Efficient Bayesian inference for harmonic models via adaptive posterior factorization

Emmanuel Vincent, Mark Plumbley

Efficient Bayesian inference for harmonic models via adaptive posterior factorization

Emmanuel Vincent [1,*]

METISS group
IRISA-INRIA
Campus de Beaulieu, 35042 Rennes Cedex, France

Mark D. Plumbley

Centre for Digital Music, Department of Electronic Engineering
Queen Mary University of London
Mile End Road, London E1 4NS, United Kingdom

Abstract

Harmonic sinusoidal models are an essential tool for music audio signal analysis. Bayesian harmonic models are particularly interesting, since they allow the joint exploitation of various priors on the model parameters. However existing inference methods often rely on specific prior distributions and remain computationally demanding for realistic data. In this article, we investigate a generic inference method based on approximate factorization of the joint posterior into a product of independent distributions on small subsets of parameters. We discuss the conditions under which this factorization holds true and propose two criteria to choose these subsets adaptively. We evaluate the resulting performance experimentally for the task of multiple pitch estimation using different levels of factorization.

Key words: Bayesian inference, harmonic model, adaptive factorization, posterior dependence

* Corresponding author.
Email addresses: emmanuel.vincent@irisa.fr (Emmanuel Vincent), mark.plumbley@elec.qmul.ac.uk (Mark D. Plumbley).
[1] E. Vincent was funded by EPSRC grant GR/S75802/01.

1 Introduction

Music and speech involve different types of sounds, including periodic, transient and noisy sounds. Short-term stationary periodic sounds composed of sinusoidal partials at harmonic or near-harmonic frequencies are perceptually essential, since they contain most of the energy of musical notes and vowels. Harmonicity means that at each instant the frequencies of the partials are multiples of a single frequency called the fundamental frequency. Estimating the periodic sounds underlying a given signal, *i.e.* estimating their fundamental frequencies and the amplitudes and phases of their partials, is required or useful for many applications, such as speech prosody analysis [1], multiple pitch estimation and instrument recognition [2] and low bit-rate compression [3]. This problem is particularly difficult for polyphonic signals, *i.e.* signals containing several concurrent periodic sounds, since different periodic sounds may exhibit partials overlapping at the same frequencies.

Existing methods for polyphonic fundamental frequency estimation are often based on one of two approaches [2]: either validation of fundamental frequency candidates given by the peaks of a short-term auto-correlation function [4–6] or inference of the hidden states of a probabilistic model of the signal short-term power spectrum based on learned template spectra [7–9]. These approaches have achieved limited performance on complex polyphonic signals so far [2,6]. Moreover neither approach provides estimates for the amplitudes and phases of the partials, which are needed for musical instrument recognition or low bit-rate compression.

A promising way to address these issues is to rely on a probabilistic model of the signal waveform incorporating various prior knowledge. Two families of such models have been proposed in the literature for music signals. One family introduced in [10,11] models each musical note signal in state-space form by a discrete fundamental frequency and a fixed number of damped oscillators at harmonic frequencies with independent transition noises. Decoding is achieved either via linear Kalman filtering or variational approximation [12], depending whether the damping factors are fixed or subject to additional transition noises. These inference methods restrict the prior distribution of the transition noises to be Gaussian or from a class of conjugate priors [13] respectively. Another family of models described in [14–16] represents musical note signals by continuous fundamental frequency, amplitude and phase parameters, inferred using Markov Chain Monte Carlo (MCMC) methods [13]. These methods are applicable to all prior distributions in theory, but tend to be computationally demanding in practice. Thus the chosen priors are mostly motivated by computational issues [16]. In particular, the amplitudes of the partials are modeled by independent uniform priors or by conjugate zero-mean Gaussian priors, so that analytical marginalization can be performed.

For both families of models, the above priors exhibit some differences with the empirical parameter distributions. In particular, they do not penalize partials with zero amplitude. This typically leads to missing estimated notes for signals composed of several notes at integer fundamental frequency ratios [14,16] or to erroneous fundamental frequency estimates equal to a multiple or a submultiple of the true fundamental frequencies [16]. To help solving these limitations, we recently designed a probabilistic harmonic model involving priors motivated by empirical parameter distributions and proposed a variant of the diagonal Laplace method for fast inference [3], since analytical marginalization was no longer feasible with these priors.

In this article, we propose an alternative fast inference method for probabilistic harmonic models, based on approximate factorization of the joint posterior into a product of independent distributions on subsets of parameters. This method is designed for models of the form described in [14–16,3], involving explicit frequency, amplitude and phase parameters. It is generic, in that it can be applied to a wide range of priors, and adaptive, since the level of factorization depends on the observed signal and the hypothesized notes. This constitutes a crucial difference compared to variational approximation methods, where the terms of the factorization are fixed *a priori* and their parameters can only be computed for certain classes of priors. We complete our preliminary work [17] by discussing the extension of this method to nongaussian likelihood and alternative model structures, investigating a new criterion for the choice of the parameter subsets and providing a detailed experimental evaluation.

The structure of the rest of the article is as follows. In section 2, we present a possible Bayesian network structure for harmonic models and make some mild assumptions about the parameter priors. Then, we describe the proposed inference method in section 3 and extend it to alternative model structures. In section 4, we evaluate its performance for the task of multiple pitch estimation on short time frames. We conclude in section 5 and suggest some perspectives for future research.

2 Assumptions about the model

The harmonic models in [14–16,3] are variations of the same concept. They all represent the observed music signal as a sum of note signals, each composed of several sinusoidal partials parametrized by a sequence of random variables spanning successive time frames. However, the chosen variables and their conditional dependency structure are slightly different for each model. For the sake of clarity, we first discuss our approach for the model structure in [3], which involves fewer variables. Also, we consider each signal frame separately,

thus omitting temporal dependencies. Such dependencies could be taken into account by replacing the priors over the variables in a frame by conditional priors given previous frames.

2.1 Bayesian network structure

On each time frame, the model described in [3] exhibits the four-layer Bayesian network structure shown in Figure 1. The observed signal frame $x(t)$ is assumed to be obtained by windowing the whole signal with a window $w(t)$ of length T. Each layer models $x(t)$ at a different abstraction level.

The bottom layer represents the active notes in this frame on a discrete pitch scale. In western music, the fundamental frequency f_p of each note expressed relatively to the sampling frequency F_s may vary across frames but remains close to a discrete pitch of the form

$$\mu_p = \frac{440}{F_s} 2^{\frac{p-69}{12}} \tag{1}$$

where p is an integer on the MIDI semitone scale. Assuming no unison, *i.e.* that several notes corresponding to the same discrete pitch cannot be present at the same time, each discrete pitch p is associated with a binary activity state S_p determining whether a note with that pitch is active or not. The number of active notes and their pitches are thus represented by the activity state vector $S = \{S_p : p_{\text{low}} \leq p \leq p_{\text{high}}\}$ where p_{low} and p_{high} are the lowest and highest pitches among possible instruments.

