MDPI

Article

A Weakly Informative Prior for Resonance Frequencies

Marnix Van Soom * and Bart de Boer

AI Lab, Vrije Universiteit Brussel. Pleinlaan 2, 1050 Brussels, Belgium.
* Correspondence: marnix@ai.vub.ac.be

Abstract: We derive a weakly informative prior for a set of ordered resonance frequencies from Jaynes' principle of maximum entropy. The prior facilitates model selection problems in which both the number and the values of the resonance frequencies are unknown. It encodes a weakly inductive bias, provides a reasonable density everywhere, is easily parametrizable, and is easy to sample. We hope that this prior can enable the use of robust evidence-based methods for a new class of problems, even in the presence of multiplets of arbitrary order.

Keywords: weakly uninformative prior; resonance frequency; model selection; maximum entropy

Citation: Van Soom, M.; de Boer, B. A Weakly Informative Prior for Resonance Frequencies. *Proceedings* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

An important problem in the natural sciences is the accurate measurement of resonance frequencies. The problem can be formalized by the following probabilistic model:

$$p(D, \boldsymbol{x}|I) = p(D|\boldsymbol{x})p(\boldsymbol{x}|I) \equiv \mathcal{L}(\boldsymbol{x})\pi(\boldsymbol{x}), \tag{1}$$

where D is the data, $\boldsymbol{x} = \{x_k\}_{k=1}^K$ are the K resonance frequencies of interest, and I is the assumed prior information about the possible values of $\boldsymbol{x}$. Note that we do not condition explicitly on other prior information I', such as the model underlying the data, noise properties, and the values of various hyperparameters. As an example instance of (1), we refer to the vocal tract resonance (VTR) problem discussed in Section 5, for which D is audio recorded from the mouth of a speaker, the $\boldsymbol{x}$ are K VTR frequencies, and the underlying model is a sinusoidal regression model. Furthermore, any realistic problem will include additional model parameters $\boldsymbol{\theta}$, but these have been silently ignored by formally integrating them out of (1), i.e., $p(D, \boldsymbol{x}|I) = \int \mathrm{d}\boldsymbol{\theta}\, p(D, \boldsymbol{x}, \boldsymbol{\theta}|I)$.

In this paper we assume that the likelihood $\mathcal{L}(\boldsymbol{x}) \equiv p(D|\boldsymbol{x})$ is given and that only the prior $\pi(\boldsymbol{x}) \equiv p(\boldsymbol{x}|I)$ remains to be chosen from knowledge of I. In addition, we are interested only in so-called uninformative or weakly informative choices of π, which implies that we shall take I to mean only limited prior information about the possible values of K and $\boldsymbol{x}$. In practice, this assumption induces a remarkable conflict between π and I, which is that *assuming limited prior information I actually precludes the uninformative priors π most commonly chosen to express that I.*

The goal of this paper is to describe this conflict and show how it can be resolved by adopting a specific choice for π. This allows robust inference of the number of resonances K in the important case of limited prior information I, which in turn enables accurate measurement of the resonance frequencies $\boldsymbol{x}$ with standard methods such as nested sampling [1] or reversible jump MCMC [2].

2. Notation

The symbol π is intended to convey a vague notion of a generally uninformative or weakly informative prior conditioned on limited prior information I. In contrast, definite *choices* for π and I are indicated with the subscript i. We consider three of them in this paper:

$$\pi_i(\boldsymbol{x}) \equiv p(\boldsymbol{x}|\boldsymbol{\beta}_i, I_i), \qquad (i = 1, 2, 3), \tag{2}$$

Table 1. The values of the hyperparameters $\boldsymbol{\beta}_i$ used throughout the paper. All quantities are given in units of Hz.

$k \rightarrow$	0	1	2	3	4	5	6	7	8	9	10
$\boldsymbol{a} = \{a_k\}$		200	600	1400	2900	3500					
$\boldsymbol{b} = \{b_k\}$		1100	3500	4000	4500	5500					
$\overline{\boldsymbol{x_0}} = \{\overline{x_k}\}$	200	500	1000	1500	2000	2500	3000	3500	4000	4500	5000
other	$x_0 = 200$						$x_{\mathsf{max}} = 5500$				

where $\boldsymbol{\beta}_i$ is a placeholder for the hyperparameter specific to π_i. Note that in the plots below and for the experiments in Section 5 the values of the $\boldsymbol{\beta}_i$ are always set according to Table 1. The conditioning prior information I_i in (2) is characterized in Appendix A.

Each π_i uniquely determines a number of important high-level quantities, since the likelihood $\mathcal{L}(\boldsymbol{x})$ and data D are assumed given. These quantities are the *evidence* for the model with K resonances

$$Z_i(K) = \int \mathrm{d}^K\boldsymbol{x}\, \mathcal{L}(\boldsymbol{x})\pi_i(\boldsymbol{x}), \tag{3}$$

the *posterior*

$$P_i(\boldsymbol{x}) = \frac{\mathcal{L}(\boldsymbol{x})\pi_i(\boldsymbol{x})}{Z_i(K)}, \tag{4}$$

and the *information*

$$H_i(K) = \int \mathrm{d}^K\boldsymbol{x}\, P_i(\boldsymbol{x}) \log \frac{P_i(\boldsymbol{x})}{\pi_i(\boldsymbol{x})}, \tag{5}$$

which measures the amount of information obtained by updating from prior π_i to posterior P_i, i.e., $H_i(K) \equiv D_{\mathrm{KL}}(P_i|\pi_i)$, where D_{KL} is the Kullback-Leibler divergence [3].

3. Conflict

For convenience, we repeat the conflict as worded in the Introduction here: *assuming limited prior information I actually precludes the uninformative priors π most commonly chosen to express that I.* To understand it, we will take a closer look at the two elements involved.

First, the *uninformative priors* π in question are of the independent and identically distributed type,

$$\pi(\boldsymbol{x}) = \prod_{k=1}^{K} g(x_k|\boldsymbol{\beta}), \tag{6}$$

where $g(x|\boldsymbol{\beta})$ is any wide distribution with hyperparameters $\boldsymbol{\beta}$. A typical choice for g is the uniform distribution over the full frequency bandwidth; other examples are diffuse Gaussians or Jeffreys priors [e.g., 4–10].

Second, the *limited prior information I* about K implies that the problem will involve model selection, since each value of K implicitly corresponds to a different model for the data [11]. It is thus necessary to evaluate and compare the evidence $Z(K) = \int \mathrm{d}^K\boldsymbol{x}\, \mathcal{L}(\boldsymbol{x})\pi(\boldsymbol{x})$ for each plausible K.

The conflict between these two elements is due to the *label switching problem*, which is a well-known issue in mixture modeling [e.g., 12]. The likelihood functions $\mathcal{L}(\boldsymbol{x})$ used in models parametrized by resonance frequencies are typically invariant to switching the label k; i.e., the index k of the frequency x_k has no distinguishable meaning in the model underlying the data. The posterior $P(\boldsymbol{x}) \propto \mathcal{L}(\boldsymbol{x})\pi(\boldsymbol{x})$ will inherit this *exchange symmetry* if the prior is of type (6). Thus, if the model parameters $\boldsymbol{x}$ are well determined by the data D, the posterior landscape will consist of one *primary* mode, which is defined as a mode living in the *ordered region*

$$\mathcal{R}_K(x_0) = \{\boldsymbol{x}|x_0 \leq x_1 \leq x_2 \leq \cdots \leq x_K\} \quad \text{with} \quad x_0 > 0, \tag{7}$$

