
\theta_k | \theta_{k-1}, r_k &\sim \underbrace{[r_k = 0]\mathcal{N}(A\theta_{k-1}, Q)}_{\text{reg}} + \underbrace{[r_k = 1]\mathcal{N}(0, S)}_{\text{new}} \\
y_k | \theta_k &\sim \mathcal{N}(C\theta_k, R)
\end{aligned}
$$



$$
A = \begin{pmatrix} G_\omega & & & \\ & G_\omega^2 & & \\ & & \ddots & \\ & & & G_\omega^H \end{pmatrix}^N \qquad G_\omega = \rho_k \begin{pmatrix} \cos(\omega) & -\sin(\omega) \\ \sin(\omega) & \cos(\omega) \end{pmatrix}
$$

damping factor $0 < \rho_k < 1$, framelength N and damped sinusoidal basis matrix C of size $N \times 2H$

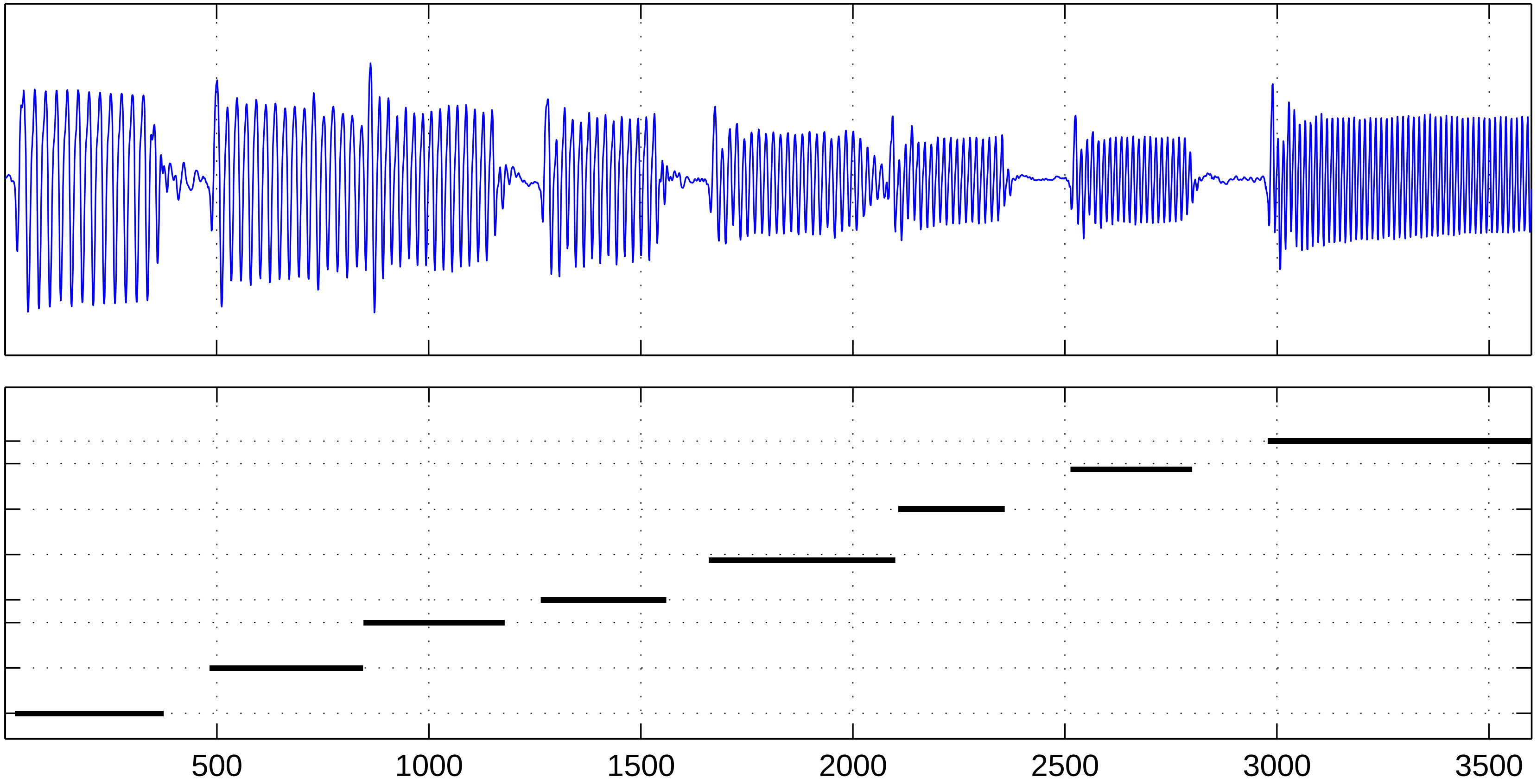
Harmonic model with changepoints



- Each changepoint denotes the onset of a new audio event

Monophonic transcription

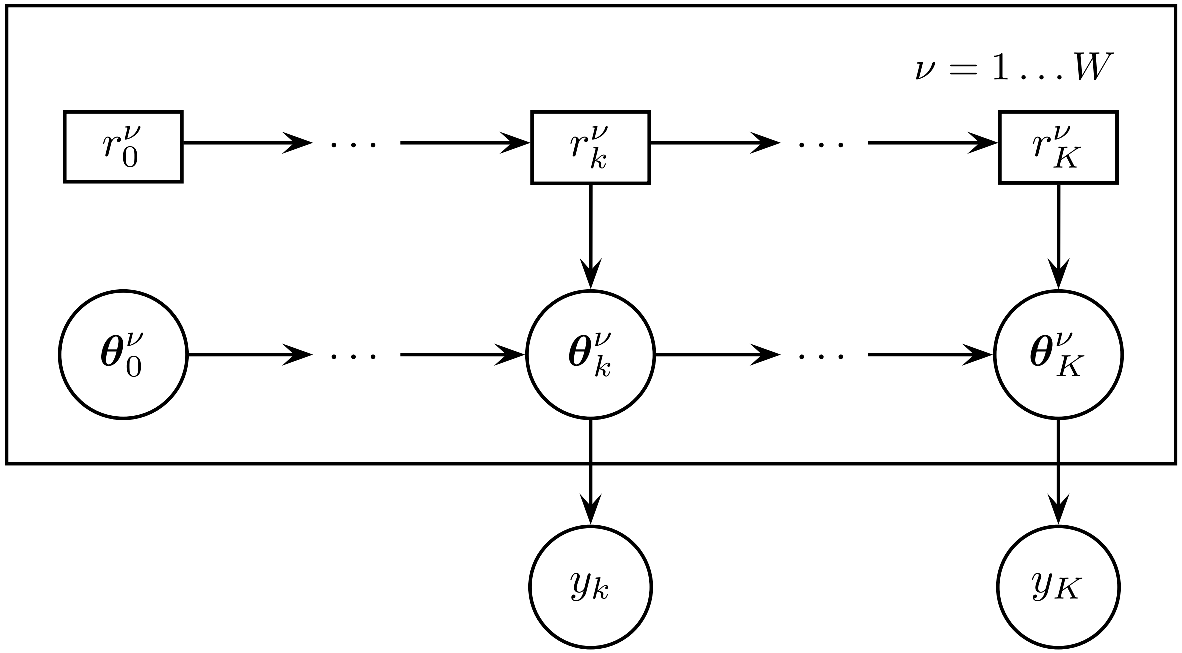
- Detecting onsets, offsets and pitch (Cemgil et. al. 2006, IEEE TSALP)