The signal $s_p(t)$ corresponding to each active note is then defined in the middle layers for $0 \leq t \leq T-1$ by

$$s_p(t) = w(t) \sum_{m=1}^{M_p} a_{pm} \cos(2\pi m f_p t + \phi_{pm}) \tag{2}$$

where f_p, a_{pm} and ϕ_{pm} are respectively its normalized fundamental frequency and the amplitude and phase of its m-th harmonic partial. The number of partials M_p is constrained as a function of the note pitch p to

$$M_p = \min\left(\frac{1}{2\mu_p}, M_{\max}\right) \tag{3}$$

so that the partials fill the whole observed frequency range up to a maximum number of partials $M_{\max}$. The amplitudes of the partials are assumed to depend on an amplitude scale factor r_p accounting for the total power of the note. The vectors of frequency, scale factor, amplitude and phase parameters for all notes are denoted respectively by $f = \{f_p : S_p = 1\}$, $r = \{r_p : S_p = 1\}$,

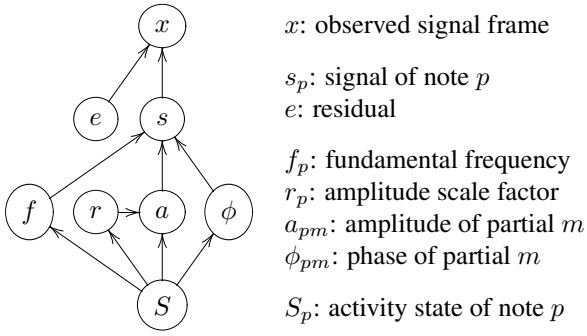

Fig. 1. Bayesian network structure of the harmonic model in [3] on one signal frame. Circles denote vector random variables (some of variable size) and arrows conditional dependencies.

$a = \{a_{pm} : S_p = 1, 1 \leq m \leq M_p\}$ and $\phi = \{\phi_{pm} : S_p = 1, 1 \leq m \leq M_p\}$. Finally, the observed signal is modeled in the top layer as

$$x(t) = \sum_{p \text{ s.t.} S_p = 1} s_p(t) + e(t) \tag{4}$$

where $e(t)$ is the residual.

2.2 Assumptions about the parameter priors

The inference method proposed below is valid given some mild assumptions about the parameter priors. Classically, we assume that the fundamental frequencies f_p of different notes p and the phases ϕ_{pm} of different partials (p, m) are independent *a priori* and that the amplitudes a_{pm} of different partials are independent *a priori* given the scale factors r_p. We also assume that the prior distribution of each fundamental frequency f_p is close to zero outside the interval $[2^{-1/12}\mu_p, 2^{1/12}\mu_p]$, so that it enforces proximity to the underlying discrete pitch. Finally, we make the hypothesis that the residual $e(t)$ has a continuous distribution and that its values at distinct frequencies are independent *a priori*, so that the likelihood $P(x|f, a, \phi)$ factors as

$$P(x|f, a, \phi) = \prod_{\nu=0}^{T-1} P(E_\nu) \tag{5}$$

where E_ν are the Discrete Fourier Transform (DFT) coefficients of $e(t)$.

Note that the ubiquitous time-domain Gaussian i.i.d. distribution satisfies this hypothesis, since it is equivalent to a Gaussian i.i.d. distribution on the DFT coefficients $P(x|f, a, \phi) = (2\pi\sigma^2)^{-T/2} \prod_{\nu=0}^{T-1} \exp(-|E_\nu|^2/(2\sigma^2))$. A more general distribution in the form of (5) and of particular interest in the following is the

frequency-weighted Gaussian [3]

$$P(x|f,a,\phi) = (2\pi\sigma^2)^{-T/2} \prod_{\nu=0}^{T-1} \exp\left(-\frac{\gamma_\nu |E_\nu|^2}{2\sigma^2}\right) \tag{6}$$

where γ_ν are constant positive weights. This can be rewritten as $P(x|f,a,\phi) = (2\pi\sigma^2)^{-T/2} \exp(-\|e\|_\gamma^2/(2\sigma^2))$ where $\|e\|_\gamma^2 = \sum_{\nu=0}^{T-1} \gamma_\nu |E_\nu|^2$ is the squared weighted Euclidean norm of the DFT coefficients.

3 Bayesian inference via adaptive posterior factorization

Harmonic models are typically employed to solve the multiple pitch estimation task, which consists of estimating the number of active notes and their pitches on each time frame. In the present framework, this task translates into finding the Maximum *A Posteriori* (MAP) activity state vector $\hat{S} = \arg\max P(S|x)$, which is achieved by trying a number of candidate vectors S, computing their posterior probabilities $P(S|x)$ and selecting the largest. These probabilities are defined by

$$P(S|x) = \int P(S,f,r,a,\phi|x)\, df\, dr\, da\, d\phi \tag{7}$$

where the joint posterior $P(S,f,r,a,\phi|x)$ is given by Bayes law

$$P(S,f,r,a,\phi|x) \propto P(x|f,a,\phi)P(\phi|S)P(a|r,S)P(r|S)P(f|S)P(S). \tag{8}$$

In the following, we focus on the computation of the integral in (7), which is known as the Bayesian marginalization problem [12]. We briefly recall some existing integration methods, then introduce the proposed method in a simple context and extend it to a more general context later on.

3.1 Sampling-based vs. full factorization-based integration

A number of sampling techniques are available to compute such integrals [12]. However they appear unsatisfactory in this context. Numerical integration on a uniform grid is accurate for distributions of a few parameters, but intractable here since the number of parameters is typically of the order of one hundred. Integration via importance sampling [18] is computationally demanding, since the variance of the importance weights, which is proportional to that of the estimate, increases sharply with the number of parameters [12]. Sampling of the joint posterior via reversible jump MCMC [13] is also demanding [16].

Fast inference can be achieved at the cost of lower accuracy by estimating the MAP parameter values $(\hat{f},\hat{r},\hat{a},\hat{\phi}) = \arg\max_{f,r,a,\phi} P(S,f,r,a,\phi|x)$ associated

with each candidate activity state vector S using some nonlinear optimization algorithm and approximating the joint posterior around these values by a simpler distribution which can be integrated analytically or by tabulation. The fastest techniques include the diagonal Laplace approximation [19], which relies on full factorization of the posterior into a product of parameter-wise univariate Gaussian distributions, and its variant proposed in [3] with a specific univariate nongaussian distribution for the phase parameters. The full Laplace approximation [19] performs poorly here due to unbounded integration over the phase parameters [17].

The proposed inference method aims to bridge the gap between sampling-based and full factorization-based techniques by partially factoring the joint posterior into a product of distributions over subsets of a few parameters and integrating these distributions via sampling. Various levels of factorization can be obtained depending on the MAP parameter values.

