Gaussian Processes for Underdetermined Source Separation

Antoine Liutkus, Roland Badeau, *Senior Member, IEEE,* Gaël Richard, *Senior Member, IEEE*

***Abstract*—Gaussian process (GP) models are very popular for machine learning and regression and they are widely used to account for spatial or temporal relationships between multivariate random variables. In this paper, we propose a general formulation of underdetermined source separation as a problem involving GP regression. The advantage of the proposed unified view is firstly to describe the different underdetermined source separation problems as particular cases of a more general framework. Secondly, it provides a flexible means to include a variety of prior information concerning the sources such as smoothness, local stationarity or periodicity through the use of adequate covariance functions. Thirdly, given the model, it provides an optimal solution in the minimum mean squared error (MMSE) sense to the source separation problem. In order to make the GP models tractable for very large signals, we introduce *framing* as a GP approximation and we show that computations for *regularly sampled* and *locally stationary* GPs can be done very efficiently in the frequency domain. These findings establish a deep connection between GP and Nonnegative Tensor Factorizations with the Itakura-Saito distance and lead to effective methods to learn GP hyperparameters for very large and regularly sampled signals.**

***Index Terms*—Gaussian Processes, NMF, NTF, Source Separation, Probability Theory, Regression, Kriging, Cokriging**

I. Introduction

Gaussian processes [28], [35], [36], [44] are commonly used to model functions whose mean and covariances are known. Given some learning points, they enable us to estimate the values taken by the function at any other points of interest. Their main advantages are to provide a simple and effective probabilistic framework for regression and classification as well as an effective means to optimize a model's parameters through maximization of the *marginal likelihood* of the observations. For these reasons, they are widely used in many areas to model dependencies between multivariate random variables and their use can be traced back at least to works by Wiener in 1941 [43]. They have also been known in geostatistics under the name of *kriging* for almost 40 years [29]. A great surge of interest for Gaussian Process (GP) models occurred when they were expressed as a general purpose framework for regression as well as for classification (see [35] for a review). Their relation to other methods commonly used in machine learning such as multi-layer perceptrons, spline interpolation or support vector machines are now well understood.

Copyright (c) 2011 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

Authors are with Institut Telecom, Telecom ParisTech, CNRS LTCI, France.
This work is partly funded by the French National Research Agency (ANR) as a part of the DReaM project (ANR-09-CORD-006-03) and partly supported by the Quaero Programme, funded by OSEO, French State agency for innovation.

Source separation is another very intense field of research (see [10] for a review) where the objective is to recover several unknown signals called *sources* that were mixed together in observable *mixtures*. Source separation problems arise in many fields such as sound processing, telecommunications and image processing. They differ mainly in the relative number of mixtures per source signal and in the nature of the mixing process. The latter is generally modeled as *convolutive*, i.e. as a linear filtering of the sources into the mixtures. When the mixing filters reduce to a single amplification gain, the mixing is called *instantaneous*. When there are more mixtures than sources, the problem is called *overdetermined* and algorithms may rely on beamforming techniques to perform source separation. When there are fewer mixtures than sources, the problem is said to be *underdetermined* and is notably known to be very difficult. Indeed, in this case there are less observable signals than necessary to solve the underlying mixing equations. Many models were hence studied to address this problem and they all either restrict the set of possible source signals or assign prior probabilities to them in a Bayesian setting. Among the most popular approaches, we can mention Independent Component Analysis [6] that focuses both on probabilistic independence between the source signals and on high order statistics. We can also cite Non-negative Matrix Factorization (NMF) source separation that models the sources as locally stationary with constant normalized power spectra and time-varying energy [16], [27].

In this study, we revisit underdetermined source separation (USS) as a problem involving GP regression. To our knowledge, no unified treatment of the different underdetermined linear source separation problems in terms of classical GP is available to date and we thus propose here an attempt at providing such a formulation whose advantages are numerous. Firstly, it provides a unified framework for handling the different USS problems as particular cases, including convolutive or instantaneous mixing as well as single or multiple mixtures. Secondly, when prior information such as smoothness, local stationarity or periodicity is available, it can be taken into account through appropriate *covariance functions*, thus providing a significant expressive power to the model. Thirdly, it yields an optimal way in the minimum mean squared error (MMSE) sense to proceed to the separation of the sources given the model.

In spite of all their interesting features, GP models come at a high $\mathcal{O}\left(n^3\right)$ computational cost where n is the number of training points. For many applications such as audio signal processing where $n \approx 10^7$ is common, this cost is prohibitive. Hence, the GP framework has to come along with effective methods to simplify the computations in order to be of

practical use. Over the years, many approximation methods have been proposed [31], [34], [37], [38], [40] to address this issue and we show that the common practice of *framing* in audio signal processing can precisely be understood in terms of GP modeling as a particular choice for GP approximation. In particular, we give its connections with recently published Partially Independent Conditional (PIC) approximation [37] and Compact Support (CS) covariance functions [31], [40]. For the special case of locally stationary and regularly sampled signals, we furthermore show that computations can be performed extremely efficiently in the frequency domain and we establish a novel connection between GP models and the emerging techniques of Nonnegative Tensor Factorizations (NTF) [9] using the Itakura-Saito divergence.

The article is organized as follows. First, we present GP and particularly Gaussian Process Regression (GPR) in section II. Then, we set out the various linear underdetermined source separation problems in terms of GPR in section III. In order to make the GP models tractable for very large signals, we introduce *framing* as a GP approximation and we show that computations for *regularly sampled* and *locally stationary* GPs can be done very efficiently in the frequency domain in section IV. Finally, we illustrate the performance of the methods on synthetic and real data in section V and draw some conclusions in section VI.

II. Gaussian processes

A. Introduction

A Gaussian process [28], [35], [36], [44] is a possibly infinite set of scalar random variables $\{f(x)\}_{x\in\mathcal{X}}$ indexed by an *input space* $\mathcal{X}$, typically $\mathcal{X}=\mathbb{R}^D$, and taking values on $\mathbb{R}$, such that for any *finite* set of inputs $X=\{x_1\cdots x_n\}\in\mathcal{X}^n$, $\mathbf{f}\triangleq[f(x_1)\cdots f(x_n)]^\top$ is distributed with respect to a multivariate Gaussian distribution[1]. A GP is thus completely determined by a mean function $m(x)=\mathbb{E}[f(x)]$ and a covariance function $k(x,x')=\mathbb{E}[(f(x)-m(x))(f(x')-m(x'))]$.

More fundamentally, a GP may be understood as a process whose mean $m(x)$ and covariance $k(x,x')$ between any two inputs are known. Given only this prior information, assigning a multivariate Gaussian distribution to $\mathbf{f}$ given any finite set X of inputs from $\mathcal{X}$ is a sensible choice, since it is the probability distribution that maximizes entropy when only the first two moments are known [23].

It has been shown that the class of valid covariance functions coincides with the class of positive definite functions [1]. Let X be a finite set of elements from $\mathcal{X}$ that is possibly randomly drawn as in [19], the covariance matrix $K_{f,XX}$ is defined as $[K_{f,XX}]_{i,j}=k(x_i,x_j)$ and the probability of $\mathbf{f}$ given X is then given by[2]:

$$p(\mathbf{f}\mid X)=\frac{1}{(2\pi)^{\frac{n}{2}}\left|K_{f,XX}\right|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{f}-\mathbf{m})^\top K_{f,XX}^{-1}(\mathbf{f}-\mathbf{m})\right) \tag{1}$$

[1]The symbol $\triangleq$ denotes a definition.

[2]Positive *semi*-definite covariance matrices are possible. In the case of singular $K_{f,XX}$, a characterization involving the characteristic function instead of (1) is required.

where $\mathbf{m}\triangleq[m(x_1)\cdots m(x_n)]^\top$. This is usually written:

$$f\sim\mathcal{GP}(m(x),k(x,x'))$$

Most studies in underdetermined source separation focus on the *single sensor* scenario $\mathcal{X}=\mathbb{R}$. Still, there is no difficulty involved in considering the general case $\mathcal{X}=\mathbb{R}^D$ and we will see examples of GPs defined on a multidimensional input space in sections III-C and IV. This framework thus easily allows modeling multivariate functions defined on arbitrary input spaces and many studies have used Gaussian processes for regression ($f(x)\in\mathbb{R}$) as well as for classification ($f(x)\in\mathbb{N}$). Their main advantages are to provide a probabilistic interpretation and a way to compute the variances of the estimates. From now on, we will focus on GPR, since our objective is to highlight the connections between GP and source separation, which is usually stated in terms of processes taking values in $\mathbb{R}$. For the sake of notational simplicity, we will assume *a priori* centered signals, i.e. $\forall x\in\mathcal{X}, m(x)=0$, as it is very common for audio signals. Still, there is no particular issue raised when considering arbitrary mean functions.