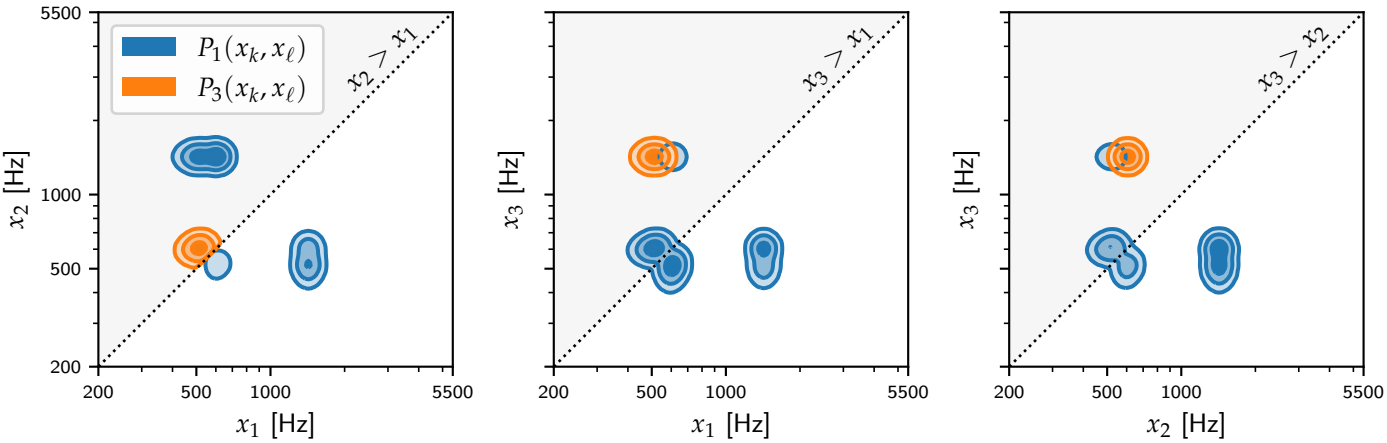


Figure 1. The label switching problem (P_1) and its solution (P_3) for a well-determined instance of the VTR problem from Section 5 with $K := 3$. The pairwise marginal posteriors $P_i(x_k, x_\ell)$ are shown using the isocontours of kernel density approximations calculated from posterior samples of $\boldsymbol{x}$. For each panel the diagonal $x_k = x_\ell$ is plotted as a dotted line and the ordered region $\mathcal{R}_3(x_0)$ is shaded in grey.

and $(K! - 1)$ *induced* modes, which are identical to the primary mode up to a permutation of the labels k and thus live outside of the region $\mathcal{R}_K(x_0)$. The trouble is that correctly taking into account these induced modes during the evaluation of $Z(K)$ requires a surprising amount of extra work besides tuning the MCMC method of choice, and that is the label switching problem in our setting. In fact there is currently no widely accepted solution for the label switching problem in the context of mixture models either [13,14]. This is, then, how uninformative priors π are "precluded" by the limited information I: the latter implies model selection, which in turn implies evaluating $Z(K)$, which is hampered by the label switching problem due to the exchange symmetry of the former. Therefore, it seems better to try to avoid it by encoding our preference for primary modes directly into the prior. This leads to abandoning the uninformative prior π in favor of the weakly informative prior π_3, which is proposed in Section 4 as a solution to the conflict.

We use the VTR problem to briefly illustrate the label switching problem in Figure 1. The likelihood $\mathcal{L}(\boldsymbol{x})$ is described implicitly in Section 5 and is invariant to switching the labels k because the underlying model function (22) of the regression model is essentially a sum of sinusoids, one for each x_k. As frequencies can be profitably thought of as scale variables [15, App. A], the uninformative prior (6) is represented by

$$\pi_1(\boldsymbol{x}) \equiv p(\boldsymbol{x}|x_0, x_{\max}, I_1) = \prod_{k=1}^{K} h(x_k|x_0, x_{\max}), \tag{8}$$

where $\boldsymbol{\beta}_1 \equiv (x_0, x_{\max})$ are a common lower and upper bound, and

$$h(x|a,b) = \begin{cases} \frac{1}{\log(b/a)} \frac{1}{x} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \quad \text{with} \quad \begin{matrix} a > 0 \\ b < \infty \end{matrix} \tag{9}$$

is the Jeffreys prior, the conventional uninformative prior for a scale variable [16–18]. We have visualized the posterior landscape $P_1(\boldsymbol{x})$ in Figure 1 using the pairwise marginal posteriors $P_1(x_k, x_\ell)$ plotted in blue. Note the exchange symmetry of P_1, which manifests as an (imperfect) reflection symmetry around the dotted diagonal $x_k = x_\ell$ bordering the ordered region $\mathcal{R}_3(x_0)$. The primary mode is plotted in orange; all other blue modes are induced modes. [This is because it just so happened that the primary mode in P_1 was missed by the MCMC exploration; while convenient for visualization purposes, this is expected behaviour only for $K \gtrsim 4$, as the number of induced modes grows as $K!$.]

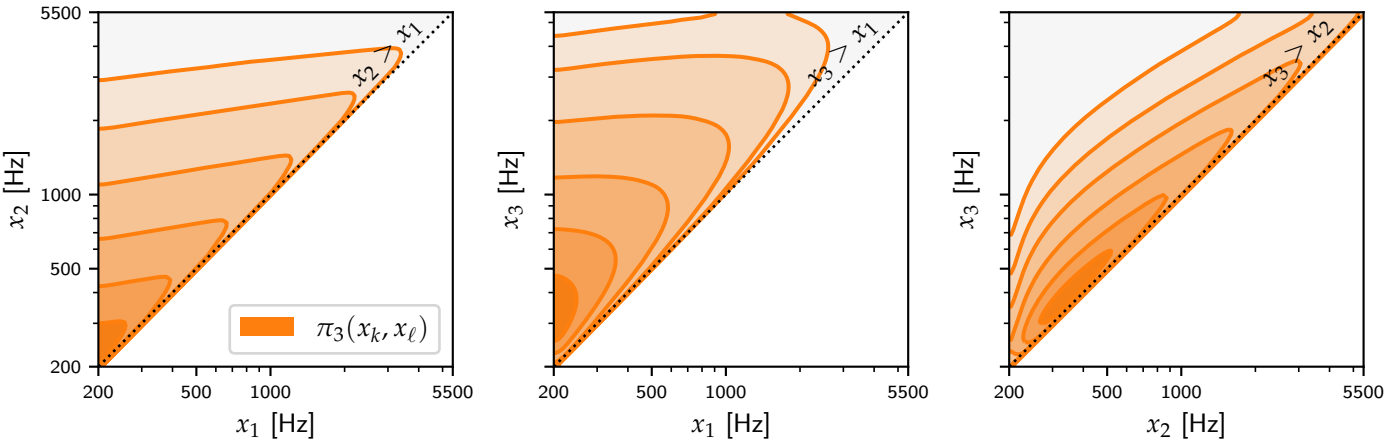


Figure 2. Representation of π_3 by the pairwise marginal priors $\pi_3(x_k, x_\ell)$ for the case $K := 3$, plotted similarly to Figure 1. The marginal $\pi_3(x_k, x_\ell)$ is obtained by integrating out the third frequency; for example, $\pi_3(x_1, x_2) = \iint \mathrm{d}x_3\, \pi_3(\boldsymbol{x})$.

3.1. A Simple Way Out?

A simple way out of the conflict is to break the exchange symmetry by assuming specialized bounds for each x_k:

$$\pi_2(\boldsymbol{x}) \equiv p(\boldsymbol{x}|\boldsymbol{a}, \boldsymbol{b}, I_2) = \prod_{k=1}^{K} h(x_k|a_k, b_k), \tag{10}$$

where $\boldsymbol{\beta}_2 \equiv (\boldsymbol{a}, \boldsymbol{b})$ with $\boldsymbol{a} = \{a_k\}_{k=1}^K$ and $\boldsymbol{b} = \{b_k\}_{k=1}^K$ are hyperparameters specifying the individual bounds. However, in order to enable the model to detect doublets (a resolved pair of two close frequencies such as the primary mode in the leftmost panel in Figure 1), it is necessary to assign overlapping bounds in $(\boldsymbol{a}, \boldsymbol{b})$, presumably using some heuristic. The necessary degree of overlap increases as the detection of higher order multiplets like triplets (which can and do occur) is desired, but the more overlap in $(\boldsymbol{a}, \boldsymbol{b})$, the more the label switching problem returns. Despite this issue, there will be cases where we have sufficient prior information I to set the $(\boldsymbol{a}, \boldsymbol{b})$ hyperparameters without too much trouble – in fact, the VTR problem is such a case, for which the overlapping values of $(\boldsymbol{a}, \boldsymbol{b})$ up to $K = 5$ are given in Table 1.