3.2 Conditional posterior factorization over the partials

For simplicity, let us assume initially that the harmonic partials of the hypothesized active notes have "different enough" frequencies and that the likelihood is a frequency-weighted Gaussian as in (6). These two assumptions are relaxed later on. The former is generally true for a single hypothesized active note, but almost never for several active notes. Mathematically, it leads to the assumption that the windowed complex sinusoidal signals

$$z_{pm}(t) = w(t)e^{2i\pi m f_p t} \tag{9}$$

corresponding to different partials are mutually orthogonal

$$\langle z_{pm}, z_{p'm'}\rangle_\gamma = 0 \quad \forall (p,m) \neq (p',m') \tag{10}$$

according to the dot product $\langle .,. \rangle_\gamma$ consistent with the weighted Euclidean norm $\|.\|_\gamma$ defined in Section 2.2. This dot product is defined for two signals $z(t)$ and $z'(t)$ by

$$\langle z, z' \rangle_\gamma = \sum_{\nu=0}^{T-1} \gamma_\nu Z_\nu \bar{Z}'_\nu \tag{11}$$

where Z_ν and Z'_ν are the DFT coefficients of $z(t)$ and $z'(t)$ and $\bar{Z}'_\nu$ is the complex conjugate of Z'_ν. The orthogonality property (10) formalizes the fact that partials with "different enough" frequencies have almost disjoint frequency supports and can be assumed to hold true for all possible frequency weights γ_ν. When the frequencies of the partials are not too close to Nyquist, the negative frequency sinusoidal signals $\bar{z}_{pm}(t) = w(t)e^{-2i\pi m f_p t}$ are also orthogonal to their positive counterparts: $\langle z_{pm}, \bar{z}_{p'm'}\rangle_\gamma = 0$ for all (p,m) and (p',m'). The observed signal $x(t)$ can then be decomposed into a sum of sinusoidal signals

at the frequencies of the hypothesized partials by orthogonal projection onto the two-dimensional subspaces spanned by $(z_{pm}, \bar{z}_{pm})$

$$x(t) = \frac{1}{2}\sum_{p,m} \tilde{a}_{pm}(e^{i\tilde{\phi}_{pm}} z_{pm}(t) + e^{-i\tilde{\phi}_{pm}} \bar{z}_{pm}(t)) + \tilde{e}(t). \tag{12}$$

The projection coefficients given by

$$\tilde{a}_{pm} e^{i\tilde{\phi}_{pm}} = 2\frac{\langle x, z_{pm}\rangle_\gamma}{\|z_{pm}\|_\gamma^2} \tag{13}$$

represent the amplitude and phase values of each partial leading to the minimum-norm residual $\tilde{e}(t)$. Given hypothesized values a_{pm} and ϕ_{pm}, the corresponding residual $e(t)$ can be decomposed as a sum of mutually orthogonal terms

$$\begin{aligned} e(t) = \frac{1}{2}\sum_{p,m} \left(\tilde{a}_{pm}e^{i\tilde{\phi}_{pm}} - a_{pm}e^{i\phi_{pm}}\right) z_{pm}(t) \\ + \left(\tilde{a}_{pm}e^{-i\tilde{\phi}_{pm}} - a_{pm}e^{-i\phi_{pm}}\right) \bar{z}_{pm}(t) + \tilde{e}(t). \end{aligned} \tag{14}$$

The squared norm of the residual then equals by analytical computation

$$\|e\|_\gamma^2 = \sum_{p,m} D_{pm} + D_0 \tag{15}$$

with $D_0 = \|\tilde{e}\|_\gamma^2$ and

$$D_{pm} = \frac{1}{2}\|z_{pm}\|_\gamma^2 \left((a_{pm} - \tilde{a}_{pm})^2 + 4\tilde{a}_{pm}a_{pm}\sin^2\frac{\phi_{pm} - \tilde{\phi}_{pm}}{2}\right). \tag{16}$$

Using (8) and the relationship between $P(x|f, a, \phi)$ and $\|e\|_\gamma^2$, this leads to the exact factorization of the joint posterior into a product of partial-wise bivariate conditional distributions over amplitude and phase parameters

$$\begin{aligned} P(S, f, r, a, \phi|x) \propto P_0(x, f)P(r|S)P(f|S)P(S) \\ \times \prod_{p,m} P_{pm}(a_{pm}, \phi_{pm}; x, f_p)P(a_{pm}|r_p)P(\phi_{pm}) \end{aligned} \tag{17}$$

where $P_0(x, f) = (2\pi\sigma^2)^{-T/2}\, e^{-D_0/(2\sigma^2)}$ depends on f only and $P_{pm}(a_{pm}, \phi_{pm}; x, f_p) = \exp(-D_{pm}/(2\sigma^2))$ is a bivariate parametric distribution that can be quickly computed, since it depends on three hyper-parameters only: $\|z_{pm}\|_\gamma^2$, $\tilde{a}_{pm}$ and $\tilde{\phi}_{pm}$. The top part of Figure 2 illustrates the validity of this factorization.

Denoting by $(\hat{a}, \hat{\phi}) = \arg\max_{a,\phi} P(S, f, r, a, \phi|x)$ the vectors of estimated MAP amplitude and phase parameter values associated with S, f and r and by $\hat{a}_{\overline{pm}} = \{\hat{a}_{p'm'} : (p', m') \neq (p, m)\}$ and $\hat{\phi}_{\overline{pm}} = \{\hat{\phi}_{p'm'} : (p', m') \neq (p, m)\}$ the same vectors minus one coefficient corresponding to partial (p, m), the above

expression can be equivalently rewritten as proved in Appendix A as

$$P(S, f, r, a, \phi|x) = P(S, f, r, \hat{a}, \hat{\phi}|x) \prod_{p,m} \frac{P(a_{pm}, \phi_{pm}|S, f, r_p, \hat{a}_{\overline{pm}}, \hat{\phi}_{\overline{pm}}, x)}{P(\hat{a}_{pm}, \hat{\phi}_{pm}|S, f, r_p, \hat{a}_{\overline{pm}}, \hat{\phi}_{\overline{pm}}, x)}. \quad (18)$$

This equation holds under the assumptions that the likelihood is frequency-weighted Gaussian and that the partials of the hypothesized active notes are orthogonal signals. It admits the following interpretation: the first term is the maximum of the joint posterior with fixed S, f and r and the quotient terms describe the decrease of this posterior around its maximum as proportional to the posterior distribution of the parameters of each partial with the parameters of other partials being fixed. In theory, this equation holds also for any value of $\hat{a}$ and $\hat{\phi}$ distinct from the actual MAP parameter values.

In practice, perfect orthogonality never happens. Nevertheless, this equation remains approximately valid under the more general assumptions that the partials of the hypothesized active notes have "different enough" frequencies and that the likelihood satisfies (5), although quick computation of the quotient terms by orthogonal projection is not feasible anymore with nongaussian likelihood. Indeed, when the amplitude parameters are not too far from their MAP values, the DFT coefficients of each partial signal are close to zero except for a few DFT bins ν around the frequency of that partial. The sets of bins associated with different partials are disjoint. Therefore the likelihood $P(E_\nu)$ in a given bin depends mostly on the parameters of a single partial, which leads to (18) after simple analytical computation. Note that this factorization holds as soon as the MAP amplitude values are not grossly overestimated, otherwise the frequency support of different partials might overlap due to secondary lobes. In particular, it holds when the joint posterior is multimodal and a local maximum was estimated instead of the global maximum.