B. Gaussian processes regression

Suppose we observe $y(x)=f(x)+\epsilon(x)$, with $f(x)$ being the signal of interest and $\epsilon(x)$ being some additive signal — usually called *noise* — that is independent from $f(x)$, for a finite set X of input points from $\mathcal{X}$: $X=\{x_1\cdots x_n\}\in\mathcal{X}^n$. We want to estimate the values taken by f on a finite and possibly different set $X^*=\{x_1^*\cdots x_{n^*}^*\}\in\mathcal{X}^{n^*}$ of input points from $\mathcal{X}$. Let us furthermore assume that $f\sim\mathcal{GP}(0,k_f(x,x'))$ and $\epsilon\sim\mathcal{GP}(0,k_\epsilon(x,x'))$ where the covariance functions k_f and k_ϵ are known. As f and ϵ are supposed independent, we have:

$$f+\epsilon\sim\mathcal{GP}(0,k_f(x,x')+k_\epsilon(x,x')) \tag{2}$$

Let K_{f,XX^*} be the covariance matrix defined by $[K_{f,XX^*}]_{ij}=k_f(x_i,x_j^*)$. We define K_{f,X^*X}, K_{f,X^*X^*}, $K_{\epsilon,XX}$ in the same way. Let $\mathbf{f}\triangleq[f(x_1)\cdots f(x_n)]^\top$, $\mathbf{f}^*\triangleq[f(x_1^*)\cdots f(x_n^*)]^\top$ and similarly for $\mathbf{y}$. We have:

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{f}^*\end{bmatrix}\sim\mathcal{N}\left(\mathbf{0},\begin{bmatrix}K_{f,XX}+K_{\epsilon,XX} & K_{f,XX^*}\\ K_{f,X^*X} & K_{f,X^*X^*}\end{bmatrix}\right)$$

Classical probability results then assert that the conditional distribution of $\mathbf{f}^*$ given $\mathbf{y}$ is (see [35]):

$$\mathbf{f}^*\mid\mathbf{y}\sim\mathcal{N}\left(\overline{\mathbf{f}^*},\mathrm{cov}\mathbf{f}^*\right) \tag{3}$$

with[3]:

$$\overline{\mathbf{f}^*}=K_{f,X^*X}\left[K_{f,XX}+K_{\epsilon,XX}\right]^{-1}\mathbf{y} \tag{4}$$

and

$$\mathrm{cov}\mathbf{f}^*=K_{f,X^*X^*}-K_{f,X^*X}\left[K_{f,XX}+K_{\epsilon,XX}\right]^{-1}K_{f,XX^*} \tag{5}$$

These expressions show that the maximum likelihood estimate $\hat{\mathbf{f}}^*$ of $\mathbf{f}^*\mid\mathbf{y}$ is found by setting $\hat{\mathbf{f}}^*=\overline{\mathbf{f}^*}$, which is also the Minimum Mean Squared Error (MMSE) estimate in the Gaussian case. This result will be fundamental when performing source separation using Gaussian processes. We can furthermore compute the covariance of the estimates.

[3]In the case of singular covariance matrix $K_{f,XX}+K_{\epsilon,XX}$, numerical methods such as Moore-Penrose pseudo-inversion may be used.

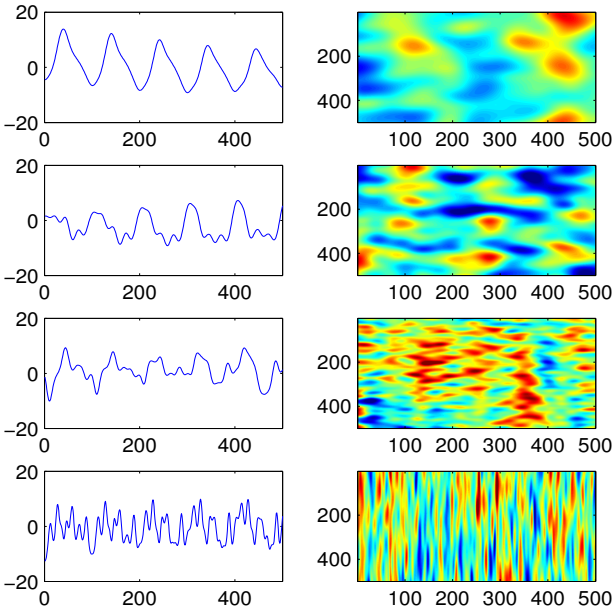

Figure 1. Typical realizations of GP with SE and periodic covariance functions with different values of the hyperparameters. On the left column, $D = 1$ and the processes are only parameterized by 3 scalars. On the right, $D = 2$ and the processes are parameterized by 6 scalars.

C. Covariance functions

Many studies (see [1] for a review) concern valid covariance functions. They belong to the general family of *kernels* and must be definite positive. Some properties of interest have been demonstrated:

- Sums and products of valid covariance functions are valid covariance functions.
- When it is *stationary*, i.e. when it can be expressed as a function of $\tau = x - x'$, then the covariance function can be parameterized by its Fourier transform.

Two examples of covariance functions for $\mathcal{X} = \mathbb{R}^D$ are:

- The Squared Exponential (SE) covariance function defined by: $k_{\text{SE}}(x, x'|\sigma, M) = \sigma^2 \exp\left(-\frac{(x-x')^\top M (x-x')}{2}\right)$ with $\sigma^2 > 0$ and M positive semidefinite. When $D = 1$, we have $k_{\text{SE}}(x, x'|\sigma, \lambda) = \sigma^2 \exp\left(-\frac{(x-x')^2}{2\lambda^2}\right)$. λ is called a *characteristic length scale* in the sense that $|x - x'| \gg \lambda$ is required for two points x and x' of the process to be independent.
- The less common periodic covariance function of period T given by: $k_{\text{periodic}}(x, x'|T, \lambda) = \sigma^2 \exp\left(-\frac{2\sin^2 \frac{\pi(x-x')}{T}}{\lambda^2}\right)$.

As can be seen, covariance functions are generally parameterized by a set of scalar values such as their Fourier transform when they are stationary, their characteristic length scales, their period, etc. These scalars are often called *hyperparameters* and are usually gathered in a *hyperparameter set* Θ. Typical realizations of GPs with SE and periodic covariance functions are given for $D = 1$ and $D = 2$ in Figure 1.

D. Optimization of the hyperparameters

In a Bayesian context, we may need to find the hyperparameters that maximize the marginal likelihood of the observations. In other words, we may need to find Θ^* such that $p(\mathbf{y} \mid X, \Theta^*)$ is maximum. Indeed, even if we may well guess the covariance functions that are adequate to the problem at hand, such as stationary covariance functions parameterized by their Fourier transform or SE covariance functions, it is likely that the hyperparameters that best explain the observations are not exactly known.

To this purpose, we can compute the closed-form expression of the marginal log-likelihood of the observations, $\log p(\mathbf{y} \mid X, \Theta)$ as (see [35]):

$$\log p(\mathbf{y} \mid X, \Theta) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\mathbf{y}^\top \left[K_{f,XX} + K_{\epsilon,XX}\right]^{-1}\mathbf{y} - \frac{1}{2}\log\left|K_{f,XX} + K_{\epsilon,XX}\right| \tag{6}$$

where each covariance matrix depends on Θ. Using the opposite of (6) as a cost function, we can proceed to the optimization of the hyperparameters using classical optimization algorithms, in a principled probabilistic framework. Note that depending on the covariance function and the hyperparameter considered, the corresponding optimization problem may or may not be convex.

III. Gaussian processes for source separation

A. Single mixture with instantaneous mixing

The presentation of GPR given in section II-B is actually slightly more general than what is usual in the literature. Indeed, it is often assumed that the covariance function k_ϵ of the additive signal ϵ is given by $k_\epsilon(x, x') = \sigma^2 \delta_{xx'}$ where $\delta_{xx'} = 1$ if and only if $x = x'$ and zero otherwise. This assumption corresponds to additive independent and identically distributed (i.i.d.) white Gaussian noise of variance σ^2.

In our presentation, the additive signal $\epsilon(x)$ is a GP itself and is potentially very complex. In any case, its covariance function is given by k_ϵ and the only assumption made is its independence with the signal of interest $f(x)$. A particular example of a model where k_ϵ non trivially depends on x and x' was for example studied in [20].

The results obtained can very well be generalized to the situation where y is the sum of M independent latent Gaussian processes:

$$\forall x \in \mathcal{X}, y(x) = \sum_{m=1}^{M} f_m(x)$$

with

$$f_m \sim \mathcal{GP}(0, k_m(x, x'))$$

In this case, if our objective is to extract the signal corresponding to the source m_0, we only need to replace k_f with k_{m_0} and k_ϵ with $\sum_{m \neq m_0} k_m$ in section II-B. Note that inversion of $K_{f,XX} + K_{\epsilon,XX}$ is needed only once for the extraction of all sources. Similarly, we can also jointly

optimize the hyperparameters of all covariance functions using exactly the same framework as in section II-D. We now consider the case of convolutive mixtures of independent GPs.

B. Single mixture with convolutive mixing

An important fact, which has already been noticed in the literature [2], [4], is that the convolution of a GP, as a linear combination of Gaussian random variables, remains a GP. Indeed, let us consider some GP $f_{0,m} \sim \mathcal{GP}\left(0, k_{0,m}\left(x, x'\right)\right)$ and let us define

$$f_m\left(x\right) = \int_{\mathcal{X}} a_m\left(x - z\right) f_{0,m}\left(z\right) dz \triangleq \left(a_m * f_{0,m}\right)\left(x\right)$$

where $a_m : \mathcal{X} \to \mathbb{R}$ is a stable *mixing filter* from $f_{0,m}$ to f_m. If the mean function of $f_{0,m}$ is identically 0, the mean function of f_m is easily seen to also be identically 0. The covariance function of f_m can be computed as $k_m\left(x, x'\right) = \mathbb{E}\left[f_m\left(x\right) f_m\left(x'\right)\right]$, that is:

$$k_m\left(x, x'\right) = \int_{\mathcal{X}} \int_{\mathcal{X}} a_m\left(x - z\right) a_m\left(x' - z'\right) k_{0,m}\left(z, z'\right) dz dz'$$

which is the convolution of $k_{0,m}$ by $a_m \times a_m \triangleq \left(x, x'\right) \in \mathcal{X}^2 \mapsto a_m\left(x\right) a_m\left(x'\right)$:

$$k_m\left(x, x'\right) = \left(\left(a_m \times a_m\right) * k_{0,m}\right)\left(x, x'\right) \tag{7}$$

Moreover, if several convolved GPs $\left\{f_m = \left(a_m * f_{0,m}\right)\right\}_{m=1\cdots M}$ are summed up in a mixture, it can readily be shown that the f_m are independent if the $f_{0,m}$ are independent. We thus get back to the instantaneous mixing model using modified covariance functions (7).