4. Solution

Our solution to the conflict discussed in Section 3 is a chain of K coupled Pareto distributions:

$$\pi_3(\boldsymbol{x}) \equiv p(\boldsymbol{x}|\overline{\boldsymbol{x_0}}, I_3) = \prod_{k=1}^{K} \mathsf{Pareto}(x_k|x_{k-1}, \lambda_k) \tag{11}$$

where

$$\mathsf{Pareto}(x|x_*, \lambda) = \begin{cases} \frac{\lambda x_*^\lambda}{x^{\lambda+1}} & \text{if } x \geq x_* \\ 0 & \text{otherwise} \end{cases} \quad \text{with} \quad \begin{matrix} x_* > 0 \\ \lambda > 0, \end{matrix} \tag{12}$$

and the hyperparameter $\boldsymbol{\beta}_3 \equiv \overline{\boldsymbol{x_0}}$ is defined as

$$\overline{\boldsymbol{x_0}} \equiv (\overline{x_0}, \overline{\boldsymbol{x}}), \quad \overline{x_0} := x_0, \quad \overline{\boldsymbol{x}} = \{\overline{x_k}\}_{k=1}^K, \quad \lambda_k = \frac{\overline{x_k}}{\overline{x_k} - \overline{x_{k-1}}}. \tag{13}$$

The expression for π_3 (11) is the main contribution of the paper; it is derived in Section 4.1 and illustrated in Figures 2 and 3.

It can be seen that π_3 encodes weakly informative knowledge about K *ordered frequencies*, because (11) and (12) together imply that $\pi_3(\boldsymbol{x})$ is defined only for $\overline{\boldsymbol{x}} \in$

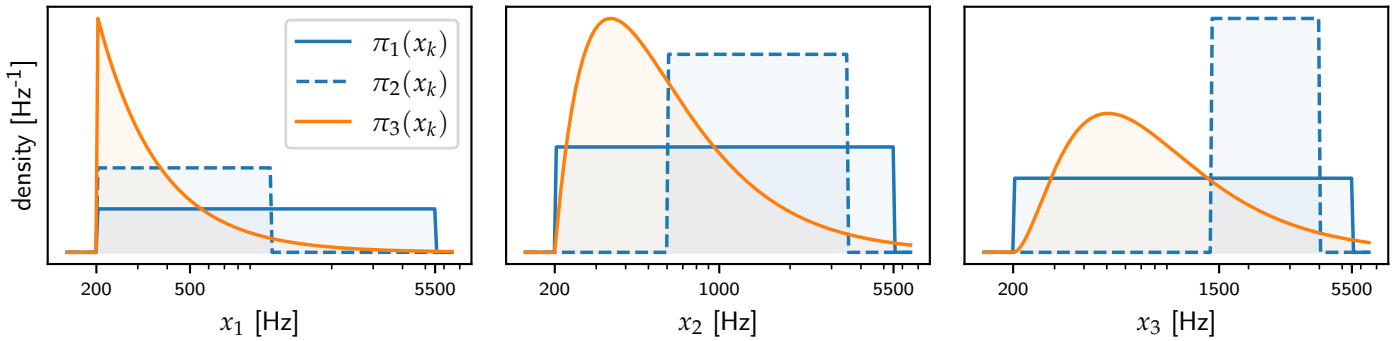


Figure 3. Comparison of π_1, π_2 and π_3 in terms of the marginal priors $\pi_i(x_k)$ for the case $K := 3$. The priors are defined in (8, 10, 11), respectively. The marginal $\pi_i(x_k)$ is obtained by integrating out the two other frequencies; for example, $\pi_i(x_1) = \iint \mathrm{d}x_2\, \mathrm{d}x_3\, \pi_i(\boldsymbol{x})$.

$\mathcal{R}_K(x_0)$, while nonzero only for $\boldsymbol{x} \in \mathcal{R}_K(x_0)$. In other words, its support is precisely the ordered region $\mathcal{R}_K(x_0)$, which solves the label switching problem underlying the conflict automatically. This is illustrated in Figure 1, where P_3 contracts to a single primary mode, which is just what we would like.

The $K+1$ hyperparameters $\overline{\boldsymbol{x_0}}$ in (13) are a *common lower bound* x_0 plus K *expected values of the resonance frequencies* $\overline{\boldsymbol{x}}$. While the former is generally easily determined, the latter may seem difficult to set, given the premise of this paper that we dispose only of limited prior information I. More precisely, why do we claim that π_3 is only weakly informative, if it is parametrized by the expected values of the very things it is supposed to be only weakly informative about? The answer is that for any reasonable amount of data, inference based on π_3 is completely insensitive to the exact values of $\overline{\boldsymbol{x}}$. This fact makes it easy to set $\overline{\boldsymbol{x}}$ after all, as any reasonable guess will suffice in practice. For example, for the VTR problem we simply applied a heuristic where we take $\overline{x_k} = k \times 500$ Hz (see Table 1). This insensitivity is due to the maximum entropy status of π_3, and indicates the weak inductive bias it entails. On a more prosaic level, the heavy tails of the Pareto distributions in (11) ensure that the prior will be eventually overwhelmed by the data, no matter how a priori improbable the true value of $\boldsymbol{x}$ is. More prosaic still, below we show quantitatively that for the VTR problem π_3 is about as (un)informative as π_2 [Figure 5(b)].

4.1. Derivation of π_3

We now give a rather brief derivation of (11), due to limited space. Our ansatz consists of interpreting the $\boldsymbol{x}$ as a set of K *ordered* scale variables which are bounded from below by x_0. Starting from (8) and not bothering with the bounds (a, b), we obtain the improper pdf

$$m(\boldsymbol{x}) \propto \begin{cases} \prod_{k=1}^{K} \frac{1}{x_k} & \boldsymbol{x} \in \mathcal{R}_K(x_0) \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

We can simplify (14) using the one-to-one transformation $\boldsymbol{x} \leftrightarrow \boldsymbol{u}$ defined as

$$\begin{aligned} \boldsymbol{x} \to \boldsymbol{u}: \quad & u_k = \log \frac{x_k}{x_{k-1}} && (k = 1, 2, \ldots, K) \\ \boldsymbol{u} \to \boldsymbol{x}: \quad & x_k = x_0 \exp \sum_{\kappa=1}^{k} u_\kappa && (k = 1, 2, \ldots, K) \end{aligned} \tag{15}$$

which yields (with abuse of notation for brevity)

$$m(\boldsymbol{u}) \propto \begin{cases} 1 & \boldsymbol{u} \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

where $\boldsymbol{u} \geq 0$ is shorthand for $u_1 \geq 0, u_2 \geq 0, \ldots, u_K \geq 0$.

Since model selection requires proper priors, we need to normalize $m(\boldsymbol{u})$ by adding extra information (i.e., constraints) to it; we propose to simply fix the K first moments $\langle \boldsymbol{u} \rangle = \{\langle u_k \rangle\}_{k=1}^K$. This will yield the Pareto chain prior $\pi_3(\boldsymbol{u})$ directly, expressed in $\boldsymbol{u}$ space rather than $\boldsymbol{x}$ space. The expression for $\pi_3(\boldsymbol{u})$ is found by minimizing the Kullback-Leibler divergence [19]

$$D_{\mathrm{KL}}(\pi_3|m) = \int \mathrm{d}^K\boldsymbol{u}\, \pi_3(\boldsymbol{u}) \log \frac{\pi_3(\boldsymbol{u})}{m(\boldsymbol{u})}, \quad \text{subject to} \quad \langle \boldsymbol{u} \rangle \equiv \int \mathrm{d}^K\boldsymbol{u}\, \boldsymbol{u} \pi_3(\boldsymbol{u}) = \overline{\boldsymbol{u}}, \tag{17}$$

where $\overline{\boldsymbol{u}} = \{\overline{u_k}\}_{k=1}^K$ are the supplied first moments. This variational problem is equivalent to finding $\pi_3(\boldsymbol{u})$ by means of Jaynes' *principle of maximum entropy* [20] with $m(\boldsymbol{u})$ serving as the invariant measure [21]. Since the exponential distribution $\mathsf{Exp}(x|\lambda)$ is the maximum entropy distribution for a random variable $x \geq 0$ with a fixed first moment $\langle x \rangle = 1/\lambda$, the solution to (17) is

$$\pi_3(\boldsymbol{u}) = \prod_{k=1}^K \mathsf{Exp}(u_k|\lambda_k), \tag{18}$$

where the rate hyperparameters $\lambda_k = 1/\overline{u_k}$ and

$$\mathsf{Exp}(x|\lambda) = \begin{cases} \lambda \exp\{-\lambda x\} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{with} \quad \lambda > 0. \tag{19}$$

Transforming (18) to $\boldsymbol{x}$ space using (15) finally yields (11), but we still need to express λ_k in terms of $\overline{\boldsymbol{x}}$ – we might find it hard to pick reasonable values of $\overline{u_k} = \overline{\log x_k/x_{k-1}}$ from limited prior information I. For this we will need the identity (see Appendix B for a proof)

$$\langle x_k \rangle \equiv \int \mathrm{d}^K\boldsymbol{x}\, x_k \pi_3(\boldsymbol{x}) = \frac{\lambda_k}{\lambda_k - 1} \langle x_{k-1} \rangle \qquad (k = 1, 2, \ldots, K). \tag{20}$$

Constraining $\langle x_k \rangle = \overline{x_k}$ and solving for λ_k, we obtain $\lambda_k = \overline{x_k}/(\overline{x_k} - \overline{x_{k-1}})$, in agreement with (13). Note that the existence of the first marginal moments $\langle x_k \rangle$ requires that $\lambda_k > 1$.

