MODELLING ELECTROPHEROGRAM DATA FOR DNA SEQUENCING USING VARIABLE DIMENSION MCMC

N.M.Haan and S.J.Godsill

Signal Processing Group,
Department of Engineering, University of Cambridge,
Cambridge, U.K.
nmh28@eng.cam.ac.uk, sjg@eng.cam.ac.uk

ABSTRACT

DNA sequencing may be considered as a two stage process: the generation of noisy data indicative of DNA sequence by using advanced chemical techniques; and the interpretation of that data. In this paper we present an algorithm for interpretation, or "base calling", which accurately models the underlying process, and is able to incorporate most of the prior information we possess in a mathematically tractable and minimally ad-hoc manner. Our algorithm is framed within a fully Bayesian probabilistic framework, thereby allowing representation of the random nature of the generative process, using a Reversible Jump Metropolis Hastings algorithm [8] and the Gibbs sampler [5] to traverse the variable dimension parameter space. The techniques used to construct our algorithm have only recently become feasible for use in such applications, due to their inherent computational requirements.

1. INTRODUCTION

Much, if not all, of the information which determines our physical function is stored within our cells within the double helical chemical structure known as deoxyribonucleic acid (DNA). For our purposes, DNA can be thought of as a string of symbols (in reality, chemical bases) taken from a four letter alphabet: A (Adenine), G (Guanine), C (Cytosine), or T (Thymine). The sequence of bases, via a triplet encoding, defines the composition of all proteins created within us and consequently forms the key to our biochemical function. In this paper we are concerned with the determination of that sequence.

In 1974, Sanger proposed a method for DNA sequencing which, with technical improvements, has since been almost universally accepted [13]. The idea behind the process is simple. Initially, via a process of replication and truncation the DNA sequence of interest is used to form a large population of partial replicas. Each replica is identical to the sequence of interest over a range of bases, always commencing with the first base of the initial sequence, and terminating some random distance down the strand. That is, for the sequence ACGGG the population would contain a number of each of the following: A, AC, ACG, ACGG, and ACGGG. Each fragment is fluorescently labelled according to its terminating base. Subsequently, the entire population is aligned at the start of a large rectangular gel, and an electric field is applied. The fragments progress through the gel at rates approximately inversely proportional to their length, resulting in the various subpopulations arriving at the end of the gel in sequence order. A laser positioned near the end of the gel excites the fluorescent labels, allowing an emission detector to estimate the number of fragments of each type passing at a given time.

After some preprocessing, four data sets are obtained (henceforth, lanes), corresponding to the variation of fragment concentration with time for each of the four terminating bases. This collection of data is known as an electropherogram and is quite clearly indicative of the underlying base sequence. The superposition of four real data sets is shown in figure 2.

Automated analysis of the electropherogram is a more difficult problem than at first it seems. Ideally we would like a sequence of constant amplitude, evenly spaced, distinct peaks of constant spread, with each peak corresponding to a single base. Such data would be trivial to interpret. Unfortunately, in reality the data is subject to several processes leading to serious degradation of data quality, particularly near the end of the sequence. For example, during a typical sequencing run an electropherogram exhibits variation in the processes which define peak shape, spacing, amplitude, and the nature of the noise. Leakage between lanes is also problematic.

A range of prior information is available to constrain the problem (see, for example, [10]), but in our opinion has yet to be satisfactorily used in combination with modern signal processing techniques.

Providing a synopsis of previous work is difficult since some authors of the more effective algorithms have chosen not to publish their work. Notably, the software developed by ABI which until recently has formed the benchmark by which all others must be judged, has been described in little detail [1]. It does, however, appear to be based on deterministic consideration of the data.

Other algorithms proposed include those described in [3], [6], [7], and most recently (and successfully) [4]. These use a combination of techniques ranging from the use of peak detection algorithms, and frequency domain techniques for peak position detection and channel deconvolution, to the use of neural networks.

In this paper, we present an alternative algorithm which we believe better models the underlying process, is able to incorporate the prior information we possess in a mathematically tractable and minimally ad-hoc manner, and further, which provides suitable error metrics.