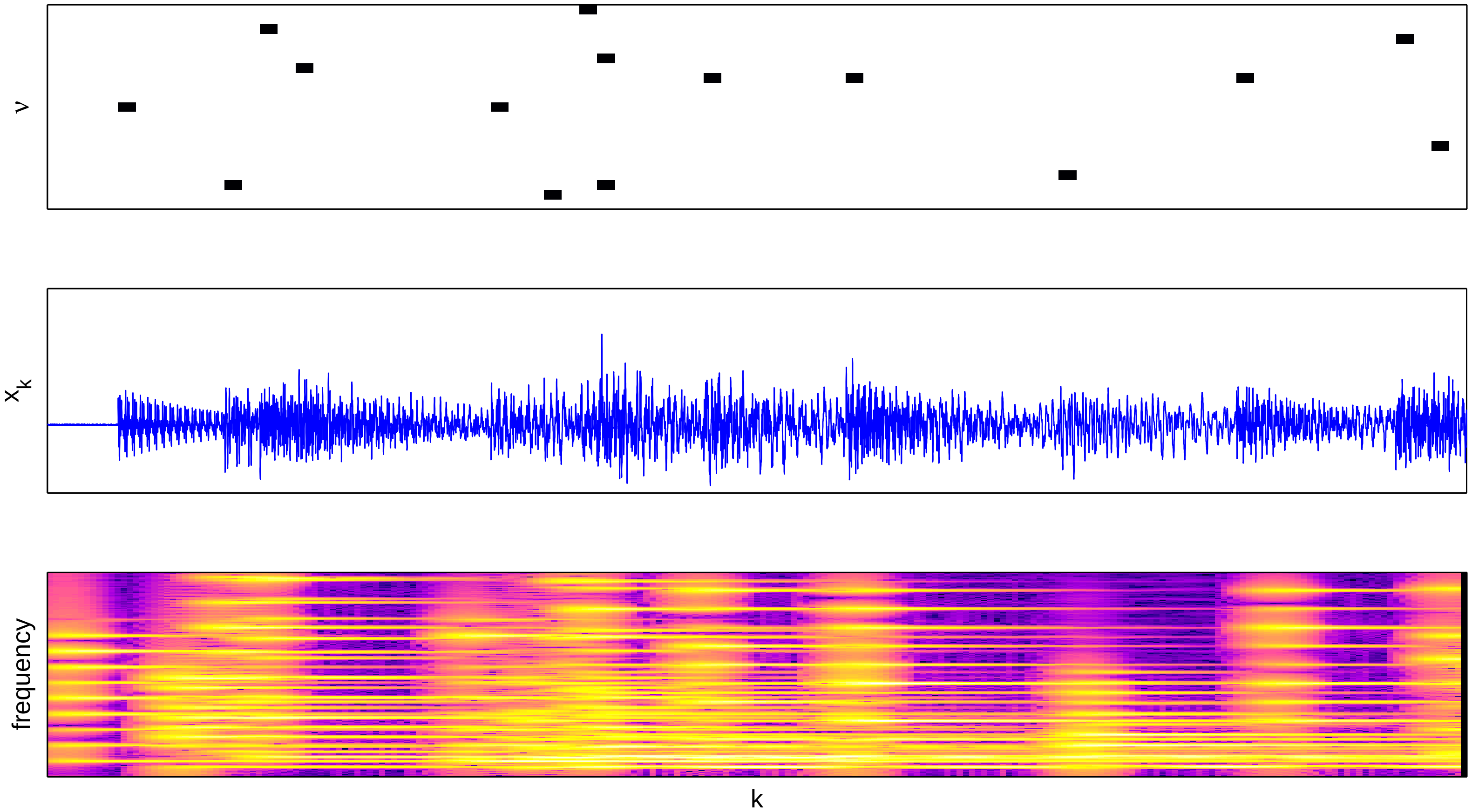
Exact inference is possible

Factorial Changepoint model

$$
\begin{aligned}
r_{0,\nu} &\sim \mathcal{C}(r_{0,\nu}; \pi_{0,\nu}) \\
\theta_{0,\nu} &\sim \mathcal{N}(\theta_{0,\nu}; \mu_\nu, P_\nu) \\
r_{k,\nu}|r_{k-1,\nu} &\sim \mathcal{C}(r_{k,\nu}; \pi_\nu(r_{t-1,\nu})) && \text{Changepoint indicator} \\
\theta_{k,\nu}|\theta_{k-1,\nu} &\sim \mathcal{N}(\theta_{k,\nu}; A_\nu(r_k)\theta_{k-1,\nu}, Q_\nu(r_k)) && \text{Latent state} \\
y_k|\theta_{k,1:W} &\sim \mathcal{N}(y_k; C_k\theta_{k,1:W}, R) && \text{Observation}
\end{aligned}
$$



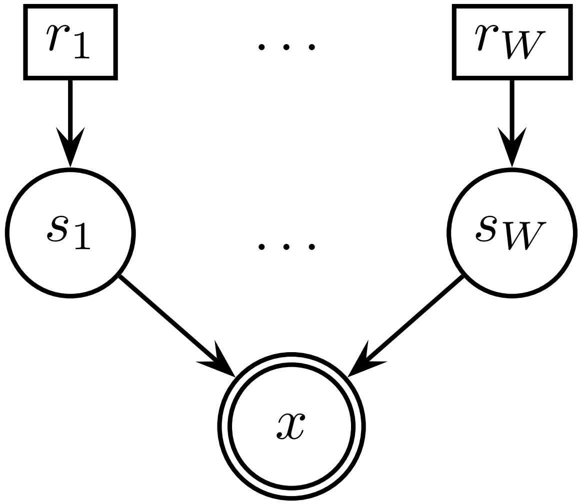
Application: Analysis of Polyphonic Audio