3.3 Conditional posterior factorization over subsets of partials

In the general case where some partials of the hypothesized active notes may have close frequencies, the terms of (18) can still be computed but this equation may not hold true anymore, as shown in the middle part of Figure 2. It is however possible to group partials into disjoint subsets such that partials from different subsets have "different enough" frequencies. These subsets can be iteratively created as follows: a partial (p, m) is assigned to a previously created subset g if there exists a partial $(p', m') \in g$ such that

$$|mf_p - m'f_{p'}| \leq f_{\max} \quad (19)$$

where $f_{\max}$ is a manually set frequency threshold, otherwise it forms a new singleton subset. Provided that the likelihood satisfies (5) and that the MAP am-

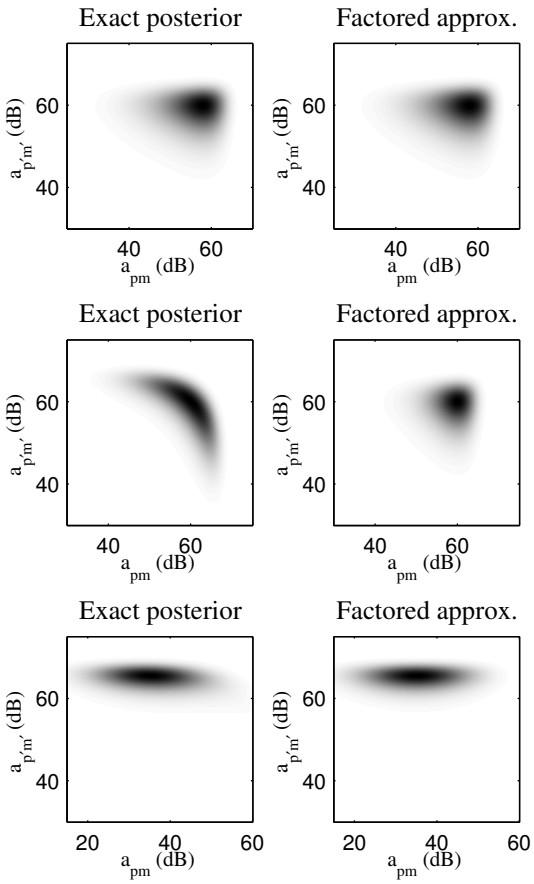

Fig. 2. Joint posterior distribution $P(a_{pm}, a_{p'm'}|S, \hat{f}, \hat{r}, \hat{a}_{\overline{pmp'm'}}, \hat{\phi}, x)$ of the amplitudes of two partials given the MAP values of other parameters, using the priors defined in [3] and assuming 60 dB ground truth amplitudes with no other partials at the same frequencies. Dark areas denote high probability. Top: partials at different frequencies with mean prior amplitudes of 50 dB and 60 dB. Middle: partials at the same frequency with identical mean prior amplitudes of 60 dB. Bottom: partials at the same frequency with mean prior amplitudes of 40 dB and 60 dB. The posterior dependence between a_{pm} and $a_{p'm'}$ equals 0, 0.99 and 0.093 bits respectively.

plitude values are not grossly overestimated, similar arguments as above lead to the approximate factorization of the posterior into a product of multivariate conditional distributions over subsets of amplitude and phase parameters $a_g = \{a_{pm}, (p,m) \in g\}$ and $\phi_g = \{\phi_{pm}, (p,m) \in g\}$

$$P(S, f, r, a, \phi|x) \approx P(S, f, r, \hat{a}, \hat{\phi}|x) \prod_g \frac{P(a_g, \phi_g|S, f, r, \hat{a}_{\overline{g}}, \hat{\phi}_{\overline{g}}, x)}{P(\hat{a}_g, \hat{\phi}_g|S, f, r, \hat{a}_{\overline{g}}, \hat{\phi}_{\overline{g}}, x)}. \tag{20}$$

Each quotient term can be quickly computed by orthogonal projection in the particular case where the likelihood is frequency-weighted Gaussian.

A higher threshold $f_{\max}$ increases the accuracy of this equation, but also leads to larger subsets. In practice, it is often possible to obtain a factored expression

of similar accuracy with smaller subsets. Indeed there exist some situations where partials at close frequencies may still be associated with different subsets. An example of such a situation is given in the bottom part of Figure 2 and discussed in [17]. Denoting by the vectors y and y' two disjoint subsets of variables and by $\hat{y}$ and $\hat{y}'$ their estimated MAP values given the rest of the variables y'', we assess the accuracy of the approximation of the joint posterior distribution $P(y, y'|y'')$ by the factored distribution $P(y|\hat{y}', y'')P(y'|\hat{y}, y'')$ using the Kullback-Leibler divergence [12]

$$\mathcal{D}(y, y') = \int P(y, y'|y'') \log_2 \frac{P(y, y'|y'')}{P(y|\hat{y}', y'')P(y'|\hat{y}, y'')} \, dy \, dy'. \tag{21}$$

This quantity is always positive and equal to zero only when the approximation is exact. It can be seen as a measure of the local posterior dependence between y and y' expressed in bits. Indeed, it is analogous to mutual information [12], except that the marginal distribution of each variable is replaced here by its posterior distribution given the estimated MAP value of the other. This suggests that partials (p, m) and (p', m') are to be grouped in the same subset if

$$\mathcal{D}(\{a_{pm}, \phi_{pm}\}, \{a_{p'm'}, \phi_{p'm'}\}) \geq c_{\min} \tag{22}$$

where $c_{\min}$ is a manually set threshold. Compared to mutual information, this criterion is tractable for a wide range of distributions. However, it is less accurate when three or more partials overlap at a given frequency and the joint posterior is multimodal, since misestimation of the MAP parameter values of one partial may affect the estimated posterior dependence between the parameters of other partials. This may in turn affect the determination of the parameter subsets, hence the accuracy of (20).

Figure 3 depicts the distribution of posterior dependence between the parameters of two partials as a function of their frequency difference measured in number of DFT bins. The choice of the frame length T affects the measured frequency difference. However, the distribution for a given frequency difference remains roughly independent of the frame length. On average, posterior dependence tends to decrease with increasing frequency difference and exhibits smaller values for differences corresponding to certain zeroes of the DFT of the window $w(t)$. Most importantly, for a given frequency difference, posterior dependence values differing by up to three orders of magnitude can be observed. This shows that there exist many situations in practice where the parameters of two partials at close frequencies are much less dependent than average, so that the size of the parameter subsets can effectively be reduced. This figure can also be exploited to speed up the estimation of the subsets by avoiding the computation of the posterior dependence between partials whose frequency difference is above a certain threshold, chosen so that the posterior dependence is guaranteed to be larger than $c_{\min}$.

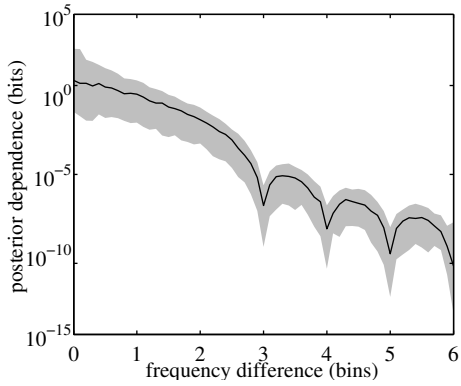

Fig. 3. Posterior dependence between the parameters of two partials as a function of their frequency difference measured in number of DFT bins. The black curve and the gray area denote respectively the median and the two-tailed 95th percentile of the values computed for all the data of Section 4.1.