2. MODEL FORMULATION

The basis of the process has already been described in the introduction to this paper. Ideally, associated with each fragment is a peak in the lane corresponding to the terminating base, the magnitude of which varies in time according to the changing number of fragments passing under the detection window. In reality, there is significant leakage between lanes, resulting in secondary peaks in each of the remaining lanes. It therefore seems reasonable to model the system as a mixture of peaks in each of the four lanes, each quadtuplet of peaks linked to a single base. The shape of the peaks varies between sequencing chemistries ranging from approximately symmetric Gaussian shaped, to asymmetric heavy tailed forms.

Let the underlying base sequence be of length k, and the corresponding electropherogram be defined at N points in time in each of the four lanes. We define the $4N \times 1$ vector $\mathbf{y} \triangleq \{\mathbf{y}_1^t, ..., \mathbf{y}_N^t\}^t$, where $\mathbf{y}_n \triangleq \{y_{n,1}, y_{n,2}, y_{n,3}, y_{n,4}\}^t$ contains the data in each of the lanes at time n. The following parametric model is proposed:

$$\begin{aligned} \mathbf{y}_n &= \sum_{i=1}^{k} \mathbf{a}_i g\left(\frac{n-p_i}{\sigma_i^2}\right) + \mathbf{e}_n \quad n \in \{1,2,...,N\} && (1) \\ &= \mathbf{x}_n + \mathbf{e}_n && (2) \end{aligned}$$

The function $g(.)$ defines the peak shape we wish to use to fit the data. The peaks corresponding to base i are defined through a position parameter, p_i, and a peak spread (width) parameter, σ_i^2. The amplitude of the peaks corresponding to that base are contained in the 4×1 vector $\mathbf{a}_i \triangleq \{a_{i,1}, a_{i,2}, a_{i,3}, a_{i,4}\}^t$. $\mathbf{a}$, $\mathbf{p}$, and $\boldsymbol{\sigma}$ are the column vectors,
$\{\mathbf{a}_1^t, ..., \mathbf{a}_k^t\}^t$, $\{p_1, ..., p_k\}^t$, and $\{\sigma_1^2, ..., \sigma_k^2\}^t$.

The total effect of the various noise sources (eg. detector noise) is assumed independent Gaussian in each of the four lanes.

$$\mathbf{e} \sim \mathcal{N}\left(\mathbf{e}|\mathbf{1}_{4N\times 1}\boldsymbol{\mu}_e^t, \sigma_e^2 \mathrm{I}_{4N\times 4N}\right) \qquad (3)$$

with mean in the four lanes defined by
$\boldsymbol{\mu}_e \triangleq \{\mu_{e1}, \mu_{e2}, \mu_{e3}, \mu_{e4}\}^t$ and common variance, σ_e^2.

The likelihood, $p(\mathbf{y}|\mathbf{x})$, takes the standard Gaussian form:

$$\left(2\pi\sigma_e^2\right)^{-2N} \exp\left\{-\frac{1}{2\sigma_e^2}\sum_{i=1}^{k}\left[\mathbf{y}_i - \mathbf{x}_i\right]^t\left[\mathbf{y}_i - \mathbf{x}_i\right]\right\} \qquad (4)$$

We have so far not included the underlying base sequence in our model. We now introduce the state vector, $\mathbf{s} \triangleq \{s_1, ..., s_k\}$, as the underlying hidden sequence of a Hidden Markov model (HMM), where each state has a random output process in the variables $\mathbf{a}$, $\mathbf{p}$, and $\boldsymbol{\sigma}$, defined through the probability density,
$p(\mathbf{a}, \mathbf{p}, \boldsymbol{\sigma}|\mathbf{s}, k)$. Each element, s_i, uniquely indicates the type of the corresponding base. In the simplest case, $s_i \in \{A, G, C, T\}$, the state sequence directly corresponding to base sequence. However, in order to incorporate sequence dependent information, we will usually expand the possible state space to include particular special sequences; for example, we could introduce an A state which occurs only when preceded by a G state.