- Each latent changepoint process $\nu = 1 \ldots W$ corresponds to a "piano key". Indicators $r_{1:W,1:K}$ encode a latent "piano roll"

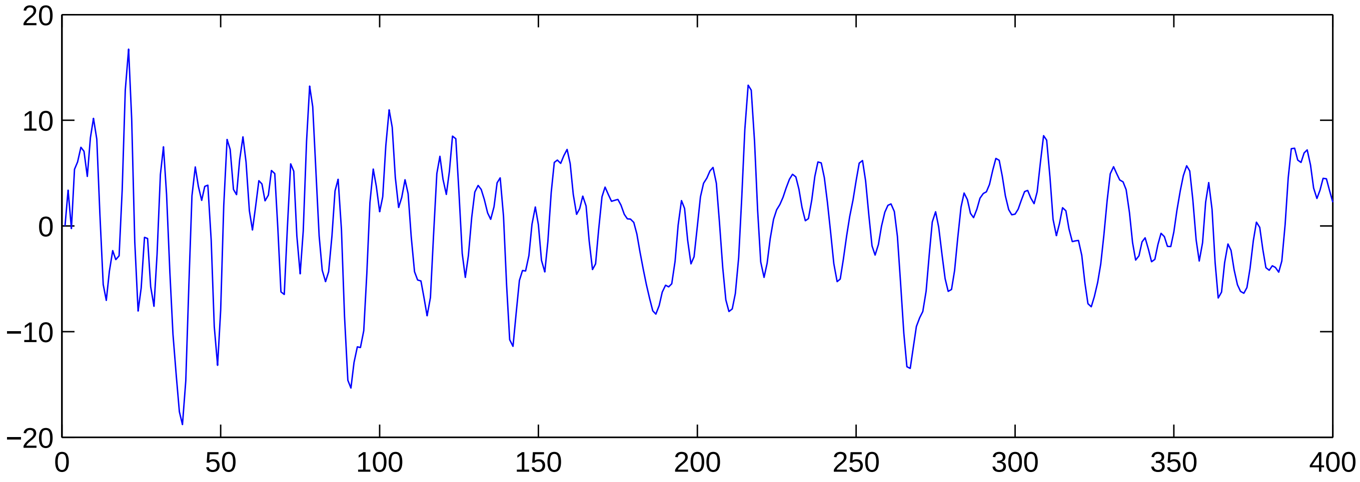
Single time slice - Bayesian Variable Selection

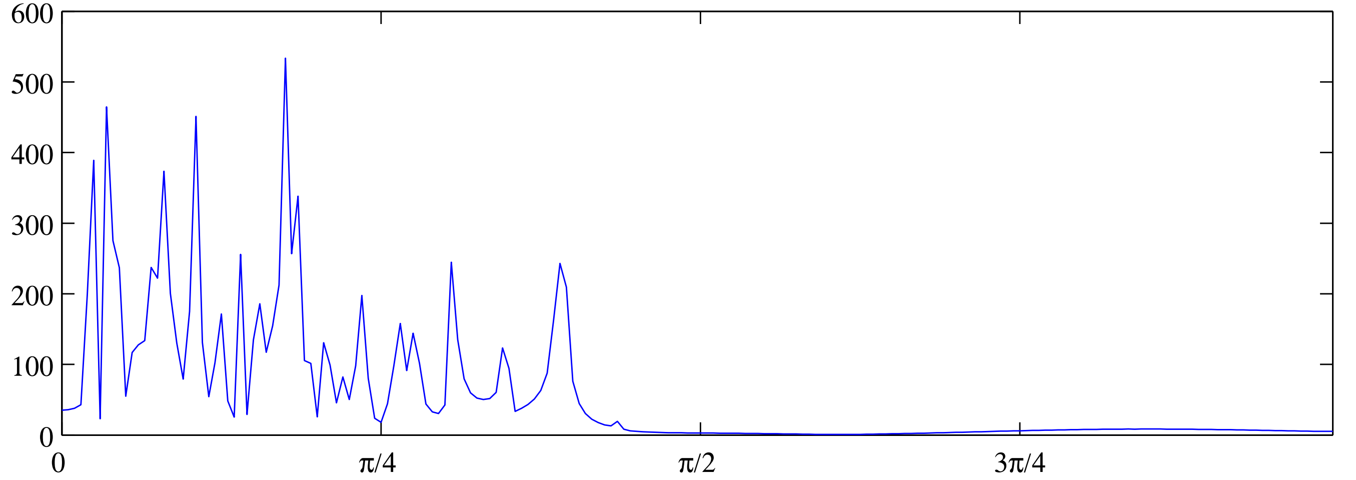
$$
\begin{aligned}
r_i &\sim \mathcal{C}(r_i; \pi_{\text{on}}, \pi_{\text{off}}) \\
s_i|r_i &\sim [r_i = \text{on}]\mathcal{N}(s_i; 0, \Sigma) + [r_i \neq \text{on}]\delta(s_i) \\
\mathbf{x}|s_{1:W} &\sim \mathcal{N}(\mathbf{x}; C s_{1:W}, R) \\
C &\equiv [\ C_1 \quad \dots \quad C_i \quad \dots \quad C_W\ ]
\end{aligned}
$$



