RESEARCH ARTICLE

A data augmentation approach for a class of statistical inference problems

Rodrigo Carvajal[1]*, **Rafael Orellana**[1,2], **Dimitrios Katselis**[3], **Pedro Escárate**[4,5], **Juan Carlos Agüero**[1]

1 Electronics Engineering Department, Universidad Técnica Federico Santa María, Valparaíso, Chile, **2** Universidad de Los Andes, Mérida, Venezuela, **3** Coordinated Science Laboratory and Information Trust Institute, University of Illinois, Urbana-Champaign, Illinois, United States of America, **4** Large Binocular Telescope Observatory, Steward Observatory, University of Arizona, Tucson, AZ, United States of America, **5** Instituto de Electricidad y Electrónica, Facultad de Ciencias de la Ingeniería, Universidad Austral, Valdivia, Chile

* rodrigo.carvajalg@usm.cl

OPEN ACCESS

Citation: Carvajal R, Orellana R, Katselis D, Escárate P, Agüero JC (2018) A data augmentation approach for a class of statistical inference problems. PLoS ONE 13(12): e0208499. https://doi.org/10.1371/journal.pone.0208499

Editor: Yong-Hong Kuo, University of Hong Kong, HONG KONG

Received: May 22, 2018

Accepted: November 19, 2018

Published: December 10, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was partially supported by the Fondo Nacional de Desarrollo Científico y Tecnológico-Chile through grants No. 3140054 and 1181158. This work was also partially supported by the Comisión Nacional de Investigación Científica y Tecnológica, Advanced Center for Electrical and Electronic Engineering (AC3E, Proyecto Basal FB0008), Chile. The work of R. Orellana was partially supported by the PIIC

Abstract

We present an algorithm for a class of statistical inference problems. The main idea is to reformulate the inference problem as an optimization procedure, based on the generation of surrogate (auxiliary) functions. This approach is motivated by the MM algorithm, combined with the systematic and iterative structure of the Expectation-Maximization algorithm. The resulting algorithm can deal with hidden variables in Maximum Likelihood and Maximum a Posteriori estimation problems, Instrumental Variables, Regularized Optimization and Constrained Optimization problems. The advantage of the proposed algorithm is to provide a systematic procedure to build surrogate functions for a class of problems where hidden variables are usually involved. Numerical examples show the benefits of the proposed approach.

1 Introduction

Problems in statistics and system identification often involve variables for which measurements are not available. Among others, real-life examples can be found in communication systems [1, 2] and systems with quantized data [3, 4]. In Maximum Likelihood (ML) estimation problems, the *likelihood function* is in general difficult to optimize by using closed-form expressions, and numerical approximations are usually cumbersome. These difficulties are traditionally avoided by the utilization of the Expectation-Maximization (EM) algorithm [5], where a surrogate (auxiliary) function is optimized instead of the main objective function. This surrogate function includes the complete data, i.e. the measurements and the random variables for which there are no measurements. The incorporation of such *hidden data* or *latent variables* is usually termed as *data augmentation*, where the main goal is to obtain, in general, simple and fast algorithms [6].

On the other hand, the MM (MM stands for Maximization-Minorization or Minimization-Majorization, depending on the optimization problem that needs to be solved) algorithm

program of the Universidad Técnica Federico Santa María, scholarship 015/2018. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

[7] is generally employed for solving more general optimization problems, not only for ML and Maximum a Posteriori (MAP) estimation problems. In general, the main motivation for using the MM algorithm is the lack of closed-form expressions for the solution of the optimization problem or dealing with objective cost functions that are not convex. Applications where the MM algorithm has been utilized include communication systems problems [8] and image processing [9]. For constrained optimization problems, an elegant solution is presented by Marks and Wright [10], where the constraints are incorporated via the formulation of surrogate functions. Surprisingly, Marks' approach has not received the same attention from the scientific community when it comes to compare it with the EM and the MM algorithms. In fact, these three approaches are contemporary, but the EM algorithm has attracted most of the attention (out of the three methods), and it has been used for solving linear and nonlinear statistical inference problems in biology and engineering, see e.g. [11–16], amongst others. On the other hand, as shown in [7], the MM algorithm has obtained much less attention, while Marks' approach is mostly known to a limited audience in the the Communication Systems community. These three approaches have important similarities: i) a surrogate function is defined and optimized in place of the original optimization problem, and ii) the solution is obtained iteratively. In general, these algorithms are "principles and recipes" [17] or a "philosophy" [7] for constructing solutions to a broad variety of optimization problems.