4.2. *Sampling from π_3*

Sampling from π_3 is trivial because of the independence of the u_k in $\boldsymbol{u}$ space (18). To produce a sample $\boldsymbol{x}' \sim \pi_3(\boldsymbol{x})$ given the hyperparameter $\overline{\boldsymbol{x_0}}$, compute the corresponding rate parameters $\{\lambda_k\}_{k=1}^K$ from (13), and use them in (18) to obtain a sample $\boldsymbol{u}' \sim \pi_3(\boldsymbol{u})$. The desired $\boldsymbol{x}'$ is then obtained from $\boldsymbol{u}'$ using the transformation (15).

Example Python code is given in Appendix C.

5. Application: The VTR Problem

We now present a relatively simple – but real – instance of the problem of measuring resonance frequencies, which will allow us to illustrate the above ideas. The VTR problem consists of measuring human vocal tract resonance (VTR) frequencies $\boldsymbol{x}$ for each of five representative vowel sounds taken from the CMU ARCTIC database [22]. The VTR frequencies $\boldsymbol{x}$ describe the *vocal tract transfer function* $T(x)$ and are fundamental quantities in acoustic phonetics [23]. The five vowel sounds are recorded utterances of the first vowel in the words $W = \{\mathsf{shore}, \mathsf{that}, \mathsf{you}, \mathsf{little}, \mathsf{until}\}$. In order to achieve high-quality VTR frequency estimates $\hat{\boldsymbol{x}}$, only the quasi-periodic *steady-state* part of the vowel sound is considered for the measurement. The data D thus consists of a string of highly correlated *pitch periods*. See Figure 4 for an illustration of these concepts.

The measurement itself is formalized as inference using the probabilistic model (1). The model assumed to underlie the data is the sinusoidal regression model introduced in [24]; due to limited space, we only describe it implicitly, because full specification of

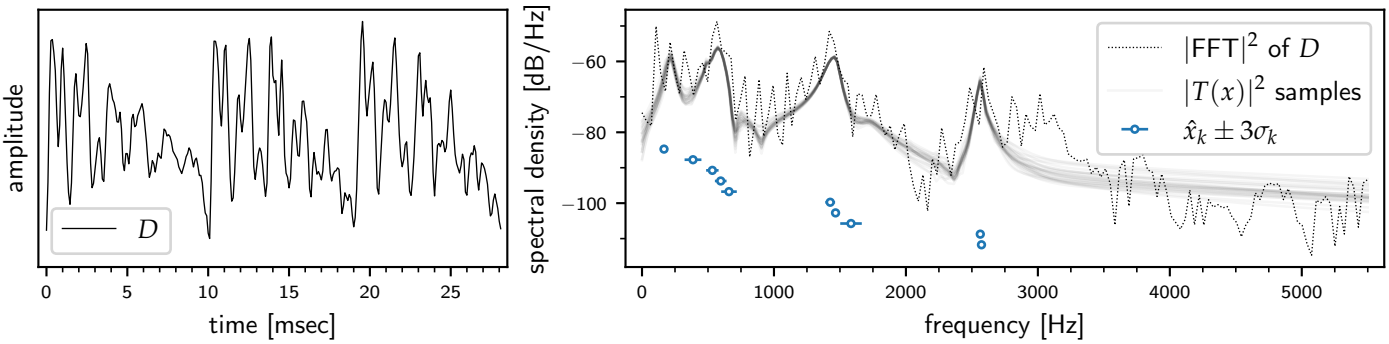


Figure 4. The VTR problem for the case ($D :=$ until, $K := 10$). Left panel: The data D, i.e., the quasi-periodic steady-state part consisting of 3 highly correlated pitch periods. Right panel: Inferred VTR frequency estimates $\{\hat{x}_k\}_{k=1}^{K}$ for $K := 10$ at 3 sigma. They describe the power spectral density of the vocal tract transfer function $|T(x)|^2$, represented here by 25 posterior samples and compared to the Fast Fourier Transform (FFT) of D. All $\hat{x}_k$ are well resolved and most have error bars too small to be seen on this scale.

the likelihood $\mathcal{L}(\boldsymbol{x})$ would introduce much superfluous detail. The sinusoidal regression model assumes each pitch period $\boldsymbol{d} \in D$ can be modeled as

$$d_t = f(t; \boldsymbol{A}, \boldsymbol{\alpha}, \boldsymbol{x}) + \sigma e_t \quad \text{where} \quad e_t \sim \mathcal{N}(0,1), \qquad (t = 1, 2, \ldots, T), \tag{21}$$

where $\boldsymbol{d} = \{d_t\}_{t=1}^{T}$ is a time series consisting of T samples. The model function

$$f(t; \boldsymbol{A}, \boldsymbol{\alpha}, \boldsymbol{x}) = \sum_{k=1}^{K} [A_k \cos(x_k t) + A_{K+k} \sin(x_k t)] \exp\{-\alpha_k t\} + \sum_{\ell=1}^{L} A_{2K+\ell} t^{\ell-1} \tag{22}$$

consists of a sinusoidal part (first $\sum$) and a polynomial trend correction (second $\sum$). Note the additional model parameters $\boldsymbol{\theta} = \{\boldsymbol{A}, \boldsymbol{\alpha}, \sigma, L\}$. Formally, given the prior $p(\boldsymbol{\theta})$ [24, Sec. 2.2], the marginal likelihood $\mathcal{L}(\boldsymbol{x})$ is then obtained as $\mathcal{L}(\boldsymbol{x}) = \int \mathrm{d}\boldsymbol{\theta}\, \mathcal{L}(\boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta})$, where the complete likelihood $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\theta})$ is implicitly given by (21) and (22). Practically, we just marginalize out $\boldsymbol{\theta}$ from samples obtained from the complete problem $p(D, \boldsymbol{x}, \boldsymbol{\theta} | I)$.