In a Bayesian framework the state sequence is introduced through an additional level in our parameter hierarchy. The following conditional dependence structure is proposed:

$$\begin{aligned} p\left(\mathbf{a}, \mathbf{p}, \boldsymbol{\sigma}, \mathbf{s}, \sigma_e^2, \boldsymbol{\mu}_e, k|\mathbf{y}\right) &\propto p\left(\mathbf{y}|\mathbf{a}, \mathbf{p}, \boldsymbol{\sigma}, \sigma_e^2, k\right) \\ &\times p\left(\mathbf{a}, \mathbf{p}|\mathbf{s}, k\right) p\left(\boldsymbol{\sigma}|k\right) p\left(\mathbf{s}|k\right) \\ &\times p\left(k\right) p\left(\sigma_e^2\right) p\left(\boldsymbol{\mu}_e\right) \qquad (5) \end{aligned}$$

Introducing a number of hyperparameters (considered fixed in this treatment), we now define a flexible prior system which can be adapted to represent the available information for most sequencing chemistries.

For the mean and variance of the noise source we apply diffuse Normal ($\mathcal{N}$) and Inverted Gamma ($\mathcal{IG}$) priors:

$$\sigma_e^2 \sim \mathcal{IG}\left(\sigma_e^2|.\right)$$

$$\boldsymbol{\mu}_e \sim \mathcal{N}\left(\boldsymbol{\mu}_e|.\right)$$

The following random-walk type prior is assigned to ensure that peak spread varies slowly across the sequence:

$$\sigma_i^2 \sim \mathcal{IG}(\alpha(\sigma_{i-1}^2), \beta(\sigma_{i-1}^2))$$

The prior on $\mathbf{p}$ can be made to favour approximately equally spaced peaks, while still allowing for occasional large state dependent deviations. This is achieved using the Gamma distribution ($\mathcal{G}$):

$$p_i \sim p_{i-1} + \mathcal{G}(\alpha(s_i), \beta(s_i))$$

The amplitude prior is defined to allow for state dependent inter-lane leakage:

$$\mathbf{a}_i \sim \mathcal{N}(\boldsymbol{\mu}(s_i), \Sigma(s_i))$$

The prior on state sequence is first order Markovian. $p(s_i|s_{i-1})$ can be chosen according to information about the likely frequencies of the four bases:

$$\mathbf{s} \sim p(s_1) \prod_{i=2}^{k} p(s_i|s_{i-1})$$

A truncated uniform prior ($\mathcal{U}$)is assigned to k:

$$k \sim \mathcal{U}\left(k \in \{1, ..., 2N\}\right)$$

3. SIMULATION FROM THE POSTERIOR DISTRIBUTION

Analytic maximisation of the posterior distribution with respect to the parameters is not, to our knowledge, possible, and we therefore look to numerical techniques for estimation.

Markov Chain Monte Carlo (MCMC) techniques are ideally suited to such statistical problems. These techniques aim to simulate from a distribution of interest through construction of a Markov chain with stationary distribution equal to the required distribution. Quantities of interest can then be examined through histograms of the simulated population(eg. [2]).

3.1. Fixed Dimension Case

We begin by considering the case in which the number of bases responsible for the electropherogram, k, is a known quantity. The posterior distribution of interest becomes $p(\boldsymbol{\theta}|\mathbf{y}, k)$, where $\boldsymbol{\theta}$ denotes the complete set of unknowns. Simulation from this distribution is achieved using a variation of the Gibbs sampler [5] which exploits the structure of our model. Where it is impossible to directly simulate from conditional distributions of interest, Metropolis Hastings algorithms [9] are used.

The following algorithm is proposed:

1. Initialise the parameter vector, $\boldsymbol{\theta}^{(0)} = \left[\mathbf{a}^{(0)}, \mathbf{p}^{(0)}, \sigma^{(0)}, \mathbf{s}^{(0)}, \boldsymbol{\mu}_e^{(0)}, \sigma_e^{2^{(0)}}\right]$. Set $i = 1$.
2. Sample as follows:

   (a) $\mathbf{p}^{(i)} \sim p\left(\mathbf{p}^{(i)}|\boldsymbol{\sigma}^{(i-1)}, \mathbf{s}^{(i-1)}, \boldsymbol{\mu}_e^{(i-1)}, \sigma_e^{2^{(i-1)}}, \mathbf{y}\right)$