In this paper we adopt the ideas behind [5, 7, 10] to develop an algorithm for a special class of functions. Our approach generalizes the ones of [5, 7, 10] by reinterpreting the E-step in the EM algorithm and expressing the cost function in terms of an infinite mixture or kernel. This kind of problems can be interpreted as *inverse problems* that, in turn, can be posed as integral equations, such as channel modelling in wireless communications [18], estimation of the distribution of stellar rotational velocities [19], and mass estimation in particle physics problems [20, 21]. In particular cases, the kernel corresponds to a variance-mean Gaussian mixture (VMGM), see e.g. [22]. VMGMs, also referred to as normal variance-mean mixtures [23] and normal scale mixtures [24], have been considered in the literature for formulating EM-based approaches to solve ML [25] and MAP problems, including regularized sparse estimation problems [22, 26, 27]. Sparse estimation problems have been widely studied in the last two decades and several techniques have been developed that include different types of penalties or contraint, see e.g [28, 29] and different strategies for the formulation of those penalties/constraints [30]. In this paper we show that our proposal can also be considered for sparsity problems, however the analysis of the solution is out of the scope of the paper. Our approach is applicable to a wide class of functions, which allows for defining the likelihood function, the prior density function, and constraints as kernels, extending also the work in [10]. Thus, our work encompasses the following contributions: i) a systematic approach to constructing surrogate functions for a class of cost functions and constraints, ii) a class of kernels where the unknown quantities of the algorithm can be easily computed, and iii) a generalization of [5, 7, 10] by including the cost function and the constraints in one general expression. Our proposal is based, among other things, on a particular way to apply Jensen's inequality [31]. In addition, we provide the details on how to construct quadratic surrogate functions for cost functions and constraints.

Our algorithm is tested by two examples. In the first one we considered the problem of estimating the rotational velocities of stars. The system model corresponds to the convolution of two probability density functions (pdf's) and thus it is an infinite mixture. We show that our reinterpretation of the EM algorithm allows for the direct application of our proposal for the correct estimation of the parameter of interest. In the second example, we considered the estimation of the channel in wireless communications, where the true distribution can be either Rayleigh or Rice, depending on environment where the electromagnetic waves propagate. The problem is solved considering a sum of a Rayleigh and a Rice term, allowing for a more complex channel distribution. To select the more representative distribution, Akaike's Information Criterion [32] was also considered in order to obtain the least complex model that exhibits the best possible fitting.

2 Rudiments of the proposed approach

2.1 The EM algorithm

Let us consider an estimation problem and its corresponding log-likelihood function defined as $\ell(\boldsymbol{\theta}) = \log p(\boldsymbol{y}|\boldsymbol{\theta})$, where $p(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood function, $\boldsymbol{\theta} \in \mathbb{R}^p$, and $\boldsymbol{y} \in \mathbb{R}^N$. Denoting the *complete data* by $\boldsymbol{z} \in \Omega(\boldsymbol{y})$, and using Bayes' theorem, we can obtain:

$$\ell(\boldsymbol{\theta}) = \log p(\boldsymbol{y}|\boldsymbol{\theta}) = \log p(\boldsymbol{z}|\boldsymbol{\theta}) - \log p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta}). \tag{1}$$

Let us assume that at the ith iteration we have the estimate $\hat{\boldsymbol{\theta}}^{(i)}$. By integrating at both sides of (1) with respect to $p(\boldsymbol{z}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(i)})$ we obtain $\ell(\boldsymbol{\theta}) = \mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}) - \mathcal{H}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$, where

$$\mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}) = \int_{\Omega(\boldsymbol{y})} \log p(\boldsymbol{z}|\boldsymbol{\theta}) p(\boldsymbol{z}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(i)}) d\boldsymbol{z}, \tag{2}$$

$$\mathcal{H}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}) = \int_{\Omega(\boldsymbol{y})} \log p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta}) p(\boldsymbol{z}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(i)}) d\boldsymbol{z}. \tag{3}$$

Using Jensen's inequality [31], it is possible to show that for any value of $\boldsymbol{\theta}$, the function $\mathcal{H}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$ is decreasing. Hence, the optimization is only carried out on the auxiliary function $\mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$ because, by maximizing $\mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$, the new parameter $\hat{\boldsymbol{\theta}}^{(i+1)}$ is such that the likelihood function increases (see e.g. [5, 33]).