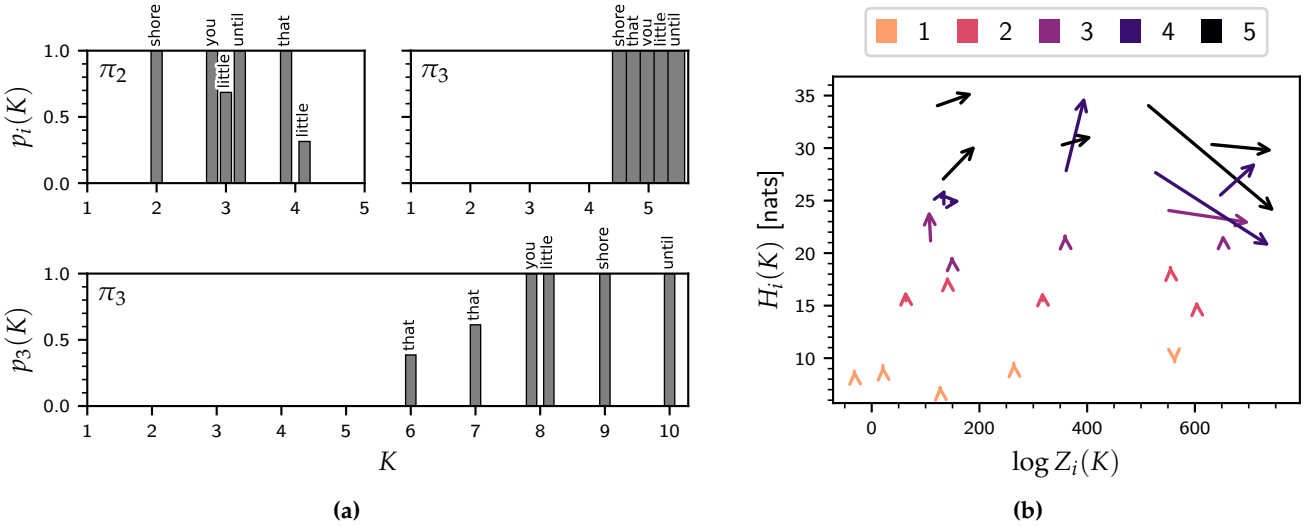
For inference, the computational method of choice is nested sampling [25] using the dynesty library [26–30]. Since the VTR problem is quite simple [$H_i(K) \sim 30$ nats], we only perform single nested sampling runs and take the obtained $\log Z_i(K)$ and $H_i(K)$ as point estimates. Full details on the experiments and data are at https://github.com/mvsoom/frequency-prior. Finally, we point out several theoretical connections between π_3 and important concepts from acoustic phonetics in Appendix D.

5.1. Experiment I: Comparing π_2 and π_3

In Experiment I, we perform a high-level comparison between π_2 and π_3 in terms of *evidence* (3) and *information* (5). We did not include π_1 in this comparison as the label switching problem prevented convergence of nested sampling runs for $K \geq 4$. The $(\boldsymbol{a}, \boldsymbol{b})$ bounds for π_2 were based on formant tables from several works [31–36]; i.e., we loosely interpreted the VTRs as formants [37], which dictated that $K_{\max} = 5$. For π_3 we simply applied a heuristic where we take $\overline{x_k} = k \times 500$ Hz. We selected x_0 empirically (although a theoretical approach is also possible [38]) and $x_{\max}$ was set to the Nyquist frequency. The role of $x_{\max}$ is to truncate π_3 in order to avoid aliasing effects, since the support of $\pi_3(x_i)$ is unbounded from above. We implemented this by using the following likelihood function in the nested sampling program:

$$\mathcal{L}'(\boldsymbol{x}) = \begin{cases} \mathcal{L}(\boldsymbol{x}) & \text{if } x_k \leq x_{\max} \text{ for all } (k = 1, 2, \ldots, K) \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

Another approach is to truncate π_3 directly with rejection sampling; see Appendix C.



(a) **(b)**

Figure 5. (a) Model selection in Experiment I (top row) and Experiment II (bottom row). **(b)** In Experiment I, π_2 and π_3 are compared in terms of evidence [$\log Z_i(K)$] and uninformativeness [$H_i(K)$] for each (D, K). The arrows point from π_2 to π_3 and are color-coded by the value of K. For small values of K, the arrow lengths are too small to be visible on this scale.

First, we compare the influence of π_2 and π_3 on model selection. Given $D \in W$, the posterior probability of the number of resonances K is given by

$$p_i(K) = \frac{Z_i(K)}{\sum_{K'} Z_i(K')} \qquad (K = 1, 2, \ldots, K_{\max}). \tag{24}$$

The results in the top row of Figure 5(a) are striking: while $p_2(K)$ shows individual preferences based on D, $p_3(K)$ prefers $K = K_{\max}$ unequivocally.

Second, in Figure 5(b) we compare π_2 and π_3 directly in terms of differences in evidence [$\log Z_i(K)$] and uninformativeness [$H_i(K)$] for each combination (D, K).

Arrows pointing *eastward* indicate $Z_3(K) > Z_2(K)$. The π_3 prior dominates the π_2 prior in terms of evidence, for almost all values of K, indicating that π_3 places its mass in regions of higher likelihood; or, equivalently, that the data was much more probable under π_3 than π_2. This implies that the hint of π_3 at more structure beyond $K > K_{\max}$ should be taken seriously – we investigate this in Section 5.2.

Arrows pointing *northward* indicate $H_3(K) > H_2(K)$, i.e., π_3 is *less* informative than π_2, since more information is gained by updating from π_3 to P_3 than from π_2 to P_2. It is seen that π_2 and π_3 are roughly comparable in terms of (un)informativeness.

5.2. Experiment II: 'Free' Analysis

We now freely look for more structure in the data by letting K go up until $K_{\max} = 10$. This goes beyond the capacities of π_1 (because of the label switching problem) and π_2 (because no data is available to set the $(\boldsymbol{a}, \boldsymbol{b})$ bounds). The great advantage of π_3 is thus that we can use a simple heuristic to set $\overline{\boldsymbol{x_0}}$ and let the model do the discovering without worrying about convergence issues or the obtained evidence values. The bottom row in Figure 5(a) shows that model selection for the VTR problem is well-defined, with the most probable values of $K \leq 10$, except for $D = \texttt{until}$. That case is investigated in Figure 4, where the need for more VTRs (higher K) is apparent from the unmodeled broad peak centered at around 3000 Hz in the FFT power spectrum (right panel). Incidentally, this spectrum also shows that spectral peaks are often resolved into more than one VTR, which underlines the importance of using a prior that enables trouble-free handling of multiplets of arbitrary order. A final observation from the spectrum is the fact that the inferred $\hat{x}_k$ differ substantially from the supplied values in $\overline{\boldsymbol{x}}$ (Table 1), which hints at the weak inductive bias underlying π_3.

6. Discussion

> It is only when the information in the prior is comparable to the information in the data that the prior probability can make any real difference in parameter estimation problems or in model selection problems [39, p. 9].

Alhough the prior π_3 is meant to be overwhelmed, its practical advantage (i.e., solving the label switching problem) will nonetheless persist, making a "real difference [...] in model selection problems" even when "the information in the prior" is much smaller than "the information in the data". In this sense π_3 is quite unlike "the prior" referenced in the above quote. Since it will be overwhelmed, all it has to do is provide a reasonable density everywhere (which it does), and be easily parametrizable (which it is), and be easy to sample from (which it is).

We thus hope that this prior can enable the use of robust evidence-based methods for a new class of problems, even in the presence of multiplets of arbitrary order. It is valid for any collection of scale variables which are intrinsically ordered, of which frequencies and wavelengths seem to be the most natural examples. Some examples of recent work where the prior could be applied directly are:

- Nuclear magnetic resonance (NMR) spectroscopy [40]
- Resonant ultrasound spectroscopy (a standard method in material science) [41]
- In the analysis of atomic spectra [42], such as X-ray diffraction [43]
- Absorption spectral-line finding in astronomy [44]
- Accurate modeling of instrument noise (in this case LIGO/Virgo noise) [45]
- Measuring high precision acoustic impedance spectra of the vocal tract [46]
- Spectral mixture kernels in Gaussian processes [47]
- Model-based Bayesian analysis in acoustics [48]

Author Contributions: TODO

Funding: This research was funded by the Flemish AI plan and by the Research Foundation Flanders (FWO) under grant number G015617N.

Acknowledgments: TODO

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Characterizing the Limited Prior Information I_i

The three resonance frequency priors π_1, π_2, π_3 discussed in this paper form a sequence of priors that represent states of knowledge I_1, I_2, I_3 that are expected to be increasingly informative about the possible values of $\boldsymbol{x}$.

The priors are all based on the Jeffreys prior $h(x) \propto 1/x$, which, by *Jaynes' transformation invariance principle* [17,21], represents a state of total ignorance about the scale parameter x. This means that we know nothing more than the fact that x is a scale parameter for the likelihood $\mathcal{L}(x)$.[1]

The distributions representing complete ignorance found by Jaynes' transformation invariance principle are improper (not normalizable), and $h(x)$ is no exception to that. In Jaynes' own words, this impropriety "arises simply from the fact that our formulation of the notation of complete ignorance was an idealization that does not strictly apply in any realistic problem" [21, p. 22]. In other words, $h(x)$ does not represent any realistic state of knowledge, and – signaled by the appearance of hyperparameters – information must be added to make a normalizable prior out of $h(x)$.