- Generalized Linear Model – Column's of C are the basis vectors
- The exact posterior is a mixture of 2^W Gaussians
- When W is large, computation of posterior features becomes intractable.
- Sparsity by construction (Olshausen and Millman, Attias, ...)

Chord detection example





Inference : Iterative Improvement

$$r^*_{1:W} = \arg\max_{r_{1:W}} \int ds_{1:W} p(y|s_{1:W}) p(s_{1:W}|r_{1:W}) p(r_{1:W})$$

iteration	r_1 … r_M	$\log p(y_{1:T}, r_{1:M})$
1	○ ○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	-1220638254
2	○ ○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○ ○	-665073975
3	○ ○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○ ●	-311983860
4	○ ○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ● ○ ●	-162334351
5	○ ○ ○ ○ ○ ○ ○ ● ● ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ● ○ ●	-43419569
6	○ ○ ○ ○ ○ ○ ○ ● ● ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ● ○ ● ○ ●	-1633593
7	○ ○ ○ ○ ○ ○ ○ ● ● ○ ○ ● ○ ○ ● ○ ○ ○ ○ ● ○ ● ○ ●	-14336
8	○ ○ ○ ○ ○ ○ ○ ● ● ○ ● ● ○ ○ ● ○ ○ ○ ○ ● ○ ● ○ ●	-5766
9	○ ○ ○ ○ ○ ○ ○ ● ○ ○ ● ● ○ ○ ● ○ ○ ○ ○ ● ○ ● ○ ●	-5210
10	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ● ● ○ ○ ● ○ ○ ○ ○ ● ○ ● ○ ●	-4664
True	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ● ● ○ ○ ● ○ ○ ○ ○ ● ○ ● ○ ●	-4664

Inference : MCMC/Gibbs sampler

- MCMC: Construct a markov chain with stationary distribution as the desired posterior $\mathcal{P}$
- Gibbs sampler: We cycle through all variables $r_\nu = 1 \dots W$ and sample from full conditionals

$$r_\nu \quad \sim \quad p(r_\nu | r_1^{(t+1)}, r_2^{(t+1)}, \dots, r_{\nu-1}^{(t+1)}, r_{\nu+1}^{(t)}, \dots, r_W^{(t)})$$

- Rao-Blackwellisation: Conditioned on $r_{1:W}$, the latent variables $s_{1:W}$ can be integrated over analytically.

Variational Bayes – Structured mean field

- VB: Approximate a complicated distribution $\mathcal{P}$ with a simpler, tractable one $\mathcal{Q}$ in the sense of

$$\mathcal{Q}^* = \operatorname*{argmin}_{\mathcal{Q}} KL(\mathcal{Q}||\mathcal{P})$$

- KL is the Kullback-Leibler divergence

$$KL(\mathcal{Q}||\mathcal{P}) \quad \equiv \quad \langle \log \mathcal{Q} \rangle_{\mathcal{Q}} - \langle \log \mathcal{P} \rangle_{\mathcal{Q}} \geq 0$$

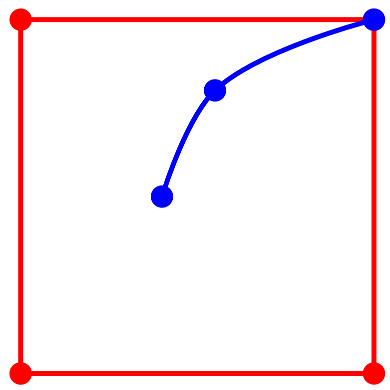
- If $\mathcal{Q}$ obeys the factorisation as $\mathcal{Q} = \prod_\nu \mathcal{Q}_\nu$ the solution is given by the fixed point

$$\mathcal{Q}_\nu \quad \propto \quad \exp\left(\langle \log \mathcal{P} \rangle_{\mathcal{Q}_{\neg \nu}} \right)$$