(b) $\boldsymbol{\sigma}^{(i)} \sim p\left(\boldsymbol{\sigma}^{(i)}|\mathbf{p}^{(i)}, \mathbf{s}^{(i-1)}, \boldsymbol{\mu}_e^{(i-1)}, \sigma_e^{2^{(i-1)}}, \mathbf{y}\right)$

(c) $\boldsymbol{\mu}_e^{(i)} \sim p\left(\boldsymbol{\mu}_e^{(i)}|\mathbf{p}^{(i)}, \boldsymbol{\sigma}^{(i)}, \mathbf{s}^{(i-1)}, \sigma_e^{2^{(i-1)}}, \mathbf{y}\right)$

(d) $\sigma_e^{2^{(i)}} \sim p\left(\sigma_e^{2^{(i)}}|\mathbf{p}^{(i)}, \boldsymbol{\sigma}^{(i)}, \mathbf{s}^{(i-1)}, \boldsymbol{\mu}_e^{(i)}, \mathbf{y}\right)$

(e) $\mathbf{a}^{(i)} \sim p\left(\mathbf{a}^{(i)}|\mathbf{p}^{(i)}, \boldsymbol{\sigma}^{(i)}, \mathbf{s}^{(i-1)}, \boldsymbol{\mu}_e^{(i)}, \sigma_e^{2^{(i)}}, \mathbf{y}\right)$

(f) $\mathbf{s}^{(i)} \sim p\left(\mathbf{s}^{(i)}|\mathbf{a}^{(i)}, \mathbf{p}^{(i)}\right)$

3. Increment i. If the required number of iterations has not been reached then return to step 2.

Our algorithm differs from the standard Gibbs sampler in one important respect. Steps 2a thru 2d are simulations from the conditionals corresponding to the marginalised posterior $p\left(\mathbf{p}, \boldsymbol{\sigma}, \mathbf{s}, \sigma_e^2, \mu_e|\mathbf{y}\right)$, rather than the full conditionals employed by the standard Gibbs sampler (note that the marginalised posterior can be obtained from the full posterior by integrating out $\mathbf{a}$). It can be easily verified that this modified kernel does indeed satisfy the required conditions for convergence, and in practice we have found it to have superior convergence qualities.

Blocking of parameters exhibiting strong dependency relationships tends to improve convergence of the Gibbs sampler. Consequently, we have implemented blocking schemes where possible; steps 2a and 2b each represent k univariate simulation steps, while steps 2e and 2f represent block simulations.

The joint simulation in step 2f is achieved using a Forward-Backward recurrence algorithm (similar to [11]). This sampling step could alternatively be replaced by a series of univariate samples from the reduced (marginalised) conditional distribution, thereby removing the need to sample for $\mathbf{a}$, but we have found that little convergence gain is achieved for the increased computational load imposed.

3.2. Variable Dimension Case

We have shown that the fixed dimension problem can be treated using standard MCMC techniques. In reality the length of the base sequence is unknown, and the algorithm must be extended to incorporate model uncertainty.

In a Bayesian setting, the key quantity of interest for model selection is $p(k|\mathbf{y})$ where k denotes one element of the entire model space, K. Application of Bayes' theorem yields:

$$p(k|\mathbf{y}) = \frac{\int p(\mathbf{y}|k, \boldsymbol{\theta})\, p(\boldsymbol{\theta}|k)\, p(k)\, d\boldsymbol{\theta}}{p(\mathbf{y})} \tag{6}$$

In this, and the majority of other models, the required integration cannot be performed analytically.

We can, however, implicitly perform this integration using a Reversible Jump Metropolis Hastings algorithm [8]. The Reversible Jump algorithm extends the standard Metropolis Hastings algorithm to the variable dimension case, constructing a variable dimension Markov chain with stationary distribution $p(\boldsymbol{\theta}_k, k|\mathbf{y})$. It may be summarised as follows:

1. Initialise parameters
2. Make a proposal from $q\left(\boldsymbol{\theta}_{k'}|k', \boldsymbol{\theta}_k, \mathbf{y}\right) J\left(k \longrightarrow k'\right)$ where q and J may be arbitrarily chosen proper distributions, $\boldsymbol{\theta}_k$ represents the set of parameters corresponding to model k.
3. Accept with probability $\alpha =$

$$\min\left\{1, \underbrace{\frac{p\left(k', \boldsymbol{\theta}'_{k'}|\mathbf{y}\right)}{p(k, \boldsymbol{\theta}_k|\mathbf{y})}}_{\text{target distn.}} \underbrace{\frac{q\left(\boldsymbol{\theta}_k|k, \boldsymbol{\theta}'_{k'}\mathbf{y}\right)}{q\left(\boldsymbol{\theta}'_{k'}|k', \boldsymbol{\theta}_k, \mathbf{y}\right)} \frac{J\left(k' \longrightarrow k\right)}{J\left(k \longrightarrow k'\right)}}_{\text{proposal distn.}}\right\}$$

4. Repeat from step 2 until convergence to target distribution

The selection of an appropriate proposal density for moves between models is essential for efficient implementation of the algorithm. We use a birth-death, split-merge type proposal structure, with proposals generated dependent upon the data $\mathbf{y}$ and the current state $(\boldsymbol{\theta}_k, k)$. The split and merge moves work by matching the 0^{th}, 1^{st}, and 2^{nd} order moments of the proposal with those of the current state of the chain (similar to [12]). To ensure sensible proposals, birth moves are made using a peak detection algorithm, combined with a random component to ensure irreducibility of the chain (required for convergence) (similar to [2], [14], [15]).

Fixed dimension moves are made using the algorithm described in the previous section.

4. RESULTS

We confine ourself to consideration of two data sets. The first exhibits some ambiguity as to the number of bases, and the second shows more complicated peak shapes, and significant leakage between lanes.

In the first case, the peak shape was assumed Gaussian, and the algorithm was run for 2000 iterations. Both the data set, and two realisations of the called sequence from the posterior distribution are shown in figure 1. We see that there is little doubt according to the posterior distribution as to whether there are 6 or 7 peaks responsible for the electropherogram, the 6 system being far more likely. This is largely a consequence of our prior structure's preference for approximately evenly spaced peaks. We emphasise that other current sequencing algorithms do not provide multiple interpretations of the data, nor error metrics inherently defined by the relevant base-calling model.

The second data set was generated using a different sequencing chemistry, and exhibits heavy tailed peak shapes, significant baseline noise, and large secondary peaks corresponding to leakage between lanes. Simple Gaussian peak shapes are not adequate to model this data and, if used, result in overestimation of the sequence length. Instead, a sum of Gaussians representation was used to represent each peak (in this case, two Gaussians), such that the heavy tail of the shape was represented. We suggest that this representation is sufficiently flexible for any sequencing chemistry, noting that if little information is available, the peak shape can be estimated much like any other parameter. The resultant base call after 2000 iterations is shown in figure 2.

Although not shown due to space considerations, the algorithm has also been run on far longer and more complicated data sets exhibiting sequencing chemistry dependent effects. Here too we have had success, largely because our algorithm is able to track non-stationarity in the model parameters, where traditional heuristic algorithms relying on peak detection algorithms often fail. For example, poor tracking of increasing peak spread across the sequence will often lead to the an overestimation of the number of bases generating peaks.

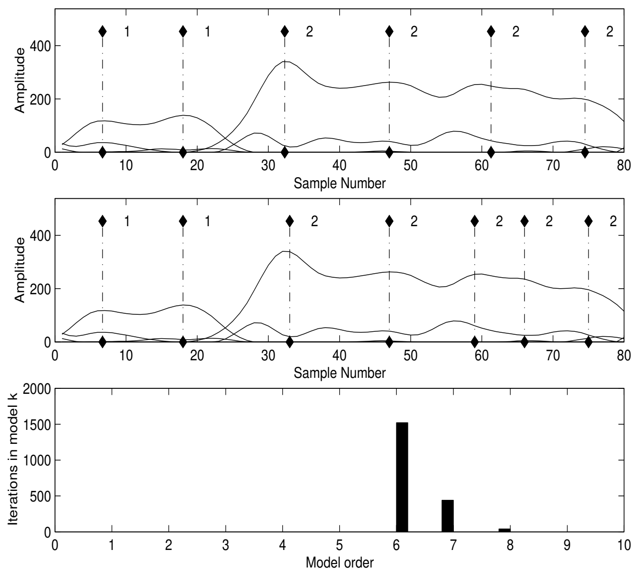

Figure 1: Top two plots: Vertical dotted lines represent the estimated peak positions, the numbers next to the diamonds denote the associated base (1=A, 2=G, 3=C, 4=T). The top plot denotes a 6 base explanation of the system, the middle plot a 7 base explanation. Bottom plot: a histrogram denoting the time spent by the Markov chain in each of the 6,7, and 8 base spaces

5. CONCLUSIONS

We have described the problems associated with automated processing of DNA sequence data in some detail, particularly noting the non-stationarity inherent in the system, and have observed that existing heuristic algorithms do not adequately model the data.