C. Multiple output GP

We have for now only considered GPs whose outputs lie in $\mathbb{R}$. A sizable body of literature focuses on possible extensions of this framework to cases where the processes of interest are multiple-valued, i.e. whose outputs lie in $\mathbb{R}^C$ for $C \in \mathbb{N}^*$. In geostatistics for example, important applications comprise the modeling of co-occurrences of minerals or pollutants in a spatial field. First attempts in this direction [24] include the so-called *linear model of coregionalization*, that considers each output as a linear combination of some latent processes. The name of *cokriging* has often been used for such systems in the field of geostatistics. If the latent processes are assumed to be GPs, the outputs are also GPs.

In the machine learning community, multiple-output GPs have been introduced [5] and popularized under the name of *dependent* GPs. Several extensions of such models have been proposed subsequently [2]–[4], [30] and we focus here on the model presented in [2] which is very close to the usual convolutive mixing model commonly used in multi-channel source separation, e.g. in [32].

Let $\left\{y_c\left(x\right)\right\}_{c=1\cdots C}$ be the C output signals called the *mixtures*. The *convolutive* GP model consists in assuming that each observable signal y_c is the sum of convolved versions of M latent GPs of interest $\left\{f_{0,m} \sim \mathcal{GP}\left(0, k_{0m}\left(x, x'\right)\right)\right\}_{m=1\cdots M}$ that we will call *sources*, plus one specific additional term $\epsilon_c \sim \mathcal{GP}\left(0, k_{\epsilon c}\left(x, x'\right)\right)$ that is often referred to as *additive noise*. We thus have:

$$y_c\left(x\right) = \sum_{m=1}^{M} \left(a_{cm} * f_{0,m}\right)\left(x\right) + \epsilon_c\left(x\right) \tag{8}$$

Instead of making a fundamental distinction between c and x, the GP framework allows us to consider that $\left\{y_c\left(x\right)\right\}_{(c,x)\in\{1\cdots C\}\times\mathcal{X}}$ is a single signal $\left\{y\left(x'\right)\right\}_{x'\in\{1\cdots C\}\times\mathcal{X}}$ indexed on an extended input space $\{1\cdots C\} \times \mathcal{X}$. If we assume that the different underlying sources $\left\{f_{0,m}\right\}_{m=1\cdots M}$ are independent, which is frequent in source separation and that the different $\left\{\epsilon_c\right\}_{c=1\cdots C}$ are also independent, we can express the covariance function $k\left(\left(c, x\right), \left(c', x'\right)\right)$ of y for two extended input points $\left(c, x\right)$ and $\left(c', x'\right)$ as:

$$k_{cc'}\left(x, x'\right) = \left(\sum_{m=1}^{M} k_{cc',m} + \delta_{cc'} k_{\epsilon c}\right)\left(x, x'\right) \tag{9}$$

$$\text{where } k_{cc',m}\left(x, x'\right) \triangleq \left(\left(a_{cm} \times a_{c'm}\right) * k_{0,m}\right)\left(x, x'\right) \tag{10}$$

For any given c, the different $\left\{f_{cm} \triangleq a_{cm} * f_{0,m}\right\}_{m=1\cdots M}$ are independent and are GPs with mean functions 0 and covariance functions $k_{cc,m}\left(x, x'\right)$. f_{cm} will be called the *contribution* of source m to mixture c. We can readily perform source separation on $\mathbf{y_c}$ to recover the different $\left\{\mathbf{f}_{cm}\right\}_{m=1\cdots M}$ using the standard formalism presented in section II-B. Let $\hat{\mathbf{f}}_{cm_0}$ be the estimate of $\mathbf{f}_{cm_0}$, we have:

$$\hat{\mathbf{f}}_{cm_0} = K_{cc,m_0}\left[\sum_{m=1}^{M} K_{cc,m} + K_{cc,\epsilon}\right]^{-1} \mathbf{y}_c \tag{11}$$

where $K_{cc,\epsilon}$ is the covariance matrix of the additive signal ϵ_c and where the covariance matrix $K_{cc,m}$ is defined as $\left[K_{cc,m}\right]_{x,x'} = k_{cc,m}\left(x, x'\right)$.

It is important to note here that even if the *sources* are the $\left\{f_{0,m}\right\}_{m=1\cdots M}$, many systems consider the signals of interest to actually be the different $\left\{f_{cm}\right\}_{c,m}$. For example, in the case of audio source separation, a stereophonic mixture can be composed of several monophonic sources such as voice, piano and drums. It is often considered sufficient to be able to separate the different instruments *within* the stereo mixtures and thus to obtain one stereo signal for each source, rather than trying to recover the original monophonic signals.

Still, for some m, given the estimates $\left\{\hat{\mathbf{f}}_{cm}\right\}_{c=1\cdots C}$ of all the different $\left\{\mathbf{f}_{cm}\right\}_{c=1\cdots C}$, we can for example estimate $\mathbf{f}_{0,m}$ using standard beamforming techniques.

D. Parameter optimization

Even in complex situations such as those presented in sections III-B or III-C, we can still use classical optimization methods to maximize the marginal log-likelihood (6) of the observations. Following [2], we will now give a simple way to include multiple output GPs in this framework.

Given a set X of n input points and the corresponding C column vectors $\left\{\mathbf{y}_c\right\}_{c=1\cdots C}$, we can build $\mathbf{y} \triangleq \left[\mathbf{y}_1^\top, \ldots, \mathbf{y}_C^\top\right]^\top$ as the Cn column vector containing all stacked outputs and use the expression (9) to build its covariance matrix K. We can ten proceed to parameters estimation through maximization of the

marginal log-likelihood $\log p(\mathbf{y} \mid X, \Theta)$ of the observations. Once more, depending on the covariance functions considered, this problem may or may not be convex.

E. Conclusion

In this section, we have derived a way to perform underdetermined source separation using GP models when the mixtures are the convolved sums of several independent GPs. Given some covariance functions and mixing filters, we saw that stating the problem in terms of GPs provides a principled way to estimate the source signals that minimize the mean squared error. In the GP framework, optimization of the hyperparameters is done through maximization of the marginal log-likelihood of the mixtures given the model.

To our knowledge, very few references are available to date on the topic. For example, [33] performs source separation using GPs in the determined case, but the covariance functions are therein applied on the outputs of the source signals rather than on the coordinates themselves (i.e. time or spatial position). A successful application of GPs to a subject close to source separation can also be found in [39] for echo cancellation.

IV. GP APPROXIMATIONS FOR LARGE SIGNALS

A. The need for approximations

The main issue with GP models is the need to invert the $n \times n$ covariance matrix of the learning points for inference (4) and for each evaluation of the observation likelihood in (6). In many areas of interest, we cannot afford to handle such a big matrix, since it is not computationally tractable. In audio signal processing for example, values such as $n \approx 10^7$ are common and GP models cannot be used without a significant reduction of the computational cost of the method.

In order to address this issue, many authors have proposed *sparse* approximation techniques [31], [34], [37], [38], [40] over the years that all aim at making GP inference possible for large datasets. As highlighted in [34], many methods rely on the choice of a small set of input points called *the inducing inputs* to approximate the posterior distribution at test points X^*. Among those methods, we can mention the Fully Independent Conditional (FIC) approximation [34], [38], that considers all the test points and the learning points independent given the inducing inputs. This leads to a very important reduction of the computational burden, but heavily relies on the density of the inducing points [37] to yield good estimates. Another approximation called Partially Independent Conditional (PIC) [37] no longer makes the assumption that both the training and test cases are independent given the inducing points, but rather that each of them not only depends on the possibly remote inducing points, but also on a limited number of other learning points nearby. This technique has the advantage of producing better estimates than FIC, while maintaining an easy inversion of the $n \times n$ covariance matrix that is now block-diagonal. Its main disadvantage is to lead to discontinuities of the estimates between the blocks, which may be problematic for some applications such as audio processing.

Another very attractive direction of research in the last few years has been the consideration of covariance functions with Compact Support (CS) [31], [40], i.e covariance functions $k(x, x')$ such that $\|x - x'\| > l \Rightarrow k(x, x') = 0$ for some given scale l. The idea underlying these techniques is to consider that if they are sufficiently far from each other, two points will be independent. If such covariance functions are used, the covariance matrix is sparse and inference through Cholesky decompositions is done much faster [40]. The main issue with this approach is to design covariance functions that correspond to some prior knowledge about the sources and that have CS at the same time.

In sections IV-B and IV-C, we introduce a general method for fast inference in GP models based on *framing* and that is a direct generalization of the common practice in audio signal processing.

Another important computational simplification is introduced in sections IV-D and IV-E when the signals are regularly sampled. In that case, we show that when the covariance functions are assumed stationary and separable, exact inference can be done extremely efficiently in the frequency domain.

When both approaches are combined into so called *locally-scaled* and *framewise-independent* stationary covariance functions, we show in section IV-F that inference and learning of hyperparameters become equivalent to recent and powerful Nonnegative Tensor Factorization (NTF) techniques [9].

B. Frames

In audio signal processing, it is common to split the signals into overlapping *frames* and to process the frames separately. Formally, the frames $\{y_i(x')\}_{i \in \mathbb{N}}$ are defined as small portions of the original signal. The advantage of the technique is that the frames are small and can be easily processed. The original signals can then be recovered through a deterministic *overlap-add* procedure: each frame is multiplied by a *weighting function* $g : \mathcal{X}' \to \mathbb{R}^+$ to ensure smooth transitions between the frames and is added to the reconstructed signal. g is often a HANN or a triangular window.