3.4 An exploitable posterior factorization

The conditional factorization (20) can be exploited for numerical integration of the posterior, either by sampling on a uniform grid or by importance sampling. Indeed integration over amplitude and phase parameters can be achieved by multiplying lower dimension integrals over the parameters of each subset of partials. Using sampling on a uniform grid and denoting by N the number of grid points for each scalar variable, P the number of hypothesized notes, $M = \sum_p M_p$ their total number of partials and G the size of the largest subset of partials, this results in a maximum complexity of $\mathcal{O}(\frac{M}{G} N^{2P+2G})$. This is smaller than the complexity of $\mathcal{O}(N^{2P+2M})$ associated with straightforward integration of the joint posterior, but still intractable.

In order to get faster integration, additional parameter dependencies must be removed. A natural approach consists of computing the MAP values $(\hat{f}, \hat{r}, \hat{a}, \hat{\phi}) = \arg\max_{f,r,a,\phi} P(S, f, r, a, \phi|x)$ of all parameters, replacing the free fundamental frequency and scale factor parameters f and r in the quotient terms of (20) by their MAP values and factoring the first term of (20) over individual parameters f_p and r_p. This gives

$$P(S, f, r, a, \phi|x) \approx P(S, \hat{f}, \hat{r}, \hat{a}, \hat{\phi}|x) \prod_p \frac{P(f_p|S, \hat{f}_{\overline{p}}, \hat{a}, \hat{\phi}, x)}{P(\hat{f}_p|S, \hat{f}_{\overline{p}}, \hat{a}, \hat{\phi}, x)} \times \prod_p \frac{P(r_p|S, \hat{a}_p)}{P(\hat{r}_p|S, \hat{a}_p)} \prod_g \frac{P(a_g, \phi_g|S, \hat{f}, \hat{r}, \hat{a}_{\overline{g}}, \hat{\phi}_{\overline{g}}, x)}{P(\hat{a}_g, \hat{\phi}_g|S, \hat{f}, \hat{r}, \hat{a}_{\overline{g}}, \hat{\phi}_{\overline{g}}, x)}. \quad (23)$$

This equation allows approximate numerical integration of the posterior with a maximum complexity of $\mathcal{O}(\frac{M}{G} N^{2G})$. Although it is not straightforward to justify mathematically, this additional approximation appears experimentally

valid under the assumption that the prior distribution of fundamental frequency parameters enforces proximity to the underlying discrete pitches. For instance, the posterior dependence between fundamental frequencies and other parameters or between scale factors and other parameters for the data of Section 4.1 was above the best setting of $c_{\min}$ determined in that section in less than 4% of the cases, so that the error introduced by factorization of the joint posterior over fundamental frequency and scale factor parameters was generally smaller than that introduced by factorization over amplitude and phase parameters. Similarly to above, this equation may become inaccurate due to a different grouping of the partials when three or more partials overlap at a given frequency and the MAP amplitude and phase values $\hat{a}$ and $\hat{\phi}$ are misestimated. However, misestimation of the MAP fundamental frequencies $\hat{f}$ has little effect, since it affects mostly the grouping of upper frequency partials which are generally independent *a posteriori* due to their small amplitude.

3.5 Summary of the proposed inference method

To sum up, the proposed inference method is as follows. For each signal frame $x(t)$ and each candidate activity state vector S

(1) estimate the MAP parameter values $(\hat{f}, \hat{r}, \hat{a}, \hat{\phi})$ by nonlinear optimization
(2) either
- group the partials into disjoint subsets g according to the frequency difference criterion (19) or
- compute the posterior dependence criterion (22) between all partials with small frequency difference and group them into disjoint subsets g according to this criterion

(3) compute the integral of each term of the factored posterior (23) via numerical integration and multiply these integrals to obtain $P(S|x)$

Various algorithms can be used to address each step. In the following, the MAP parameter values were computed using the subspace trust region optimization algorithm implemented in Matlab's lsqnonlin function [2]. This algorithm was initialized with the parameter values $f_p = \mu_p$, $a_{pm} = \tilde{a}_{pm}$ and $\phi_{pm} = \tilde{\phi}_{pm}$ defined in (1) and (13). The posterior dependence was computed for all pairs of partials with frequency difference smaller than 2.5 bins by numerical integration on a uniform grid with 11 points per variable, that is about 1.5×10^4 samples per pair of partials. Each term of the factored posterior was subsequently integrated by sampling on a uniform grid with N points per variable, resulting in a total of $N_{\text{tot}} = N^{2P} + \sum_g N^{2|g|}$ samples per candidate activity state vector where $|g|$ denotes the number of partials in subset g. We also tried

[2] http://www.mathworks.com/access/helpdesk/help/toolbox/optim/ug/lsqnonlin.html

integration of these terms via importance sampling [18], but this did not significantly affect performance, although this led to an increase in computation time.

3.6 Extension to alternative model structures

The proposed marginalization could be extended to alternative harmonic model structures, such as those described in [14–16]. Indeed, the approximate posterior independence property of partials at different frequencies remains valid.

Among all structural differences, these models consider the number of partials per note M_p as a random variable subject to a certain prior. With narrow fundamental frequency priors as in [14,15], the proposed method can be directly applied to compute the integrals of the joint posterior for each value of M_p. Note that this results in little additional cost compared to fixed M_p. Indeed, when increasing or decreasing M_p by one, only one subset of partials needs to be updated, while the integral over the other subsets remains constant. With wider fundamental frequency priors as in [16], the posterior becomes multimodal with local maxima at all fundamental frequencies present in the signal and rational multiples of these. Amplitude and phase parameters then exhibit a strong dependence with fundamental frequency parameters. The proposed method can still be applied by splitting the fundamental frequency range into disjoint narrow bands, similar to the semitone bands considered above, and summing the integrals of the joint posterior within each band.

Another difference is that the models in [15,16] involve additional parameters, namely one global inharmonicity parameter and one spectral shape parameter per note in [15] and one local inharmonicity parameter per partial in [16]. The proposed method can be directly applied in the second case by grouping local inharmonicity parameters with amplitude and phase parameters from the same partials, yielding a maximum complexity of $\mathcal{O}(\frac{M}{G}N^{3G})$. We believe that it could also be applied in the first case after additional factorization of the joint posterior over global inharmonicity and spectral shape parameters. Indeed these parameters are physically similar to fundamental frequency and scale factor parameters and should exhibit a similar level of posterior dependence with other parameters.

Finally, the models in [14–16] describe the likelihood by a Gaussian whose variance is considered as a random variable. Although this distribution does not satisfy (5), the proposed method can still be applied after additional factorization of the posterior over this variance parameter. We believe that this factorization remains approximately accurate provided that the posterior distribution of the variance is unimodal and narrow.

4 Evaluation

The precision of the integral estimates obtained by the proposed marginalization method cannot be assessed for realistic signals, since ground truth integral values are not available. However, the aim of marginalization is often not to compute the exact values of the state posteriors $P(S|x)$, but rather to provide accurate multiple pitch estimation, that is to select the right MAP activity state vector $\hat{S}$. Therefore we evaluated the performance of the proposed method with respect to the latter task. The variant of the diagonal Laplace method employed in [3] was also evaluated for comparison.