We have therefore developed a new probabilistic algorithm, set in a Bayesian framework, which we believe better models the underlying process and is able to flexibly incorporate prior information about the system. The problem is essentially reduced to a variable dimension parameter estimation problem, to which we apply a Reversible Jump Metropolis Hastings sampler.

Preliminary results from our algorithm are promising, indicating tangible improvement over the most widely used software, Phred [4], for a range of different data.

6. REFERENCES

[1] ABI. *ABI Prism, DNA sequencing analysis software, User's Manual*. Perkin Elmer Applied Biosystems, Foster City, CA, 1996.

[2] C. Andrieu and A. Doucet. Joint Bayesian model selection and estimation of noisy sinusiods via Reversible Jump MCMC. To appear in IEEE Trans. Signal Processing.

[3] A.J. Berno. A graph theoretic approach to the analysis of DNA sequencing data. *Genome Research*, 6:80–91, 1996.

[4] B. Ewing, L. Hillier, M. Wendl, and P. Green. Base-calling of automated sequencer traces using Phred. I. Accuracy Assessment. *Genome Research*, 8:175–185, 1998.

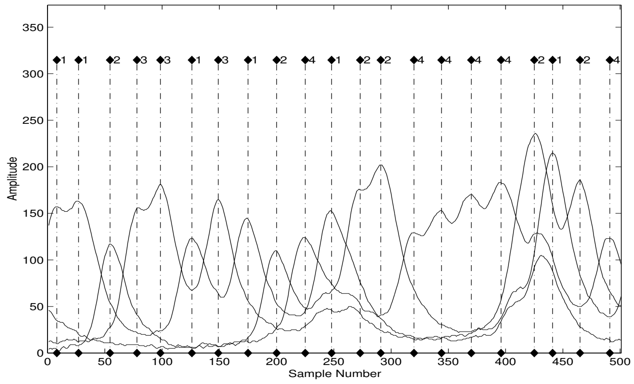

Figure 2: Data exhibiting heavy tailed peak shapes, baseline noise, and leakage between lanes

[5] S. Geman and D. Geman. Stochastic Relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Patt. Anal. and Mach. Int.*, 6(6):721–741, 1984.

[6] M. Giddings et al. An adaptive object oriented strategy for base calling in DNA sequence analysis. *Nucleic Acids Research*, 21(19):4530–4540, 93.

[7] J. Golden, E. Garcia, and C. Tibbetts. Evolutionary optimization of a neural network-based signal processor for photometric data from an automated DNA sequencer. In *Evolutionary programming IV. Proceedings of the Fourth Annual Conference on Evolutionary Programming*, pages 579–601, 1995.

[8] P.J. Green. Reversible Jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[9] W. Hastings. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[10] Beckman Research Institute. Optimal data looks like this. On World Wide Web - bricoh.coh.org/ DNAseq/trouble.html.

[11] F. Muri, D. Chauveau, and D. Cellier. Convergence assessment in latent variable models: DNA applications. In C.P. Robert, editor, *Discretization and MCMC Convergence Assessment*, volume 135 of *Lecture Notes in Statistics*. Springer, 1998.

[12] S. Richardson and P.J. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Roy. Stat. Soc. Series B*, 59(4):731–758, 1997.

[13] F. Sanger, S. Nicklen, and A.R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci*, 74:5463–5467, 1977.

[14] P.T. Troughton and S.J. Godsill. A Reversible Jump sampler for autoregressive time series. In *Proceedings of IEEE ICASSP*, volume IV, pages 2257–2260, 1998.

[15] P.J. Walmsley, S.J. Godsill, and P.J.W. Rayner. Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters. In *Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1999.