In the case of $\pi_1(\boldsymbol{x})$ and $\pi_2(\boldsymbol{x})$, this extra information comes simply in the form of bounds on the range of $\boldsymbol{x}$, either global (π_1) or specialized (π_2).

[1] This statement can be made precise in terms of transformation groups; see [49, Sec. 12.4.1]. Bretthorst [50, App. A] shows how (resonance) frequencies can be interpreted as scale variables by demanding that the likelihood $\mathcal{L}(x)$ be invariant in form under a rescaling of the conjugate time variable.

Table A1. Verbalization of the information represented by the I_i symbols in $\pi_i \equiv p(\boldsymbol{x}|\boldsymbol{\beta}_i, I_i)$.

I_1	$\equiv$	$\boldsymbol{x}$ are K scale variables with global bounds $(x_0, x_{\max})$ ignorance of $\boldsymbol{x}$ within support
I_2	$\equiv$	$\boldsymbol{x}$ are K scale variables with specialized bounds $(\boldsymbol{a}, \boldsymbol{b})$ ignorance of $\boldsymbol{x}$ within support
I_3	$\equiv$	$\boldsymbol{x}$ are K ordered scale variables bounded from below by x_0 ignorance of $\boldsymbol{x}$ within support, to the extent that $\boldsymbol{x}$ is expected to be around $\overline{\boldsymbol{x}}$

In the case of $\pi_3(\boldsymbol{x})$, information is added through ordering $\boldsymbol{x}$ (which is a kind of bounding) and constraining the first moments $\langle \boldsymbol{x} \rangle = \overline{\boldsymbol{x}}$ using Jaynes' principle of maximum entropy. This in effect ensures that π_3 is as 'spread-out' as possible while still agreeing with the given first moments, which in turn renders inference with π_3 highly insensitive to the actual values of $\overline{\boldsymbol{x}}$ supplied in the presence of a reasonable amount of data.

Thus we can characterize the limited prior information I_i by combining the idealized ignorance represented by $h(x)$ with the information added for each of the π_i to make it proper. A summary is given in Table A1.

Finally, we emphasize that although we have suggestively ordered the priors π_1, π_2, π_3 in terms of 'expected' increasing informativeness, the actual (un)informativeness of a prior $H_i(K)$, as with the evidence $Z_i(K)$, depends on the details of the problem at hand, including the choice of hyperparameters $\boldsymbol{\beta}_i$. Indeed, in the particular experiment of Section 5.1 we saw that π_2 and π_3 were comparable in terms of informativeness (while π_3 dominated π_2 in terms of evidence).

Appendix B. Proof of (20)

We start out by showing that π_3 is consistent under adding a new frequency; i.e., marginalizing out the last (highest) frequency is equivalent to having set up π_3 without knowledge of that frequency. Symbolically,

$$\pi_3(\boldsymbol{x}_{-K}|K) = \pi_3(\boldsymbol{x}_{-K}|K-1), \tag{A1}$$

where we have conditioned on K explicitly and used the 'cavity notation' $\boldsymbol{x}_{-\ell} = \{x_k\}_{k\in\{1\cdots K\}\setminus\{\ell\}}$, i.e., $\boldsymbol{x}_{-\ell}$ is $\boldsymbol{x}$ with the ℓth element missing. The proof of (A1) is trivial:

$$\begin{aligned}\pi_3(\boldsymbol{x}_{-K}|K) &\equiv \int \mathrm{d}x_K\, \pi_3(\boldsymbol{x}|K) = \int \mathrm{d}x_K \prod_{k=1}^{K} \mathsf{Pareto}(x_k|x_{k-1}, \lambda_k) \\ &= \prod_{k=1}^{K-1} \mathsf{Pareto}(x_k|x_{k-1}, \lambda_k) = \pi_3(\boldsymbol{x}_{-K}|K-1).\end{aligned} \tag{A2}$$

Next, we prove a special case of (20); namely, for x_K:

$$\begin{aligned}\langle x_K|K\rangle &\equiv \int \mathrm{d}^K\boldsymbol{x}\, x_K \pi_3(\boldsymbol{x}|K) \\ &= \int \mathrm{d}^{K-1}\boldsymbol{x}_{-K}\, \pi_3(\boldsymbol{x}_{-K}|K) \int \mathrm{d}x_K\, x_K \underbrace{\pi_3(x_K|\boldsymbol{x}_{-K}, K)}_{=\,\mathsf{Pareto}(x_K|x_{K-1},\lambda_K)} \\ &= \frac{\lambda_K}{\lambda_K - 1} \int \mathrm{d}^{K-1}\boldsymbol{x}_{-K}\, x_{K-1} \underbrace{\pi_3(\boldsymbol{x}_{-K}|K)}_{=\,\pi_3(\boldsymbol{x}_{-K}|K-1)} \\ &= \frac{\lambda_K}{\lambda_K - 1} \langle x_{K-1}|K-1\rangle\end{aligned} \tag{A3}$$

The proof of (20) for general $k \in \{1, 2, \ldots, K\}$ is then completed by noting that the special case (A3) actually applies to any value of k, due to the consistency (A1): by marginalizing out all higher frequencies $\{x_{k+1}, x_{k+2}, \ldots, x_k\}$ we obtain $\pi_3(x_1, x_2, \ldots, x_k|K) = \pi_3(x_1, x_2, \ldots, x_k|k)$, to which (A3) then applies.

Appendix C. Python Code for Sampling

```
import numpy as np
import scipy.stats

def sample(x0, xbar, size=1):
    K = len(xbar)
    X = [x0, *xbar]

    # Calculate scale parameters for the u ~ Exp(beta)
    beta = [(X[k+1] - X[k])/X[k+1] for k in range(K)]

    # Draw the u
    u = scipy.stats.expon.rvs(scale=beta, size=(size,K))

    # Transform to x
    x = x0*np.exp(np.cumsum(u, axis=1))

    return x # (size, K)

def sample_truncated(x0, xbar, xmax, size=1):
    def get_batch(size):
        x = sample(x0, xbar, size)
        keep = np.all(x <= xmax, axis=1)
        return x[keep,:]

    accept = get_batch(size)
    p = max(accept.shape[0]/size, 1/20)

    while accept.shape[0] < size:
        new = int((size - accept.shape[0])/p)
        batch = get_batch(new)
        accept = np.concatenate((accept, batch), axis=0)

    return accept[:size,:] # (size, K)
```

Appendix D. Interpretation of π_3 in Acoustic Phonetics

It is easily verified that all three priors π_i are scale-free distributions, i.e., $\pi_i(c\boldsymbol{x}) \propto \pi_i(\boldsymbol{x})$ with $c > 0$ [51]. Assuming the uniform scaling hypothesis [e.g., 52], the scale transformation $\boldsymbol{x} \to c\boldsymbol{x}$ corresponds to a uniform rescaling of the vocal tract such that its length $L \to L/c$.[2] The π_i thus succeed in representing information about the resonance frequencies $\boldsymbol{x}$ in a way that is *independent of the speaker's vocal tract length*, which is the major source of inter-speaker variability after vowel type [54].[3]

This is true even for π_3, despite the increased amount of prior information it would typically represent (Table A1); in general, in the maximum entropy framework, the symmetries of the invariant measure are not preserved under adding constraints as in

[2] For example, [53] has estimated that L for females is about 20% shorter than L for males, and indeed one finds that on average female formants are about 20% higher than male ones [52].

[3] The scale-free criterion $p(cx|\beta) \propto p(x|\beta)$ is not to be confused with requiring invariance of functional form under a given transformation $\{x, \beta\} \to \{x', \beta'\}$ as in Jaynes' transformation invariance principle (see Appendix A); indeed, the former is much more stringent than the latter. In Jaynes' method, invariance of functional form is required under transformations *between problems*, such that the sample and parameter space $\{x, \beta\}$ are transformed simultaneously [55, App. A]. In contrast, the scale-free criterion only involves a transformation of the sample space x (i.e., $x \to cx$), without a possible 'countertransformation' of the parameter space $\beta \to \beta'$ to 'compensate' for the transformation of the sample space. For example, any distribution of the form $p(x|\beta) = (1/\beta)h(x/\beta)$ (e.g., a zero-mean Gaussian) is invariant in form under $\{x, \beta\} \to \{cx, c\beta\}$, while not necessarily scale-free; in fact, the only one-dimensional scale-free distribution is the Pareto distribution (12) [51].