This idea can very well be generalized in any dimension D. Instead of considering the original signal y, we can split it into overlapping frames of smaller dimension. To this end, we consider a *frame input set* $\mathcal{X}' \subset \mathcal{X}$, a summable *weighting function* $g : \mathcal{X}' \to \mathbb{R}^+$ and a set of *frame positions* $\{t_i \in \mathcal{X}\}_{i \in \mathbb{N}}$ such that:

$$\forall x \in \mathcal{X}, I_x \triangleq \{i \in \mathbb{N} : x - t_i \in \mathcal{X}'\} \neq \varnothing \tag{12}$$

I_x is thus the set of frame numbers to which the input point x is mapped. Condition (12) ensures that each point of the signal is represented in at least one frame. Finally, given some signal $\{y(x)\}_{x \in \mathcal{X}}$, we can make the assumption that there is a set of *frames* $\{y_i(x)\}_{i \in \mathbb{N}, x \in \mathcal{X}'}$, also noted $\mathcal{G}\{y\}$ in the following, such that:

$$\forall x \in \mathcal{X}, y(x) = \frac{1}{\sum_{i \in I_x} g(x - t_i)} \sum_{i \in I_x} g(x - t_i)\, y_i(x - t_i) \tag{13}$$

When considering a finite set X of n input points, we only need to consider the frames $I \triangleq \bigcup \{I_x\}_{x \in X}$

that contain at least one input point from X. Let $X'_i = \{x' \in \mathcal{X}' \mid \exists x \in X : x' = x - t_i\}$ be the finite set of points from $\mathcal{X}'$ to which the elements of X in the scope of frame i are mapped and let[4] $L_i = \#(X'_i)$. Given some signal $\{\mathbf{y}(x)\}_{x \in X}$, it is always possible to build a set of frames obeying (13). This can be achieved by choosing X'_i and $\mathbf{y}_i(x)$ such that:

$$\forall (i, x') \in I \times \mathcal{X}', (t_i + x' \in X) \Rightarrow \begin{cases} x' \in X'_i \\ \mathbf{y}_i(x') = \mathbf{y}(t_i + x') \end{cases} \tag{14}$$

When they make use of framing, usual methods focus on the frames $\mathcal{G}\{y\}$ as the signals of interest rather than on y. Indeed, a good model for $\mathcal{G}\{y\}$ is *de facto* a good model for y since it can be computed deterministically from $\mathcal{G}\{y\}$. In such methods based on framing, the set (14) of frames is usually taken as being the observation. From our point of view, the frames are simply another process which is indexed on $\mathbb{N} \times \mathcal{X}'$ and from which we can deterministically recover y which is indexed on $\mathcal{X}$.

C. Frame-wise independence assumption

Given a signal $\{y(x)\}_{x \in \mathcal{X}}$ and a corresponding set of frames $\{y_i\}_{i \in \mathbb{N}}$, a classical assumption consists in writing that the different y_i are independent. As y can be deterministically computed from $\{y_i\}_{i \in \mathbb{N}}$, this is written:

$$\log p(\mathbf{y} \mid X, \Theta) = \sum_{i \in I} \log p(\mathbf{y}_i \mid X'_i, \Theta) \tag{15}$$

If $\mathcal{G}\{y\}$ is modeled as a GP, the frame-wise independence assumption is equivalent to modeling the covariance function $k((i,x),(i',x'))$ of $\mathcal{G}\{y\}$ as:

$$k((i,x),(i',x')) = \delta_{ii'} k_i(x, x') \tag{16}$$

with k_i being the covariance function of the GP $\{y_i(x)\}_{x \in \mathcal{X}'}$. Let X be a finite set of input points from $\mathcal{X}$, $\mathbf{y}$ a process indexed on X and $I \triangleq \bigcup \{I_x\}_{x \in X}$ be the corresponding frame indexes for a framing $\mathcal{G}$. Let n_I be the number of frames. If we model $\mathcal{G}\{\mathbf{y}\}$ as a GP, we readily see that it is equivalent to a multiple output GP as seen in section III-C with n_I outputs whose input set is $\mathcal{X}'$. We can thus stack its outputs and observe that the corresponding covariance matrix is block-diagonal due to the frame-wise independence assumption. Its inverse is thus easily computed.

The main computational trick involved by framing is hence to split the signal into overlapping frames, with a synthesis scheme that allows perfect reconstruction. Then, the frames are supposed to be independent and the corresponding covariance matrix becomes block diagonal. The advantage of this method is that when the frames are overlapping, each point estimate is a smoothing of several estimates computed in the different frames that contain this point, thus avoiding systematic discontinuities. In Figure 2, we illustrate this advantage of framing over PIC to produce smooth estimates in a very simple regression problem.

[4]For a countable set X, $\#(X)$ denotes the number of elements in X.

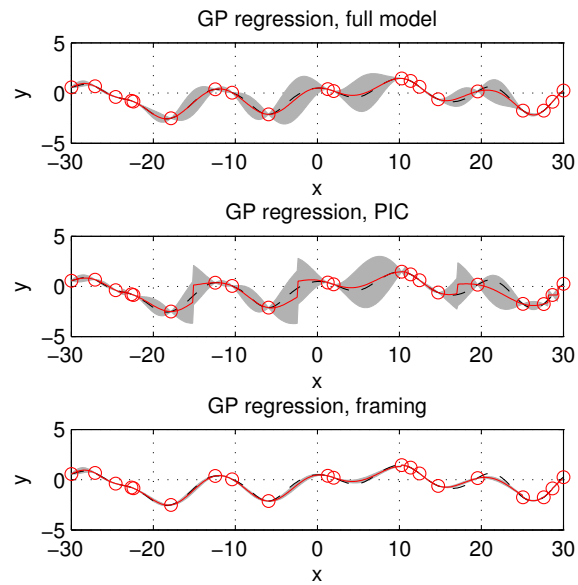

Figure 2. A simple regression example with a full GP model (up), PIC (middle) and framing (bottom) . The dotted and red lines are respectively the true and estimated signal. The grey area represents three standard deviations of the estimates around the mean and the circles stand for observed points.

The connections between *frame-wise independent* correlation functions and other existing models such as PIC or CS covariance functions [31], [40] are numerous. Firstly, we see that framing without overlap between the frames is very similar to PIC except for the fact that it does not take inducing points outside the frames into account. In framing, such remote inducing points are handled through the use of overlapping frames of different scales. Secondly, when considering (13), which gives the expression of the signal given its constitutive frames, we can straightforwardly compute the covariance function of the signal y itself given the covariance functions of the different frames. This computation is actually very similar to that led in [31] where the *basis function* used in the computation becomes the *weighting function* g we considered here. The basis function proposed in [31] is precisely the HANN window which is a very popular choice for g in audio processing.

Of course, as in practice the frames are built using expression (14), there is a duplication of the samples that belong to overlapping frames and the independance assumption between the frames may seem unjustified. Nevertheless, the idea underlying framing is that even if the occurrence of one point from X in some frame gets duplicated in another, and even if the corresponding observed values are connected or equal, they are supposed to be produced by two different underlying processes that do not share the same covariance function.

Still, there are interesting conceptual issues raised by framing that should be more thoroughly studied in the future. In particular, contrary to PIC, framing as it was exposed here suffers from overconfidence. This can be seen by computing the variance of the estimates for y given by (13) and then noticing that the more a point will get duplicated in different frames because of overlap, the smaller the a posteriori variance

of this point will become. This is due to the independence assumption between the frames. If this assumption was strictly legitimate, this diminution of the variance would be justified. However, as the overlap between the frames gets large, independence cannot be a valid assumption anymore and the variances get underestimated. This is illustrated in Figure 2 where the overlap was large.

Even if overlapping the different frames is common practice in audio signal processing, its consequences on statistical models has been largely neglected. In [26], LE ROUX et al. devise practical ways of consistently handling the dependencies between the frames as a postprocessing step. Still, to our knowledge, practical statistical models that fully take the overlap between the frames into account while remaining computationally tractable are yet to be proposed.

D. Stationarity assumption for regularly sampled signals

In this section, we assume that the signals are defined on $\mathcal{X} = \mathbb{R}^D$ for $D \geqslant 1$ and that $x \in \mathcal{X}$ can be written $x = (x_1, \cdots, x_D)$. We will moreover assume that all the covariance functions k that we consider are *separable*, i.e. there are D covariance functions $k^{(d)}$ such that:

$$\forall (x, x') \in \mathcal{X}^2, k(x, x') = \prod_{d=1}^{D} k^{(d)}(x_d, x'_d). \tag{17}$$

It is readily shown that this assumption implies that all the covariance matrices K considered can be expressed as a Kronecker product[5] of D covariance matrices $K^{(d)}$ of lower dimensions:

$$K = K^{(1)} \otimes K^{(2)} \cdots \otimes K^{(D)} \triangleq \bigotimes_{d=1}^{D} K^{(d)}. \tag{18}$$

From now on, we suppose that the points are regularly sampled. This is equivalent to assuming that any signal $\mathbf{y}$, $\mathbf{f}_m$ or $\mathbf{k}$ considered is the vectorization[6] of a corresponding underlying D-dimensional tensor $\underline{y}$, $\underline{f_m}$ or $\underline{k}$. Indeed, we will show in section IV-D1 that computations can be very concisely written using these tensors, which are actually natural to consider. For example, when $D = 2$, it makes sense to directly think of regularly sampled signals as matrices instead of their vectorized counterpart.

1) GPs for the separation of stationary mixtures: As seen in section II-C, a *stationary* covariance function $k(x, x')$ between two input points can be expressed as a function of their difference $\tau = x - x'$. It is noted $k(x - x')$. If all covariance functions considered are stationary, the computations become particularly simple.

Indeed, let us assume that a mixture $\{y(x)\}_{x \in \mathcal{X}}$ is the sum of several GPs $\{f_m(x)\}_{m=1 \ldots M, x \in \mathcal{X}}$ whose covariance functions $k_m(x - x')$ are all stationary, and let us furthermore suppose that we are interested in separating the different sources for all points in X, thus having $X^* = X$. The covariance matrix K_y of y is given by : $K_y = \sum_{m=1}^{M} K_m$ where K_m is the covariance matrix of source m.