4.1 Training data and evaluation procedure

The parameter priors were chosen as in [3], without assuming knowledge of the true number of active notes on each frame: the activity states S_p were modeled by Bernoulli priors, the fundamental frequencies f_p, the scale factors r_p and the amplitudes of the partials a_{pm} by log-Gaussian priors, the phases of the partials ϕ_{pm} by uniform priors and the likelihood by a frequency-weighted Gaussian. Note that the prior over a_{pm} helps to avoid partials with zero amplitude. The means and variances of these priors were learned on a subset of the RWC Musical Instrument Database [3] consisting of isolated notes from five wind instruments (flute, oboe, clarinet, trombone and bassoon) with MIDI pitches ranging from $p_{\text{low}} = 34$ to $p_{\text{high}} = 96$ and about 500 to 3000 signal frames per pitch.

For each frame, each active note in the MAP state vector $\hat{S}$ was considered to be correct if it was actually present in the test signal. Performance was then classically assessed by the F-measure $F = 2RP/(R+P)$ in percent, where the recall R is the ratio of the total number of correct notes over all frames divided by the true number of active notes and the precision P is the proportion of correct notes among the estimated active notes [20,6]. The computation time was measured for a Matlab 7.5 implementation on a 1.2 GHz dual core laptop.

4.2 Results with one-note and two-note signals

A first experiment was run on one-note and two-note single-frame signals generated by selecting and mixing isolated note signals from the five above wind instruments taken from the University of Iowa Musical Instrument Samples

[3] http://staff.aist.go.jp/m.goto/RWC-MDB/

database [4]. More precisely, the test set included 100 one-note signals spanning all discrete pitches from $p = 40$ to 87 and 100 two-note signals corresponding to all possible pitch intervals between 1 and 25 semitones with four different bass pitches $p = 40$, 47, 54 and 61. All signals were sampled at 22.05 kHz and cut to a single frame using a Hanning window $w(t)$ of length $T = 1024$ (46 ms). In order to avoid testing all possible activity state vectors S, 6 candidate vectors (3 with one active note and 3 with two active notes) were automatically pre-selected for each test signal as those minimizing the residual of the orthogonal projection of the observed magnitude spectrum onto the subspace spanned by the mean magnitude spectra of active notes, derived from the amplitude prior as explained in [3]. The thresholds $f_{\max}$ and $c_{\min}$ were varied between 0 and 2 bins and between 10^3 and $10^{-1.5}$ bits respectively, resulting in a variation of the maximum number of partials per subset from one to three. The average number N_{tot} of integration samples per test signal and per candidate was varied between 10^5 and 10^7.

With one-note signals, the proposed method gave perfect results for all settings of $f_{\max}$, $c_{\min}$ and N_{tot}, as expected by any reasonable fundamental frequency estimator. The method in [3] also provided perfect results with a faster average computation time of 0.55 s per candidate, mostly due to the optimization of the MAP parameters.

The results with two-note signals are depicted in Figure 4. The performance of the proposed method with many integration samples $N_{\text{tot}} = 10^7$ increases from F=93.4% (R=89.0%, P=98.3%) to F=96.9% (R=94.5%, P=99.5%) for both grouping criteria when the average number of partials per subset increases. This difference is statistically significant, as confirmed by a McNemar's p value [21] of 10^{-3}. By comparison, the method in [3] achieved a performance of F=92.6% (R=88.0%, P=97.8%), which is not statistically different from that of the proposed method with a single partial per subset. The best setting for the proposed method consists of using the posterior dependence criterion with $c_{\min} \simeq 10^0$ bits. Indeed, this criterion generally results in smaller subsets of partials, thus allowing a larger number N of integration samples per variable for a given total integration cost N_{tot}. The resulting increased accuracy translates into the fact that the performance curve with $N_{\text{tot}} = 10^5$ wanders less around the curve with $N_{\text{tot}} = 10^7$ for this criterion than for the frequency difference criterion. The above value of $c_{\min}$ is the largest one yielding maximum performance, resulting in as little as 7% of partials found within subsets of two partials for two-note candidates and no partials found within subsets of three or more. For this setting, the computation time with $N_{\text{tot}} = 10^5$ or equivalently $N = 15$ was equal to 1.0 s per candidate on average. This can be split into about 0.55 s for the optimization of the MAP parameters, 0.1 s for the computation of the posterior dependence between the partials and 0.35 s

[4] http://theremin.music.uiowa.edu/MIS.html

for the numerical integration of the terms of the factored posterior. This is much faster than previously reported computation times for MCMC methods with similar models, *e.g.* 1080 s per active note with $T = 6000$ using a 2.6 GHz dual core computer in [16], corresponding to about 800 s per test signal for two-note signals of length $T = 1024$ with our computer.

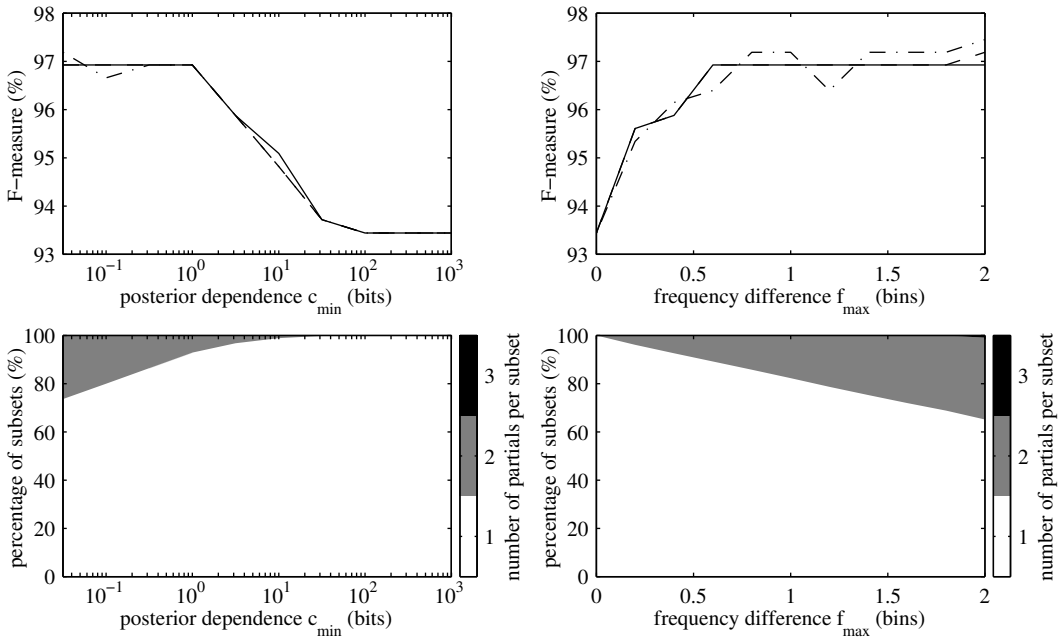

Fig. 4. Multiple pitch estimation results on two-note signals using adaptive posterior factorization based on posterior dependence (left) or frequency difference (right) between partials. Top: F-measure for various numbers of integration samples (plain: $N_{\mathrm{tot}} = 10^7$, dashed: $N_{\mathrm{tot}} = 10^6$, dash-dotted: $N_{\mathrm{tot}} = 10^5$). Bottom: Percentage of partials from two-notes candidates within subsets of size 1, 2 or 3. The percentage for subsets of size 3 equals 0.2% for $c_{\mathrm{min}} = 10^{-1.5}$ bits, 1% for $f_{\mathrm{max}} = 2$ bins and 0 in all other cases. All partials from one-note candidates form singleton subsets.