(17). That this is not the case for the invariant measure $m(\boldsymbol{u})$ is due to the fact that the scale invariance is built into the $\boldsymbol{x} \to \boldsymbol{u}$ transformation (15).

Thanks to this built-in quality, the $\boldsymbol{u}$ space is also a "natural space" to describe vowel type information [56]. This fact can be exploited, for example, when constructing a new prior $p(\boldsymbol{x}|x_0, X)$ (where X is a dataset of previously observed $\boldsymbol{x}$ samples such as [31]) which is to be informative (for example to represent prior knowledge that the data will be a open vowel) but still independent of the speaker's vocal tract length. This can be done by transforming X and processing that information in that space through, say, mixture modeling, or maximum entropy density estimation based on empirical moments $\overline{u^k}$ [57], and then transforming the obtained density back to obtain the desired $p(\boldsymbol{x}|x_0, X)$.

Furthermore, we note that the log ratio transformation $u_k = \log(x_k/x_{k-1})$ in (15) exhibits several useful properties which have been disparately observed in the literature of acoustic phonetics. For example, ratios of consecutive frequencies (x_k/x_{k-1}) are the foundation of formant ratio theory [58]. The *log* of these ratios, i.e., u_k, is the preferred representation in Miller's classical theory of vowel perception [59]. The empirical first moments $\overline{\boldsymbol{u}}$ used in (17) also play a role in vowel normalization methods [60], and we note in passing that they avoid the amplification of the error in the frequency in the denominator which is "likely to have hampered efforts to normalize for acoustic scale using formant ratios" [54, p. 2384].[4]

While these connections are of course specific to the domain of acoustic phonetics, we might expect similar advantageous connections in other fields where resonance frequencies play important roles.

Appendix D.1. Another Way of Looking at It

The maximum entropy framework is invariant under transformations, but this does not remove the arbitrariness in choosing which moments to fix. This is similar to the fact that specifying a flat prior in one coordinate frame is not flat in another: we need to find the appropriate coordinate frame, and this choice is 'arbitrary'; i.e., it is not prescribed by probability theory, because it is one of the ways information is encoded into the algebra.

From this point of view, the previous paragraphs of this Appendix not so much interpret π_3 as answer the question, "why fix the particular moments $\langle u_k \rangle = \langle \log(x_k/x_{k-1}) \rangle$ and not any other function of x?" The answer, in short, is that the theoretical properties of the function $\log(x_k/x_{k-1})$ make its expectation value a meaningful quantity to fix, at least within the domain of acoustic phonetics. The fact that the $\overline{u_k} = 1/\lambda_k$ are expressible in terms of something much more likely to be known, i.e., in terms of $\overline{\boldsymbol{x_0}}$, is an additional convenience.

References

1. Skilling, J. Nested Sampling. AIP Conference Proceedings; AIP: Garching (Germany), 2004; Vol. 735, pp. 395–405. doi:10.1063/1.1835238.
2. Green, P.J. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika* **1995**, *82*, 711–732. doi:10.2307/2337340.
3. Shore, J.; Johnson, R. Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy. *IEEE Transactions on Information Theory* **1980**, *26*, 26–37. doi:10.1109/TIT.1980.1056144.
4. Mark, Y.Z.; Hasegawa-johnson, M. Particle Filtering Approach to Bayesian Formant Tracking. Proc. IEEE Workshop on Statistical Signal Processing, 2003.
5. Yanli Zheng.; Hasegawa-Johnson, M. Formant Tracking by Mixture State Particle Filter. 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004, Vol. 1, pp. I–565. doi:10.1109/ICASSP.2004.1326048.
6. Yan, Q.; Vaseghi, S.; Zavarehei, E.; Milner, B.; Darch, J.; White, P.; Andrianakis, I. Formant Tracking Linear Prediction Model Using HMMs and Kalman Filters for Noisy Speech Processing. *Computer Speech & Language* **2007**, *21*, 543–561. doi:10.1016/j.csl.2006.11.001.

[4] The magnitude of the error of the log ratio of two numbers a and b does not depend on whether we divide b by a or a by b. This is not true for the error of the ratio alone, since $\delta \log b/a = \delta b\,/b - \delta a\,/a$ while $\delta(b/a) = (b/a)\,\delta \log b/a$.

7. Mehta, D.D.; Rudoy, D.; Wolfe, P.J. Kalman-Based Autoregressive Moving Average Modeling and Inference for Formant and Antiformant Tracking. *The Journal of the Acoustical Society of America* **2012**, *132*, 1732–1746, [https://doi.org/10.1121/1.4739462]. doi:10.1121/1.4739462.
8. Shi, Y.; Chang, E. Spectrogram-Based Formant Tracking via Particle Filters. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)., 2003, Vol. 1, pp. I–I. doi:10.1109/ICASSP.2003.1198743.
9. Deng, L.; Lee, L.J.; Attias, H.; Acero, A. Adaptive Kalman Filtering and Smoothing for Tracking Vocal Tract Resonances Using a Continuous-Valued Hidden Dynamic Model. *IEEE Transactions on Audio, Speech, and Language Processing* **2007**, *15*, 13–23. doi:10.1109/TASL.2006.876724.
10. Luberadzka, J.; Kayser, H.; Hohmann, V. Glimpsed Periodicity Features and Recursive Bayesian Estimation for Modeling Attentive Voice Tracking. *International Congress on Acoustics* **2019**, *9*, 8.
11. Knuth, K.H.; Habeck, M.; Malakar, N.K.; Mubeen, A.M.; Placek, B. Bayesian Evidence and Model Selection. *Digital Signal Processing* **2015**, *47*, 50–67. doi:10.1016/j.dsp.2015.06.012.
12. Stephens, M. Dealing with Label Switching in Mixture Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2000**, *62*, 795–809. doi:10.1111/1467-9868.00265.
13. Celeux, G.; Kamary, K.; Malsiner-Walli, G.; Marin, J.M.; Robert, C.P. Computational Solutions for Bayesian Inference in Mixture Models. *arXiv:1812.07240 [stat]* **2018**, [arXiv:stat/1812.07240].
14. Celeux, G.; Fruewirth-Schnatter, S.; Robert, C.P. Model Selection for Mixture Models - Perspectives and Strategies. *arXiv:1812.09885 [stat]* **2018**, [arXiv:stat/1812.09885].
15. Bretthorst, G.L. *Bayesian Spectrum Analysis and Parameter Estimation*; 1988.
16. Sivia, D.; Skilling, J. *Data Analysis: A Bayesian Tutorial*; OUP Oxford, 2006.
17. von von der Linden, W.; Dose, V.; von Toussaint, U. *Bayesian Probability Theory: Applications in the Physical Sciences*; Cambridge University Press: Cambridge, 2014. doi:10.1017/CBO9781139565608.
18. Gregory, P. *Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica® Support*; Cambridge University Press: Cambridge, 2005. doi:10.1017/CBO9780511791277.
19. Knuth, K.H.; Skilling, J. Foundations of Inference. *Axioms* **2012**, *1*, 38–73. doi:10.3390/axioms1010038.
20. Jaynes, E.T. Information Theory and Statistical Mechanics. *Physical Review* **1957**, *106*, 620–630.
21. Jaynes, E. Prior Probabilities. *IEEE Transactions on Systems Science and Cybernetics* **1968**, *4*, 227–241. doi:10.1109/TSSC.1968.300117.
22. Kominek, J.; Black, A.W. The CMU Arctic Speech Databases. Fifth ISCA Workshop on Speech Synthesis, 2004.
23. Van Soom, M.; de Boer, B. A New Approach to the Formant Measuring Problem. *Proceedings* **2019**, *33*, 29. doi:10.3390/proceedings201903302
24. Van Soom, M.; de Boer, B. Detrending the Waveforms of Steady-State Vowels. *Entropy* **2020**, *22*, 331. doi:10.3390/e22030331.
25. Skilling, J. Nested Sampling for General Bayesian Computation. *Bayesian Analysis* **2006**, *1*, 833–859. doi:10.1214/06-BA127.
26. Speagle, J.S. Dynesty: A Dynamic Nested Sampling Package for Estimating Bayesian Posteriors and Evidences. *arXiv:1904.02180 [astro-ph, stat]* **2019**, [arXiv:astro-ph, stat/1904.02180].
27. Feroz, F.; Hobson, M.P.; Bridges, M. MULTINEST: An Efficient and Robust Bayesian Inference Tool for Cosmology and Particle Physics. *Monthly Notices of the Royal Astronomical Society* **2009**, *398*, 1601–1614. doi:10.1111/j.1365-2966.2009.14548.x.
28. Neal, R.M. Slice Sampling. *The Annals of Statistics* **2003**, *31*, 705–767. doi:10.1214/aos/1056562461.
29. Handley, W.J.; Hobson, M.P.; Lasenby, A.N. POLYCHORD: Nested Sampling for Cosmology. *Monthly Notices of the Royal Astronomical Society* **2015**, *450*, L61–L65. doi:10.1093/mnrasl/slv047.
30. Handley, W.J.; Hobson, M.P.; Lasenby, A.N. POLYCHORD: Next-Generation Nested Sampling. *Monthly Notices of the Royal Astronomical Society* **2015**, *453*, 4384–4398. doi:10.1093/mnras/stv1911.
31. Peterson, G.E.; Barney, H.L. Control Methods Used in a Study of the Vowels. *The Journal of the acoustical society of America* **1952**, *24*, 175–184.
32. Hillenbrand, J.; Getty, L.A.; Clark, M.J.; Wheeler, K. Acoustic Characteristics of American English Vowels. *The Journal of the Acoustical society of America* **1995**, *97*, 3099–3111.
33. Vallée, N. *Systèmes Vocaliques: De La Typologie Aux Prédictions*; Thèse préparée au sein de l'Institut de la Communication Parlée (Grenoble-URA C.N.R.S. no 368), 1994.
34. Kent, R.D.; Vorperian, H.K. Static Measurements of Vowel Formant Frequencies and Bandwidths: A Review. *Journal of Communication Disorders* **2018**, *74*, 74–97. doi:10.1016/j.jcomdis.2018.05.004.
35. Vorperian, H.K.; Kent, R.D.; Lee, Y.; Bolt, D.M. Corner Vowels in Males and Females Ages 4 to 20 Years: Fundamental and F1–F4 Formant Frequencies. *The Journal of the Acoustical Society of America* **2019**, *146*, 3255–3274. doi:10.1121/1.5131271.
36. Klatt, D.H. Software for a Cascade/Parallel Formant Synthesizer. *The Journal of the Acoustical Society of America* **1980**, *67*, 971–995. doi:10.1121/1.383940.
37. Titze, I.R.; Baken, R.J.; Bozeman, K.W.; Granqvist, S.; Henrich, N.; Herbst, C.T.; Howard, D.M.; Hunter, E.J.; Kaelin, D.; Kent, R.D.; Kreiman, J.; Kob, M.; Löfqvist, A.; McCoy, S.; Miller, D.G.; Noé, H.; Scherer, R.C.; Smith, J.R.; Story, B.H.; Švec, J.G.; Ternström, S.; Wolfe, J. Toward a Consensus on Symbolic Notation of Harmonics, Resonances, and Formants in Vocalization. *The Journal of the Acoustical Society of America* **2015**, *137*, 3005–3007. doi:10.1121/1.4919349.
38. de Boer, B. Acoustic Tubes with Maximal and Minimal Resonance Frequencies. *The Journal of the Acoustical Society of America* **2008**, *123*, 3732–3732. doi:10.1121/1.2935231.
39. Bretthorst, G.L. Bayesian Analysis. II. Signal Detection and Model Selection. *Journal of Magnetic Resonance* **1990**, *88*, 552–570.