Considering (18), K_m is given by: $K_m = \bigotimes_{d=1}^{D} K_m^{(d)}$ where $\left[K_m^{(d)}\right]_{i,j} = k_m^{(d)}(x_{i,d} - x_{j,d})$. $K_m^{(d)}$ can approximately be considered as circulant[7]. It is readily shown that any circulant matrix M can be expressed as $M = W_F^* \Lambda W_F$ where W_F is the discrete Fourier transform matrix[8] and where Λ is diagonal. Thus, for all m and d, there is a diagonal positive semidefinite matrix $\mathrm{diag} S_m^{(d)}$ such that $K_m^{(d)} \approx W_F^* \mathrm{diag} S_m^{(d)} W_F$ where the vector $S_m^{(d)}$ is the discrete Fourier transform of $\tau \mapsto k_m^{(d)}(\tau)$. We can thus write K_y as:

$$K_y = \sum_{m=1}^{M} \bigotimes_{d=1}^{D} W_F^* \mathrm{diag} S_m^{(d)} W_F. \tag{19}$$

Using classical results from tensor algebra, (19) can be written:

$$K_y = \left(\bigotimes_{d=1}^{D} W_F^*\right) \left(\sum_{m=1}^{M} \bigotimes_{d=1}^{D} \mathrm{diag} S_m^{(d)}\right) \left(\bigotimes_{d=1}^{D} W_F\right). \tag{20}$$

We can use this property to extract a given source m_0, and write[9] (4) as:

$$\overline{\mathbf{f}_{m_0}^*} = \left(\bigotimes_{d=1}^{D} W_F^*\right) \left(\frac{\bigotimes_{d=1}^{D} \mathrm{diag} S_{m_0}^{(d)}}{\sum_{m=1}^{M} \bigotimes_{d=1}^{D} \mathrm{diag} S_m^{(d)}}\right) \left(\bigotimes_{d=1}^{D} W_F\right) \mathbf{y}. \tag{21}$$

Introducing the D-dimensional tensor[10]

$$\underline{S}_m = S_m^{(1)} \circ S_m^{(2)} \cdots S_m^{(D)} \triangleq \bigcirc_{d=1}^{D} S_m^{(d)} \tag{22}$$

as the *model for source* m and $\mathcal{F}_D\{\underline{y}\}$ as the D-dimensional Fourier transform of $\underline{y}$, we can simply write (21) in tensor form as:

$$\mathcal{F}_D\left\{\overline{\underline{f}_{m_0}^*}\right\} = \left(\frac{\underline{S}_{m_0}}{\sum_{m=1}^{M} \underline{S}_m}\right) \cdot \mathcal{F}_D\{\underline{y}\} \tag{23}$$

which is similar to the classical Wiener filter for stationary processes. The differences between this expression and the classical one is firstly that it is valid for any dimension D of the input space and secondly that it is not restricted to the case of only two stationary sources. The sources themselves can be recovered through an inverse D-dimensional Fourier transform. The nonnegative tensor $\underline{S}_m$ can be understood as the D-dimensional Fourier transform of the stationary covariance function k_m. Note that the complexity of this *exact* GP inference method relying on stationarity of the covariance functions and on regular sampling is $\mathcal{O}(n \log n)$, and it is dominated by the computation of Fourier transforms, for which there exist very efficient and specialized algorithms. If $\mathcal{F}_D\{\underline{y}\}$ is known beforehand, the complexity of (23) decreases to $\mathcal{O}(n)$ which is remarkable for an exact GP inference technique.

[5] See [9] for a concise introduction to tensor algebra.

[6] Vectorization is done recursively. For example, with $D = 2$ where tensors are matrices, it is done one row after the other.

[7] If the signal is regularly sampled, this approximation holds when the number n_d of points along dimension d tends to infinity or when $k^{(d)}(\tau)$ is periodic of period $\frac{n_d}{p}$ with $p \in \mathbb{N}^*$.

[8] W_F^* denotes the complex conjugate of W_F.

[9] $\frac{A}{B}$ and $A.B$ are respectively the element-wise division and multiplication of A and B.

[10] $\circ$ denotes the outer product.

2) Marginal likelihood for stationary sources: When all the covariance functions considered are stationary and parameterized by some hyperparameter set Θ that consists of their respective D-dimensional Fourier transforms, i.e. $\Theta = \{\underline{S}_1 \cdots \underline{S}_M\}$, it can readily be shown that the marginal log-likelihood $\log p(\mathbf{y} \mid X, \Theta)$ of the observations given regularly spaced input points and the hyperparameters simplifies from (6) to:

$$\log p(\mathbf{y} \mid X, \{\underline{S}_1 \cdots \underline{S}_M\}) = -\frac{1}{2} \sum_{i_1, \cdots i_D} \left[\frac{\left| \left[\mathcal{F}_D \left\{ \underline{y} \right\} \right]_{i_1, \cdots i_D} \right|^2}{\sum_{m=1}^{M} \left[\underline{S}_m \right]_{i_1, \cdots i_D}} + \log \sum_{m=1}^{M} \left[\underline{S}_m \right]_{i_1, \cdots i_D} \right] + \text{Cte} \quad (24)$$

Considering (24), we see that it is equivalent up to an additive constant independent of Θ to half the opposite IS divergence[11] between[12] $\left| \mathcal{F}_D \left\{ \underline{y} \right\} \right|^{.2}$ and $\sum_{m=1}^{M} \underline{S}_m$:

$$\log p(\mathbf{y} \mid X) = -\frac{1}{2} D_{\text{IS}} \left(\left| \mathcal{F}_D \left\{ \underline{y} \right\} \right|^{.2} \mid \sum_{m=1}^{M} \underline{S}_m \right) + \text{Cte} \quad (25)$$

The evaluation of the likelihood can be done in $\mathcal{O}(n \log n)$ operations when the signals are regularly sampled and the covariance functions are stationary. If the squared D-dimensional Fourier transform $\left| \mathcal{F}_D \left\{ \underline{y} \right\} \right|^{.2}$ of the signal is known beforehand — it is typically computed only once — the computational complexity is reduced to $\mathcal{O}(n)$.

E. Locally stationary covariance functions

Let $\{y(x)\}_{x \in \mathcal{X}}$ be a particular signal, observed on a finite input set $X \in \mathcal{X}^n$ and let $\left\{ y_i \in \mathbb{R}^{X_i'} \right\}_{i \in I}$ be a set of n_I corresponding frames. As in section IV-C, we can assume that the frames are independent and we can further suppose that the covariance function k_{im} of source m within frame i is stationary. This means that we model each source as being composed of several locally stationary frames, each of which has its own covariance function. The resulting signal is *not* supposed stationary with this assumption, only its restrictions to small regions of the input space $\mathcal{X}$ are assumed stationary.

Let us denote $\underline{\mathbf{Y}}$ the $(D+1)$-dimensional tensor whose last dimension goes over the frames and whose first D dimensions for a fixed frame contain the D-dimensional Fourier transform of the signal tensor for this frame as in section IV-D. As this tensor is called the Short Term Fourier Transform (STFT) of the signal when $D = 1$, it will be called the STFT tensor of the mixture. We define the STFT tensor $\underline{\mathbf{F}_m}$ of the sources and the STFT tensor $\underline{\mathbf{S}}_m$ of the covariance function of source m in the same way. We can use the results from the previous section for each frame and for source m_0: the MMSE estimate $\underline{\overline{\mathbf{F}}^*_{m_0}}$ of $\underline{\mathbf{F}_{m_0}}$ is given by:

$$\underline{\overline{\mathbf{F}}^*_{m_0}} = \frac{\underline{\mathbf{S}}_{m_0}}{\sum_{m=1}^{M} \underline{\mathbf{S}}_m} \cdot \underline{\mathbf{Y}}. \quad (26)$$

[11] $D_{\text{IS}}\left(\underline{x} | \underline{y}\right) \triangleq \sum_{i_1 \cdots i_D} \left[\frac{[\underline{x}]_{i_1 \cdots i_D}}{[\underline{y}]_{i_1 \cdots i_D}} - \log \frac{[\underline{x}]_{i_1 \cdots i_D}}{[\underline{y}]_{i_1 \cdots i_D}} - 1 \right]$.

[12] For a matrix M, $\left[M^{\cdot 2}\right]_{ij} \triangleq M_{ij}^2$

The sources can then be recovered by first applying an inverse D-dimensional Fourier transform to the estimate (26) for each frame, and then using the reconstruction scheme (13) to obtain the estimated sources in the original input space $\mathcal{X}$.

Let $\Theta = \{\underline{\mathbf{S}}_1, \cdots \underline{\mathbf{S}}_M\}$ be the models for the sources. The marginal likelihood $\log p(\mathbf{y} \mid X, \Theta)$ of the observations can similarly be shown to be:

$$\log p(\mathbf{y} \mid X, \Theta) = -\frac{1}{2} D_{\text{IS}} \left(|\underline{\mathbf{Y}}|^{\cdot 2} \mid \sum_{m=1}^{M} \underline{\mathbf{S}}_m \right) + Cte \quad (27)$$

where the constant is independent of Θ. This very simple expression can be computed in $\mathcal{O}(n)$ when $|\underline{\mathbf{Y}}|^{\cdot 2}$ is known and permits to efficiently proceed to hyperparameters learning as demonstrated in section IV-F.

F. Putting structures over the covariances

Given some regularly sampled signal tensor $\underline{y}$ and its corresponding STFT tensor $\underline{\mathbf{Y}}$ as defined in section IV-E, we have seen that source separation can be very efficiently performed provided some $(D+1)$-dimensional model $\underline{\mathbf{S}_m}$ is known for every source. As highlighted by CEMGIL *et al.* in [7] or [8] for the case of audio processing ($D = 1$), the important issue raised by this probabilistic framework becomes devising realistic but effective models for the nonnegative sources parameters $\underline{\mathbf{S}}_m$.