The pitch estimation errors made by the proposed method and the method in [3] are compared in Figure 5. Errors arise mostly in two situations well known to be difficult [2]: pitch intervals of 12, 19 or 24 semitones, corresponding to integer fundamental frequency ratios of 2, 3 or 4, and pitch intervals of 1 to 10 semitones with medium or low pitch bass. These errors can be explained respectively by the fact that all the partials of one note overlap with the partials of the other and that the frequency resolution is too small to distinguish multiple notes at low fundamental frequencies. The proposed method reduced the number of errors in both situations. In particular, correct estimation was achieved for some pitch intervals of 19 semitones and all intervals of 24 semitones, while such intervals would typically result in estimation errors for other models based on uniform or zero-mean Gaussian amplitude priors [14,16].

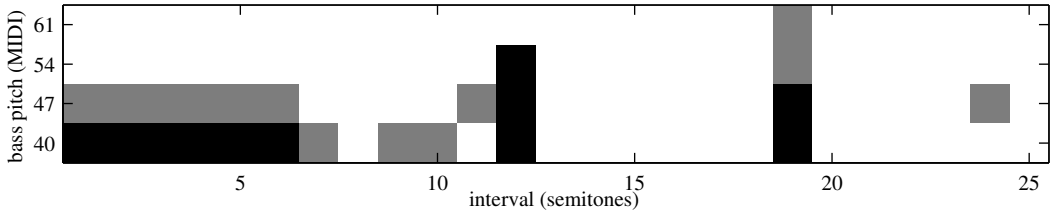

Fig. 5. Comparison of pitch estimation errors on two-note signals using adaptive posterior factorization with the best setting or the diagonal Laplace variant method in [3]. White, gray and black squares denote respectively accurate estimation for both methods, accurate estimation for the proposed method only and erroneous estimation for both methods. There are no cases where [3] was accurate but the proposed method was not.

4.3 Results with real-world signals

In order to estimate the potential performance of the proposed method on real-world data, a second experiment was run on the data for the multiple fundamental frequency estimation task of the 2007 Music Information Retrieval Evaluation Exchange (MIREX) [5]. These data consist of public and hidden excerpts of recordings and monophonic pitch annotations of individual instrument parts of a wind quintet by Beethoven. Four test signals with two to five instruments were generated by successively summing together the parts of flute, clarinet, bassoon, horn and oboe of the first 17 s of the public excerpt. These signals were then resampled at 22.05 kHz and framed with half-overlapping Hanning windows of length $T = 1024$ (46 ms). A variable number of candidate activity state vectors S with zero to five active notes was automatically pre-selected for each frame using the iterative state jump algorithm employed in [3], resulting in 23 candidates per frame on average. Inference was performed using the best settings determined above, namely $c_{\min} = 10^0$ bits and $N = 15$, leading to subsets of one to three partials.

The results are detailed in Table 1. On average, the proposed method ran in 17 h and achieved $F = 70.0\%$, while the method in [3] ran in 6 h and achieved $F = 68.7\%$. The similarity between these performance figures suggests that full posterior factorization can work nearly as well as adaptive posterior factorization in practice. One reason for this is that the pitch intervals for which an improvement was observed in the first experiment occur more rarely in this experiment. These figures are in line with the top ones reported at MIREX. For instance, the best method [6] achieved F=73% on hidden excerpts of the same data, while exploiting additional temporal priors.

[5] http://www.music-ir.org/mirex2007/

Table 1
Multiple pitch estimation results on real-world signals.

Method	Number of instruments	R (%)	P (%)	F (%)	Average time per candidate state (s)	Total time (h)
Adaptive posterior factorization	2	87.0	91.8	89.3	1.8	8
	3	60.1	86.2	70.9	5.2	24
	4	51.2	82.2	63.1	5.4	26
	5	42.2	85.4	56.5	5.5	28
Diagonal Laplace variant [3]	2	86.2	91.3	88.6	0.8	4
	3	59.3	85.9	70.1	1.4	6
	4	49.4	81.5	61.5	1.5	7
	5	40.4	84.1	54.6	1.3	6

5 Conclusion

We proposed a fast inference method for Bayesian harmonic models based on approximate factorization of the joint posterior into a product of distributions over disjoint parameter subsets and numerical integration of these distributions. A local posterior dependence criterion was exploited to determine relevant subsets. Although factorization based on this criterion is theoretically feasible for any Bayesian model, it does not necessarily provide small subsets, which are needed for subsequent numerical integration. The key property of harmonic models demonstrated here is that the parameters of partials with different frequencies are approximately independent *a posteriori*. Compared to classical inference methods such as MCMC and variational approximation, this method is tractable for a wide range of priors and allows a variable level of factorization depending on the observed signal and the hypothesized notes. However, the resulting marginal probability estimates are intrinsically less accurate. Hence we believe that it is most beneficial when classical methods are intractable due to the chosen priors. Our experiments suggest that approximate inference with priors motivated by empirical parameter distributions can provide better pitch transcription results than exact inference with standard priors motivated by computational issues.

To improve the accuracy of the factorization, it would be interesting to investigate transformations of the parameters resulting in a smaller posterior dependence. The minimization of the dependence between subsets of random variables described by a sequence of samples is known as the Independent Subspace Analysis (ISA) problem and can be solved in the case of linear transformations by Independent Component Analysis (ICA) followed by grouping of the transformed variables [22]. This approach could easily be combined with

subsequent integration based on importance sampling, and this may also allow other Bayesian models, which do not readily satisfy the posterior independence property, to benefit from the proposed inference method.

Acknowledgment

The authors would like to thank the anonymous reviewers for useful comments on the first version of this article. The first author also wishes to thank Cédric Févotte and Simon J. Godsill for motivating discussions about the influence of prior distributions on the performance of harmonic models and the practical use of MCMC methods.