- Leads to powerful generalisations of the Expectation Maximisation (EM) algorithm (Hinton and Neal 1998, Attias 2000)

MCMC versus Variational Bayes (VB)

- Each configuration of $r_{1:W}$ corresponds to a corner of a W dimensional hypercube

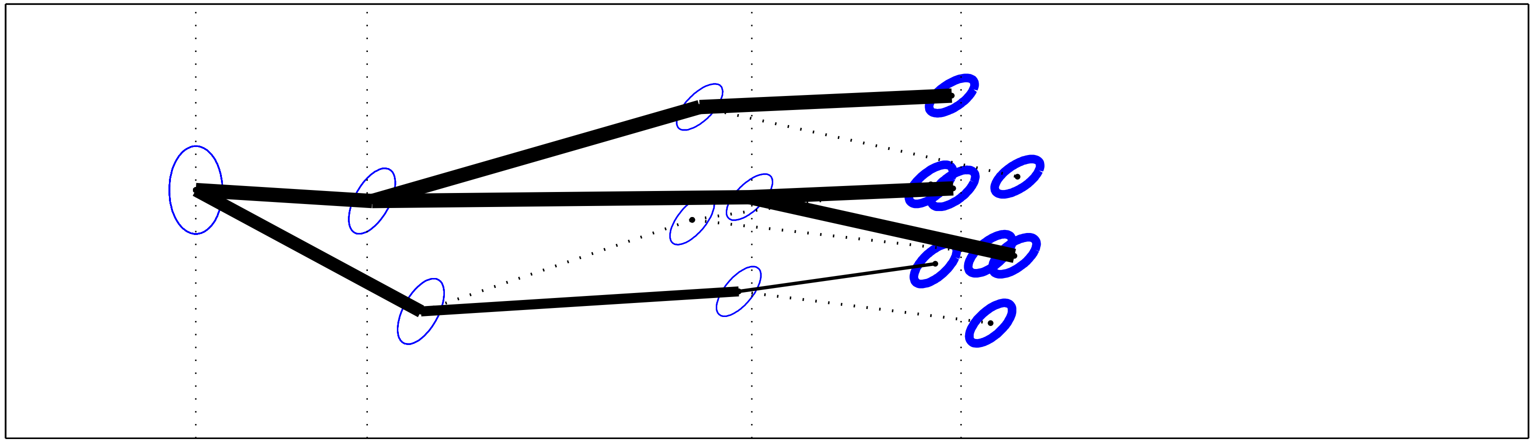


- MCMC moves along the edges stochastically

- VB moves inside the hypercube deterministically

Sequential Inference

- Filtering: Mixture Kalman Filter (Rao-Blackwellized PF) (Chen and Liu 2001)
- MMAP: Breadth-first search algorithm with greedy or randomised pruning, multi-hypothesis tracker (MHT)



- For each hypothesis, there are 2^W possible branches at each timeslice

$\Rightarrow$ Need a fast proposal to find promising branches without exhaustive evaluation

Music Processing challenges

- Computational modeling of human listening and music performance abilities
  - complex and nonstationary temporal structure, both on physical-signal and cognitive-symbolic level
  - Applications: Interactive Music performance, Musicology, Music Information Retrieval, Education

- Analysis
  - identification of individual sound events - notes, kicks
  - invariant characteristics - timbre
  - extraction of higher structure information - tempo, harmony, rhythm
  - not well defined attributes - expression, mood, genre

- Synthesis
  - design of soud synthesis models - abstract or physical
  - performance rendering: generation of a physically, perceptually or artistically feasible control policy

Issues

- What types of modelling approaches are useful for acoustic processing (e.g. hierarchical, generative, discriminative) ?
- What classes of inference algorithms are suitable for these potentially large and hybrid models of sound ?
- How can we improve the quality and speed of inference ?
- Can efficient online algorithms be developed?
- How can we learn efficient auditory codes based on independence assumptions about the generating processes?
- What can biology and cognitive science can tell us about acoustic representations and processing? (and vice versa)