In audio signal processing (D=1), the result (26) is known as adaptive or generalized Wiener filtering and many methods for source separation such as [8], [32] use this technique in a principled way to recover the sources in the frequency domain. Those studies state their probabilistic model in the frequency domain where the time-frequency bins are supposed to be distributed with respect to independent Gaussian distributions. In our approach, the model is expressed directly in the original input space. The two points of view are actually equivalent: a stationary GP has an independently distributed Gaussian representation in the frequency domain. $\underline{\mathbf{S}}_m$ can hence be seen either as the STFT tensor of a covariance function or as a tensor containing the variances of the independent components of $\underline{\mathbf{Y}}$.

Focusing on the second interpretation of $\underline{\mathbf{S}}_m$, recent studies [8], [12], [13] proposed to model these tensors as Gamma Markov Random Fields (GMRF). This is a sensible choice indeed, because such models guarantee the nonnegativity of all the elements of $\underline{\mathbf{S}}_m$ while implementing the knowledge that for a given source, the spectrum is much likely to exhibit some continuity over time, or over the frequencies, or over both. As GMRF do not provide a closed-form expression for the marginal log-likelihood of the observations, the learning of hyperparameters has to be done using approximate methods. To this end, DIKMEN *et al.* [12] propose to use contrastive divergence [22] and report good results. To our knowledge, no generalization of GMRF has yet been published for input spaces of dimension greater than 2, but GP modeling may greatly benefit from such an extension.

Another point of view is to introduce some deterministic structure into the covariance functions of the GPs. A simple assumption to this end is to consider that for a given source

m, the covariance functions of the different *independent* frames are *stationary* and *locally scaled,* i.e identical up to an amplification gain depending on the frame. The model for source m and frame i can then be written:

$$\underline{\mathbf{S}}_{im} = H_{im}\, \underline{S}_{0,m} \tag{28}$$

where $\underline{S}_{0,m} = \bigcirc_{d=1}^{D} S_{0,m}^{(d)}$ is the D-dimensional Fourier transform of some *template* covariance function $k_{0,m}$ for source m that is independent of the frame index i. We get:

$$\underline{\mathbf{S}}_{m} = \left(\bigcirc_{d=1}^{D} S_{0,m}^{(d)}\right) \circ H_m \tag{29}$$

where $H_m = (H_{1m} \cdots H_{nIm})$ denotes the amplification gains of the covariance function for source m on the different frames. Considering (29) we readily see that it is equivalent to a classical Nonnegative Tensor Factorization (NTF) model called Canonical Polyadic (CP) decomposition[13]. The different parameters become $\Theta = \left\{\left\{H_m, S_{0,m}^{(1)} \cdots S_{0,m}^{(D)}\right\}_{m=1\cdots M}\right\}$ and can be estimated by standard CP algorithms using the IS-divergence function. See [9] for a review of these models and algorithms.

G. Optimization

We have shown how GP learning can be connected to recent NTF techniques by factorizing the covariance structure of the GP model into a CP decomposition. More generally and depending on the application, the covariances can be factorized in many other ways to account for some prior knowledge we may have concerning the structure of the sources. For example, if the covariances are considered to be the outer product of some shared dictionaries, the tensor decompositions to be used become particular cases of Block Components Decompositions as introduced in [25]. Many very informative models can be designed this way, that decompose the covariance structure of the sources onto sophisticated dictionaries. In music processing ($D = 1$) for example, [41] decomposes the covariances into templates of harmonic bases. Other models of this type have also been used to model and extract singing voice signals from polyphonic mixtures with very promising results [14].

In any case, when an appropriate model has been chosen for $\{\underline{\mathbf{S}}_m\}_{m=1\cdots M}$, we have seen in section IV-E that hyperparameters learning can be done by minimizing the IS-divergence between $\sum_m \underline{\mathbf{S}}_m$ and $|\underline{\mathbf{Y}}|^{.2}$ through tensor factorizations. Efficient algorithms for IS-NTF can be found in the litterature, for example in [9].

V. Evaluation

In this section, we demonstrate the performance of the proposed approach based on GP models for the separation of real-valued mixtures. In section V-A, we first show that GP can easily be used for the separation of synthetic 2D random fields, or textures ($D = 2$). Then, we show in section V-B how GP can be used for the separation of drums signals in real polyphonic stereo recordings.

[13]CP is also called PARAFAC or CANDECOMP [9].

A. Synthetic additive textures

In this section, we set $D = 2$, which means that we aim at separating additive functions $f_m(x_1, x_2)$ defined on the plane and summed in an observable mixture signal $y(x_1, x_2)$. For this toy example, we will consider the case of one mixture ($K = 1$) that is the sum of $M = 2$ stationary sources[14]. Following the notations that were introduced in section IV-D, we will thus suppose that the mixture tensor $\underline{y}$ is the sum of two sources tensors $\underline{f}_1$ and $\underline{f}_2$. The corresponding vectors y, f_1 and f_2 will denote the vectorization of these tensors one row after the other. X denotes the corresponding coordinates.

In this experimental setup, the dimensions of the sources and mixtures tensors are 500×500 each, leading to $n = 250000$. In the following, we will assume that the covariance function of each source along each dimension is stationary. For the experiment, the covariance functions were arbitrarily set to:

$$k_m^{(d)}(x_d, x_d') = \exp\left(-\frac{2\sin^2\frac{\pi(x_d - x_d')}{T_{m,d}}}{l_{m,d}^2} - \frac{(x_d - x_d')^2}{2\lambda_{m,d}^2}\right) \tag{30}$$

where $\{T_{m,d}, l_{m,d}, \lambda_{m,d}\}_{m,d}$ are scalar parameters.

This model implements a particular prior knowledge where the sources are known to exhibit some kind of complex structure. More specifically, source m is known to be pseudo-periodic of period $(T_{m,1}, T_{m,2})$ and $(l_{m,1}, l_{m,2})$ controls the smoothness within one period. A further lengthscale $(\lambda_{m,1}, \lambda_{m,2})$ controls global covariance between two input points. In the particular example shown in Figure 3, the parameters were:

m	$\lambda_{m,1}$	$\lambda_{m,2}$	$T_{m,1}$	$T_{m,2}$	$l_{m,1}$	$l_{m,2}$
1	100	100	50	20	0.5	0.7
2	40	4	25	$+\infty$	0.7	N/A

1) Data synthesis: Generating a realization of a GP with some known covariance matrix K is generally addressed through Cholesky or Singular Value Decompositions (SVD) of the covariance matrix [35]. As we have $n = 250000$, we cannot naively implement this idea here. A simple way to circumvent this problem is to write K as in (18) and then to perform a Cholesky decomposition of each $K^{(d)}$ to get $K^{(d)} = L^{(d)} L^{(d)\top}$. The Cholesky decomposition of $K = \bigotimes_{d=1}^{D} L^{(d)} L^{(d)\top}$ is finally obtained by $K = \left(\bigotimes_{d=1}^{D} L^{(d)}\right)\left(\bigotimes_{d=1}^{D} L^{(d)}\right)^{\top}$ and a realization of this GP can be very easily generated. Indeed, let R be a vector of length n whose entries are i.i.d. Gaussian random variables of unit variance. K is the covariance matrix of $\left(\bigotimes_{d=1}^{D} L^{(d)}\right) R$.[15]

2) Source separation: In this very simple experimental setup, we consider that the 12 hyperparameters for the sources covariance functions (30) are known beforehand.

[14]This usecase is common in geostatistics: the observed signal is often modeled as the sum of the signal of interest whith a contaminating white Gaussian noise [11]. Estimating the value of the target signal through Kriging is hence a special case of source separation with GP priors.

[15]We can further speed up this computation by using the fact that for $\mathbf{c} = \mathrm{vec}(\underline{C})$ and matrices A and B of appropriate size, $(A \otimes B)\,\mathbf{c} = A\underline{C}B^{\top}$. This avoids considering such a big matrix as $\left(\bigotimes_{d=1}^{D} L^{(d)}\right)$.

We can perform separation through the exact method presented in section IV-D1. To this purpose, we can build the spectral covariance tensor $\underline{\mathbf{S}}_m$ of each source as the outer product of the Fourier transforms of (30) along each dimension and then perform separation in the frequency domain as in (23). The sources are recovered through an inverse 2-dimensional Fourier transform. It is worth noticing here that the computations are performed extremely rapidly since they only involve element-wise multiplications of 500×500 images, instead of the inversion of the 250000×250000 covariance matrix required by the basic GP setup. Overall computations for this example — synthesis and source separation — are achieved in less than 3 seconds on a standard laptop computer.

Results for one example are shown in Figure 3. The average Signal to Error ratio obtained on 50 experiments was of 8dB, which is very encouraging.

B. Separation of drums signal in polyphonic music

In this section, we apply the general framework we have presented in sections III and IV to the separation of drums signals in polyphonic music. The regularly sampled signals we consider are thus defined on the input space $\mathcal{X} = \mathbb{Z}$ of dimension $D = 1$.

Separation of the drums track from polyphonic music is a challenging task that has already been addressed in several studies such as [18], [21]. Whereas HÉLEN and VIRTANEN [21] perform a Nonnegative Matrix Factorization (NMF) of the mixture and then group the different components obtained through a classification procedure, GILLET and RICHARD [18] decompose the mixture signal with spectral templates learned from a drums database.

In section V-B1, we introduce a GP model for this task and in section V-B2, we compare its performance with the state of the art [18].

1) GP model: The observed mixtures $y(x)$ are supposed to be the sum of two independent GPs $s_d(x)$ and $s_r(x)$ corresponding respectively to the *drums* and the *musical residual* tracks. We assume that some framing $\mathcal{G}\{y\}$ with n_I frames of same length as defined in section IV-B is available for the mixtures and we aim at estimating the framings $\mathcal{G}\{s_d\}$ and $\mathcal{G}\{s_r\}$ of the different sources such that $\mathcal{G}\{y\} = \mathcal{G}\{s_d\} + \mathcal{G}\{s_r\}$.