A Proof of equation (18)

We start from the expression of joint posterior in (17). On the one hand, by evaluating (17) with $a = \hat{a}$ and $\phi = \hat{\phi}$, we obtain

$$P(S, f, r, \hat{a}, \hat{\phi}|x) \propto P_0(x, f)P(r|S)P(f|S)P(S) \times \prod_{p,m} P_{pm}(\hat{a}_{pm}, \hat{\phi}_{pm}; x, f_p)P(\hat{a}_{pm}|r_p)P(\hat{\phi}_{pm}). \quad \text{(A.1)}$$

After dividing (17) by (A.1), we get

$$\frac{P(S, f, r, a, \phi|x)}{P(S, f, r, \hat{a}, \hat{\phi}|x)} = \prod_{p,m} \frac{P_{pm}(a_{pm}, \phi_{pm}; x, f_p)P(a_{pm}|r_p)P(\phi_{pm})}{P_{pm}(\hat{a}_{pm}, \hat{\phi}_{pm}; x, f_p)P(\hat{a}_{pm}|r_p)P(\hat{\phi}_{pm})}. \quad \text{(A.2)}$$

On the other hand, by evaluating (17) with $a_{p'm'} = \hat{a}_{p'm'}$ and $\phi_{p'm'} = \hat{\phi}_{p'm'}$ for all partials (p', m') except (p, m), we have

$$P(S, f, r, a_{pm}, \hat{a}_{\overline{pm}}, \phi_{pm}, \hat{\phi}_{\overline{pm}}|x) \propto P_0(x, f)P(r|S)P(f|S)P(S) \times P_{pm}(a_{pm}, \phi_{pm}; x, f_p)P(a_{pm}|r_p)P(\phi_{pm}) \times \prod_{(p',m')\neq(p,m)} P_{p'm'}(\hat{a}_{p'm'}, \hat{\phi}_{p'm'}; x, f_{p'})P(\hat{a}_{p'm'}|r_{p'})P(\hat{\phi}_{p'm'}). \quad \text{(A.3)}$$

Using Bayes law, this leads to

$$P(a_{pm}, \phi_{pm}|S, f, r_p, \hat{a}_{\overline{pm}}, \hat{\phi}_{\overline{pm}}, x) \propto P_0(x, f)P_{pm}(a_{pm}, \phi_{pm}; x, f_p)P(a_{pm}|r_p)P(\phi_{pm}) \times \prod_{(p',m')\neq(p,m)} P_{p'm'}(\hat{a}_{p'm'}, \hat{\phi}_{p'm'}; x, f_{p'}). \quad \text{(A.4)}$$

When evaluating this expression for $a_{pm} = \hat{a}_{pm}$ and $\phi_{pm} = \hat{\phi}_{pm}$, we get

$$P(\hat{a}_{pm}, \hat{\phi}_{pm}|S, f, r_p, \hat{a}_{\overline{pm}}, \hat{\phi}_{\overline{pm}}, x) \propto P_0(x, f) P_{pm}(\hat{a}_{pm}, \hat{\phi}_{pm}; x, f_p) P(\hat{a}_{pm}|r_p) P(\hat{\phi}_{pm}) \times \prod_{(p',m') \neq (p,m)} P_{p'm'}(\hat{a}_{p'm'}, \hat{\phi}_{p'm'}; x, f_{p'}). \quad \text{(A.5)}$$

After dividing (A.4) by (A.5), we conclude that

$$\frac{P(a_{pm}, \phi_{pm}|S, f, r_p, \hat{a}_{\overline{pm}}, \hat{\phi}_{\overline{pm}}, x)}{P(\hat{a}_{pm}, \hat{\phi}_{pm}|S, f, r_p, \hat{a}_{\overline{pm}}, \hat{\phi}_{\overline{pm}}, x)} = \frac{P_{pm}(a_{pm}, \phi_{pm}; x, f_p) P(a_{pm}|r_p) P(\phi_{pm})}{P_{pm}(\hat{a}_{pm}, \hat{\phi}_{pm}; x, f_p) P(\hat{a}_{pm}|r_p) P(\hat{\phi}_{pm})}. \quad \text{(A.6)}$$

The target result (18) is then readily obtained by replacing (A.6) into (A.2). Note that the fact that $\hat{a}$ and $\hat{\phi}$ are the MAP amplitude and phase parameter values associated with S, f and r is not used within this proof.

References

[1] M. Horne (Ed.), Prosody: theory and experiment, Kluwer Academic Publishers, Boston, MA, 2000.

[2] A. Klapuri, M. Davy, Signal processing methods for music transcription, Springer, New York, NY, 2006.

[3] E. Vincent, M. D. Plumbley, Low bit-rate object coding of musical audio using Bayesian harmonic models, IEEE Trans. on Audio, Speech and Language Processing 15 (4) (2007) 1273–1282.

[4] D. P. W. Ellis, Prediction-driven computational auditory scene analysis, Ph.D. thesis, Dept. of Electrical Engineering and Computer Science, MIT, Cambridge, MA (1996).

[5] M. Wu, D. Wang, G. J. Brown, A multipitch tracking algorithm for noisy speech, IEEE Trans. on Speech and Audio Processing 11 (3) (2003) 229–241.

[6] M. P. Ryynänen, A. P. Klapuri, Polyphonic music transcription using note event modeling, in: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2005, pp. 319–322.

[7] C. Raphael, Automatic transcription of piano music, in: Proc. Int. Conf. on Music Information Retrieval (ISMIR), 2002, pp. 15–19.

[8] E. Vincent, Musical source separation using time-frequency source priors, IEEE Trans. on Audio, Speech and Language Processing 14 (1) (2006) 91–98.

[9] A. Cont, Realtime multiple pitch observation using sparse non-negative constraints, in: Proc. Int. Conf. on Music Information Retrieval (ISMIR), 2006, pp. 206–212.

[10] A. T. Cemgil, H. J. Kappen, D. Barber, A generative model for music transcription, IEEE Trans. on Audio, Speech and Language Processing 14 (2) (2006) 679–694.

[11] A. T. Cemgil, S. J. Godsill, Efficient variational inference for the dynamic harmonic model, in: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2005, pp. 271–274.

[12] D. J. C. McKay, Information theory, inference, and learning algorithms, Cambridge University Press, Cambridge, UK, 2003.

[13] G. Casella, C. P. Robert, Monte Carlo statistical methods, 2nd Edition, Springer, New York, NY, 2005.

[14] P. J. Walmsley, S. J. Godsill, P. J. W. Rayner, Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters, in: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 1999, pp. 119–122.

[15] S. J. Godsill, M. Davy, Bayesian computational models for inharmonicity in musical instruments, in: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2005, pp. 283–286.

[16] M. Davy, S. J. Godsill, J. Idier, Bayesian analysis of western tonal music, Journal of the Acoustical Society of America 119 (4) (2006) 2498–2517.

[17] E. Vincent, M. D. Plumbley, Fast factorization-based inference for Bayesian harmonic models, in: Proc. IEEE Int. Conf. on Machine Learning for Signal Processing (MLSP), 2006, pp. 117–122.

[18] R. M. Neal, Probabilistic inference using Markov chain Monte Carlo methods, Tech. Rep. CRG-TR-93-1, Department of Computer Science, University of Toronto (1993).

[19] D. M. Chickering, D. Heckerman, Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables, in: Proc. Conf. on Uncertainty in Artificial Intelligence (UAI), 1996, pp. 158–168.

[20] C. J. van Rijsbergen, Information retrieval, 2nd Edition, Butterworths, London, UK, 1979.

[21] D. J. Sheskin, Handbook of parametric and nonparametric statistical procedures, 2nd Edition, Chapman & Hall, Boca Raton, FL, 2000.

[22] J.-F. Cardoso, Multidimensional independent component analysis, in: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 1998, pp. IV–1941–1944.