40. Wilson, A.G.; Wu, Y.; Holland, D.J.; Nowozin, S.; Mantle, M.D.; Gladden, L.F.; Blake, A. Bayesian Inference for NMR Spectroscopy with Applications to Chemical Quantification. *arXiv:1402.3580 [stat]* **2014**, [arXiv:stat/1402.3580].
41. Xu, K.; Marrelec, G.; Bernard, S.; Grimal, Q. Lorentzian-Model-Based Bayesian Analysis for Automated Estimation of Attenuated Resonance Spectrum. *IEEE Transactions on Signal Processing* **2019**, *67*, 4–16. doi:10.1109/TSP.2018.2878543.
42. Trassinelli, M. Bayesian Data Analysis Tools for Atomic Physics. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* **2017**, *408*, 301–312. doi:10.1016/j.nimb.2017.05.030.
43. Fancher, C.M.; Han, Z.; Levin, I.; Page, K.; Reich, B.J.; Smith, R.C.; Wilson, A.G.; Jones, J.L. Use of Bayesian Inference in Crystallographic Structure Refinement via Full Diffraction Profile Analysis. *Scientific Reports* **2016**, *6*, 31625. doi:10.1038/srep31625.
44. Allison, J.R.; Sadler, E.M.; Whiting, M.T. Application of a Bayesian Method to Absorption Spectral-Line Finding in Simulated ASKAP Data. *Publications of the Astronomical Society of Australia* **2012/ed**, *29*, 221–228. doi:10.1071/AS11040.
45. Littenberg, T.B.; Cornish, N.J. Bayesian Inference for Spectral Estimation of Gravitational Wave Detector Noise. *Physical Review D* **2015**, *91*, 084034. doi:10.1103/PhysRevD.91.084034.
46. Hanna, N.; Smith, J.; Wolfe, J. Frequencies, Bandwidths and Magnitudes of Vocal Tract and Surrounding Tissue Resonances, Measured through the Lips during Phonation. *The Journal of the Acoustical Society of America* **2016**, *139*, 2924–2936. doi:10.1121/1.4948754.
47. Wilson, A.; Adams, R. Gaussian Process Kernels for Pattern Discovery and Extrapolation. International Conference on Machine Learning, 2013, pp. 1067–1075.
48. Xiang, N. Model-Based Bayesian Analysis in Acoustics-A Tutorial. *The Journal of the Acoustical Society of America* **2020**. doi:10.1121/10.0001731.
49. Jaynes, E.T. *Probability Theory: The Logic of Science*; Cambridge University Press: Cambridge, UK; New York, NY, 2003.
50. Bretthorst, G.L. Bayesian Spectrum Analysis on Quadrature NMR Data with Noise Correlations. In *Maximum Entropy and Bayesian Methods*; Springer, 1989; pp. 261–273.
51. Newman, M.E.J. Power Laws, Pareto Distributions and Zipf's Law. *Contemporary Physics* **2005**, *46*, 323–351. doi:10.1080/00107510500052444.
52. Fant, G. A Note on Vocal Tract Size Factors and Non-Uniform F-Pattern Scalings. *Speech Transmission Laboratory Quarterly Progress and Status Report* **1966**, *1*, 22–30.
53. Goldstein, U.G. An Articulatory Model for the Vocal Tracts of Growing Children. PhD thesis, Massachusetts Institute of Technology, 1980.
54. Turner, R.E.; Walters, T.C.; Monaghan, J.J.M.; Patterson, R.D. A Statistical, Formant-Pattern Model for Segregating Vowel Type and Vocal-Tract Length in Developmental Formant Data. *The Journal of the Acoustical Society of America* **2009**, *125*, 2374–2386. doi:10.1121/1.3079772.
55. Jaynes, E.T. Marginalization and Prior Probabilities (1980). In *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*; Rosenkrantz, R.D., Ed.; Springer Netherlands: Dordrecht, 1989; pp. 337–375. doi:10.1007/978-94-009-6581-2_12.
56. Turner, R.E.; Patterson, R.D. An Analysis of the Size Information in Classical Formant Data: Peterson and Barney (1952) Revisited. *J. Acoust. Soc. Jpn* **2003**, *33*, 585–589.
57. Bretthorst, G.L. The Maximum Entropy Method of Moments and Bayesian Probability Theory **2013**. *3*, 3–15. doi:10.1063/1.4819977.
58. Lloyd, R.J. Speech Sounds: Their Nature and Causation (I). *Phonetische Studien* **1890**, *3*, 251–278.
59. Miller, J.D. Auditory-perceptual Interpretation of the Vowel. *The Journal of the Acoustical Society of America* **1987**, *81*, S16–S16. doi:10.1121/1.2024119.
60. Johnson, K. Vocal Tract Length Normalization. *UC Berkeley PhonLab Annual Report* **2018**, *14*.