We suppose that each of the signals s_d and s_r are themselves the sum of several independent processes called *components*. In our example, the R_d different components of s_d are the five most common sources we find in a drums signal, e.g. kick drum, snare drum, hihat, bells and clap sounds. The R_r different components of the musical residual are all the other elements composing the polyphonic mixture. This assumption can be written $s_d = \sum_{m=1}^{R_d} f_m$ and $s_r = \sum_{m=R_d+1}^{R_d+R_r} f_m$.

In the model we are considering, we will assume that all the components f_m are GP whose covariance functions k_m are *locally scaled, frame-wise independent* and *stationary* as defined in section IV-F. For some frame i, they can thus be expressed as:

$$k_{im} = H_{im}\, k_{0,m} \tag{31}$$

where $k_{0,m}$ denotes the template stationary covariance function for component m and $H_m = (H_{1m} \cdots H_{n_I m})$ are the nonnegative activation gains of this component within the frames. Introducing the Fourier transform $S_{0,m}$ of $k_{0,m}$ and using the method presented in section IV-F, the MMSE estimate $\hat{\underline{F}}_d$ of the STFT $\underline{F}_d$ of the drums signal is given by:

$$\hat{\underline{F}}_d = \frac{\sum_{m=1}^{R_d} S_{0,m} \circ H_m}{\sum_{m=1}^{R_d+R_r} S_{0,m} \circ H_m} \cdot \underline{\mathbf{Y}} \tag{32}$$

where $\underline{\mathbf{Y}}$ is the STFT of the mixtures. The model $\underline{\mathbf{S}} \triangleq \sum_m \underline{\mathbf{S}}_m$ becomes $\underline{\mathbf{S}} = \sum_{m=1}^{R_d+R_r} S_{0,m} \circ H_m$. Since $D = 1$, this can be written in matrix form as $\underline{\mathbf{S}} = WH$ where $S_{0,m}$ is the mth column of W and H_m is the mth row of H. As we have shown in section IV-F, the optimization of the hyperparameters $\Theta = \{W, H\}$ through likelihood maximization is thus equivalent to the minimization of the IS distance between the power STFT $|\underline{\mathbf{Y}}|^{.2}$ and the product WH, yielding a NMF model as in [9], [17], [27], [32].

Some other kind of knowledge has now to be put into the model so that it can be useful in practice, since we have not yet made any distinction between the covariance functions of the drums components and those of the musical residual signal. A very simple and computationally cheap solution to this problem is to appropriately initialize some of the hyperparameters. In this experiment, we will focus on the meaning of the activation gains H_m of the components as introduced in (31). H_{im} can be understood as a magnitude parameter for component m into frame i. A good way to initialize all these parameters $\{H_m\}_{m=1\cdots R_d}$ for the drums signal is simply to use an *onset detector* such as [15]. Indeed, if an onset detector feature has a high magnitude in some frame i, then some drums component must be active in it. The onset detector of [15] was hence used in R_d different frequency bands of the mixture STFT, yielding R_d signals. These signals were used to initialize the activation gains $\{H_m\}_{m=1\cdots R_d}$ and all the other hyperparameters of the model were randomly initialized. A NMF was then applied using this initialization and separation was performed using (32).

2) Results: The proposed GP separation method was tested on ten 30-second excerpts sampled at 44.1kHz from the Quaero[16] source separation corpus. The excerpts featured many different kinds of music signals, including pop, electropop, rock, reggae and bossa. For each of these excerpts, the ground truth drums and musical residual signals are known for evaluation but the separation systems can only observe their mixtures. On average, the relative amplitude $20\log_{10}\frac{\sum_x |s_r(x)|}{\sum_x |s_d(x)|}$ of the musical residual signal was set to +6dB compared to the drums signal.

We applied the method proposed by GILLET and RICHARD in [18] on the same mixtures and the quality of the results were quantified through the BSSEVAL toolbox [42]. The separation quality was evaluated both on the drums signals and on the musical residual signal.

The metrics obtained through BSSEVAL include the Source to Distortion Ratio (SDR), the Source to Artifact Ratio (SAR)

[16] http://www.quaero.org

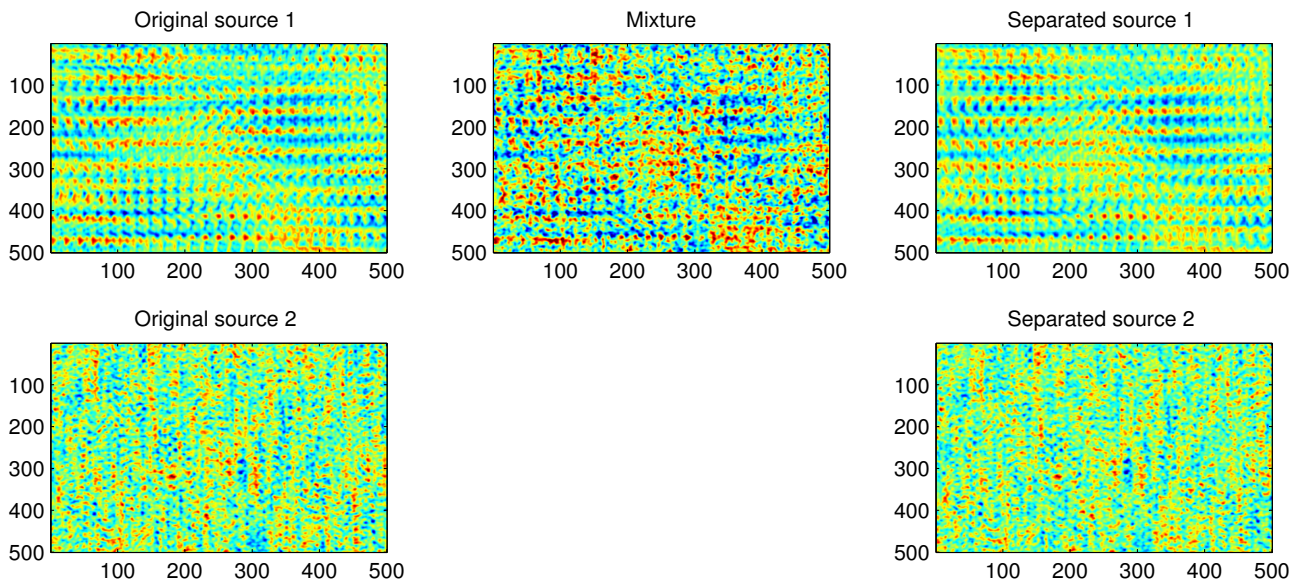

Figure 3. GP for the separation of two stationary random fields ($D = 2$) using a Gaussian Process model. On the left are the original sources. On the center is the mixture and on the right are the estimated sources.

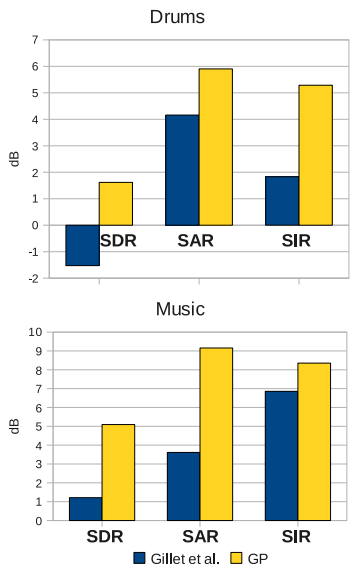

Figure 4. Evaluation of the separation quality for the extraction of the drums track (top) and the musical residual signal (bottom). Higher is better.

and the Source to Interference Ratio (SIR) that are all expressed in dB. Whereas the SDR is a global measure of separation quality, the SAR and SIR respectively measure the amount of separation/reconstruction artifacts and the amount of energy from the other sources. Results are given in Figure 4.

From Figure 4, we can see that the GP model presented in section V-B1 very well manages to separate the drums and musical residual signals on many different kinds of music and both signals are well recovered. A further feature of this technique is that it is extremely fast: on average, 30 seconds are needed to handle a 30-second long excerpt whereas 300 seconds are needed by [18]. Sound excerpts and a full implementation in Python of this separation technique are freely available on our website[17].

VI. Conclusion

In this study, we have stated the linear underdetermined, instantaneous, convolutive and multiple-output source separation problems in terms of Gaussian processes regression and have shown that it leads to simple formulas to optimally proceed to signals separation w.r.t. the MMSE. The advantages of setting out the source separation problem in terms of GP are numerous.

First, there is neither notational burden nor any conceptual issue raised when using input spaces $\mathcal{X}$ different from $\mathbb{R}$ or $\mathbb{Z}$, thus enabling a vast range of source separation problems to be handled within the same framework. Multi-dimensional signal separation may include audio, image or video sensor arrays as well as geostatistics.

Secondly, GP source separation can perfectly be used for the separation of non locally-stationary signals. Of course, some important simplifications of the computations as presented in sections IV-D and IV-E are lost when using non-stationary covariance functions. Still, the frame-wise independence assumption presented in section IV-C may nonetheless be used in order to make the estimations computationally tractable.

Thirdly, it provides a coherent probabilistic way to take many sorts of relevant prior information into account. Indeed, prior information is encapsulated in the choice of the covariance functions and the framework proposed here thus clearly distinguishes between the optimal separation methods and the particular models considered.

[17] http://www.telecom-paristech.fr/~liutkus/GPSS/

Finally, we have seen that under appropriate assumptions, optimization of the hyperparameters of a GP model is equivalent to a classical NTF using the Itakura-Saito divergence on the spectrogram tensor of the mixtures, thus enabling efficient estimation of the hyperparameters.

Setting the source separation problem in such a unified framework allows it to be considered from a larger perspective where its objective is to separate additive independent functions on arbitrary input spaces that are mathematically characterized by their first and second moments only.

REFERENCES

[1] P. Abrahamsen. A review of Gaussian random fields and correlation functions. Technical Report 878, Norsk Regnesentral, Oslo, Norway, April 1997.

[2] M. Alvarez and N. D. Lawrence. Sparse convolved Gaussian processes for multi-output regression. In *Neural Information Processing Systems (NIPS)*, pages 57–64. MIT Press, 2008.

[3] E. V. Bonilla, K. M. A. Chai, and C. K. I. Williams. Multi-task Gaussian process prediction. In *Neural Information Processing Systems (NIPS)*, pages 153–160. MIT Press, 2007.

[4] P. Boyle and M. Frean. Multiple output Gaussian process regression. Technical report, Victoria University of Wellington, April 2005.

[5] P. Boyle and M. R. Frean. Dependent Gaussian processes. In *Neural Information Processing Systems (NIPS)*, pages 217–224. MIT Press, 2004.

[6] J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 90:2009–2026, October 1998.

[7] A. T. Cemgil, S. J. Godsill, P. H. Peeling, and N. Whiteley. *The Oxford Handbook of Applied Bayesian Analysis*, chapter Bayesian Statistical Methods for Audio and Music Processing. Number ISBN13: 978-0-19-954890-3. Oxford University Press, 2010.

[8] A.T. Cemgil, P. Peeling, O. Dikmen, and S. Godsill. Prior structures for Time-Frequency energy distributions. In *Proc. of the 2007 IEEE Workshop on. App. of Signal Proc. to Audio and Acoust. (WASPAA'07)*, pages 151–154, NY, USA, October 2007.

[9] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. Wiley Publishing, September 2009.

[10] P. Comon and C. Jutten, editors. *Handbook of Blind Source Separation: Independent Component Analysis and Blind Deconvolution*. Academic Press, 2010.

[11] P.J. Diggle and P. J. Ribeiro. *Model-based Geostatistics*. Springer series in statistics. Springer, 1 edition, March 2007.

[12] O. Dikmen and A. T. Cemgil. Gamma markov random fields for audio source modelling. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):589–601, March 2010.

[13] O. Dikmen and A.T. Cemgil. Unsupervised single-channel source separation using Bayesian NMF. In *Proc. of the 2009 IEEE Workshop on. App. of Signal Proc. to Audio and Acoust. (WASPAA'09)*, pages 93–96, NY, USA, October 2009.

[14] J.-L. Durrieu, A. Ozerov., C. Févotte, G. Richard, and B. David. Main instrument separation from stereophonic audio signals using a source/filter model. In *Proc. 17th European Signal Proc. Conf. (EUSIPCO'09)*, pages 15–19, Glasgow, UK, August 2009.

[15] C. Duxbury, J.P. Bello, M. Davies, and M. Sandler. Complex domain onset detection for musical signals. In *In Proc. Digital Audio Effects Workshop (DAFx*, London, UK, September 2003.

[16] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3):793–830, March 2009.

[17] D. FitzGerald, M. Cranitch, and E. Coyle. On the use of the beta divergence for musical source separation. In *Proc. of Irish Sig. and Systems Conf. (ISSC'08)*, 2008.

[18] O. Gillet and G. Richard. Transcription and separation of drum signals from polyphonic music. *IEEE Trans. on Audio, Speech, and Language Processing,*, 16(3):529 –540, 2008.

[19] A. Girard, C. E. Rasmussen, J. Quiñonero Candela, and R. Murray-Smith. Gaussian process priors with uncertain inputs - application to multiple-step ahead time series forecasting. In *Neural Information Processing Systems (NIPS)*, pages 529–536. MIT Press, 2002.

[20] P. W. Goldberg, C. K. I. Williams, and C. M. Bishop. Regression with input-dependent noise: A Gaussian process treatment. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Neural Information Processing Systems (NIPS)*, pages 493–499. The MIT Press, 1997.

[21] M. Helén and T. Virtanen. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *Proc. 13th European Signal Processing Conference (EUSIPCO)*, Antalaya, Turkey, 2005.

[22] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, August 2002.

[23] E. T. Jaynes and G. L. Bretthorst. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[24] A.G. Journel and C.J. Huijbregts. *Mining geostatistics*. Academic Press, London ; New York, 1978.

[25] L. De Lathauwer. Decompositions of a higher-order tensor in block terms—part II: Definitions and uniqueness. *SIAM J. Matrix Anal. Appl.*, 30(3):1033–1066, September 2008.

[26] J. Le Roux, E. Vincent, Y. Mizuno, H. Kameoka, N. Ono, and S. Sagayama. Consistent Wiener filtering: Generalized time-frequency masking respecting spectrogram consistency. In *Proc. 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2010)*, pages 89–96, St. Malo, France, September 2010.

[27] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, volume 13, pages 556–562. The MIT Press, April 2001.

[28] D. MacKay. Gaussian processes - a replacement for supervised neural networks? In *Neural Information Processing Systems (NIPS)*. MIT Press, 1997.

[29] G. Matheron. The intrinsic random functions and their applications. *Advances in Applied Probability*, 5(3):439–468, 1973.

[30] A. Melkumyan and F. Ramos. Multi-kernel Gaussian processes. In *Neural Information Processing Systems (NIPS)*. MIT Press, 2009.

[31] A. Melkumyan and F. Ramos. A sparse covariance function for exact gaussian process inference in large datasets. In *IJCAI'09: Proceedings of the 21st international joint conference on Artifical intelligence*, pages 1936–1942, San Francisco, CA, USA, July 2009. Morgan Kaufmann Publishers Inc.

[32] A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Trans. on Audio, Speech and Language Processing*, 18(3):550–563, March 2010.

[33] S. Park and S. Choi. Gaussian processes for source separation. In *Proc. IEEE Intl. Conf. Acoust. Speech Signal Processing (ICASSP'08)*, volume 18, pages 1909–1912, Las Vegas, USA, March 2008.

[34] J. Quiñonero-Candela, C. E. Rasmussen, and R. Herbrich. A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, December 2005.

[35] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[36] M. Seeger. Gaussian processes for machine learning. *Int. J. Neural Syst.*, 14(2):69–106, April 2004.

[37] E. Snelson. Local and global sparse gaussian process approximations. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, volume 2, pages 524–531, San Juan, Puerto Rico, March 2007.

[38] E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Neural Information Processing Systems (NIPS)*, pages 1257–1264. MIT press, 2006.

[39] J. Ichiro Tomita and Y. Hirai. Acoustic echo cancellation using Gaussian processes. In *(ICONIP'08) 15th Int. Conf. on Neural Information Processing.*, volume 5507 of *Lecture Notes in Computer Science*, pages 353–360. Springer, 2008.

[40] J. Vanhatalo and A. Vehtari. Modelling local and global phenomena with sparse gaussian processes. In *Proc. of 24th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 571–578, Helsinky, Finland, July 2008. AUAI Press.

[41] E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE trans. on Audio, Speech and Language Proc. (TASLP)*, 18(3):528–537, March 2010.

[42] E. Vincent, C. Févotte, and R. Gribonval. Performance measurement in blind audio source separation. *IEEE Trans. on Audio, Speech and Language Processing*, 14(4):1462–1469, 2006.

[43] N. Wiener. *Extrapolation, interpolation, and smoothing of stationary time series with engineering applications.* MIT Press, 1949.

[44] C. K. I. Williams. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In M. I. Jordan, editor, *Learning and Inference in Graphical Models*. Kluwer, 1999.

Antoine Liutkus was born in France on February 23rd, 1981. He received the State Engineering degree from Telecom ParisTech, France, in 2005, and the M.Sc. degree in acoustics, computer science and signal processing applied to music (ATIAM) from the Université Pierre et Marie Curie (Paris VI), Paris, in 2005. He worked as a research engineer on source separation at Audionamix from 2007 to 2010 and is currently pursuing the Ph.D. degree in the Department of Signal and Image Processing, Telecom ParisTech.
His research interests include statistical music processing, source separation and machine learning methods applied to signal processing.

Roland Badeau (M'02-SM'10) was born in Marseille, France, on August 28, 1976. He received the State Engineering degree from the École Polytechnique, Palaiseau, France, in 1999, the State Engineering degree from the École Nationale Supérieure des Télécommunications (ENST), Paris, France, in 2001, the M.Sc. degree in applied mathematics from the École Normale Supérieure (ENS), Cachan, France, in 2001, and the Ph.D degree from the ENST in 2005, in the field of signal processing. He received the ParisTech Ph.D. Award in 2006, and the Habilitation à Diriger des Recherches degree from the Université Pierre et Marie Curie (UPMC), Paris VI, in 2010.
In 2001, he joined the Department of Signal and Image Processing, Telecom ParisTech (ENST), as an Assistant Professor, where he became Associate Professor in 2005. From November 2006 to February 2010, he was the manager of the DESAM project, funded by the French National Research Agency (ANR), whose consortium was composed of four academic partners. His research interests include high resolution methods, adaptive subspace algorithms, nonnegative factorizations, audio signal processing, and musical applications. Roland Badeau is a Senior Member of the IEEE Signal Processing Society and he is a Chief Engineer of the French Corps of Mines (foremost of the great technical corps of the French state). He is the author of 17 journal papers and 40 international conference papers.

Gaël Richard (SM'06) received the State Engineering degree from Telecom ParisTech, France (formerly ENST) in 1990, the Ph.D. degree from LIMSI-CNRS, University of Paris-XI, in 1994 in speech synthesis, and the Habilitation à Diriger des Recherches degree from the University of Paris XI in September 2001.
After the Ph.D. degree, he spent two years at the CAIP Center, Rutgers University, Piscataway, NJ, in the Speech Processing Group of Prof. J. Flanagan, where he explored innovative approaches for speech production. From 1997 to 2001, he successively worked for Matra, Bois d'Arcy, France, and for Philips, Montrouge, France. In particular, he was the Project Manager of several large scale European projects in the field of audio and multimodal signal processing. In September 2001, he joined the Department of Signal and Image Processing, Telecom ParisTech, where he is now a Full Professor in audio signal processing and Head of the Audio, Acoustics, and Waves research group. He is a coauthor of over 80 papers and inventor in a number of patents. He is also one of the experts of the European commission in the field of speech and audio signal processing.
Prof. Richard is a member of the EURASIP and an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.