In general, the EM method can be summarised as follows:

E-step: Compute the expected value of the joint *likelihood function* for the *complete data* (measurements and hidden variables) based on a given parameter estimate, $\hat{\boldsymbol{\theta}}^{(i)}$. Thus, we have (see e.g. [5]):

$$\mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}) = E[\log p(\boldsymbol{z}|\boldsymbol{\theta})|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(i)}], \tag{4}$$

M-step: Maximize the function $\mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$ (4), with respect to $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}}^{(i+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}). \tag{5}$$

This succession of estimates converges to a stationary point of the *log-likelihood* function [34].

2.2 The MM algorithm

The idea behind the MM algorithm [7] is to construct a surrogate function $g(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$, that majorizes (for minimization problems) or minorizes (for maximization problems) a given cost functions $f(\boldsymbol{\theta})$ [7] at $\hat{\boldsymbol{\theta}}^{(i)}$ such that,

$$\begin{aligned} f(\boldsymbol{\theta}) &\leq g(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}) && \text{for minimization problems, or} \\ f(\boldsymbol{\theta}) &\geq g(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}) && \text{for maximization problems, and} \\ f(\boldsymbol{\theta}) &= g(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^{(i)}), \end{aligned}$$

where $\hat{\boldsymbol{\theta}}^{(i)}$ is an estimate of $\boldsymbol{\theta}$. Then, the surrogate function is iteratively optimized until convergence. Hence, for maximizing $f(\boldsymbol{\theta})$ we have [35]

$$\hat{\boldsymbol{\theta}}^{(i+1)} = \arg\max_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}). \tag{6}$$

For the construction of the surrogate function, popular techniques include the second order Taylor approximation, the quadratic upper bound principle and Jensen's inequality for convex functions, see, e.g., [35].

Remark 1. *The iterative strategy utilized in the MM algorithm converges to a local optimum since*

$$f(\hat{\boldsymbol{\theta}}^{(i+1)}) \geq g(\hat{\boldsymbol{\theta}}^{(i+1)}, \hat{\boldsymbol{\theta}}^{(i)}) \geq g(\hat{\boldsymbol{\theta}}^{(i)}, \hat{\boldsymbol{\theta}}^{(i)}) = f(\hat{\boldsymbol{\theta}}^{(i)}).$$

2.3 Data augmentation in inference problems

Data augmentation algorithms are based on the construction of the *augmented data* and its many-to-one mapping $\Omega(y)$. This *augmented data* is assumed to describe a model from which the observed data y is obtained via marginalization [36]. That is, a system with a likelihood function $p(y|\boldsymbol{\theta})$ can be understood to arise from

$$p(y|\boldsymbol{\theta}) = \int p(y, x|\boldsymbol{\theta}) dx, \tag{7}$$

where the *augmented data* corresponds to (y, x) and x is the *latent data* [6, 36]. This idea has been utilized for supervised learning [37] and the development of the *Bayesian Lasso* [38], to mention a few examples. In those problems, the Laplace distribution is expressed as a two-level hierarchical-Bayes model. This equivalence is obtained from the representation of the Laplace distribution as a VMGM:

$$\frac{a}{2} e^{-a|\theta|} = \int_0^\infty \underbrace{\frac{1}{\sqrt{2\pi\lambda}} e^{-\theta^2/(2\lambda)}}_{p(\theta|\lambda)} \underbrace{\frac{a^2}{2} e^{-a^2\lambda/2}}_{p(\lambda)} d\lambda. \tag{8}$$

In fact, there are several pdf's than can be expressed as VMGMs, as shown in Table 1 [22], where $g(\theta)$ is the penalty term that can be expressed as a pdf. In addition, in [26] it was

Table 1. Selection of mean-variance mixture representations for penalty functions.
$p(\theta) = \int_0^\infty \mathcal{N}_{\boldsymbol{\theta}}(\mu + \lambda u, \tau^2 s^2 \lambda) p(\lambda) d\lambda.$

Penalty function	$g(\boldsymbol{\theta})$	$\boldsymbol{u}$	$\boldsymbol{\mu}$	$\boldsymbol{p(\lambda)}$
Ridge	$(\theta/\tau)^2$	0	0	λ = 1
Lasso	$\lvert\theta/\tau\rvert$	0	0	Exponential
Bridge	$\lvert\theta/\tau\rvert^\alpha$	0	0	Stable
Generalized Double-Pareto	$\left[\frac{(1+\alpha)}{\tau}\right] \log\left(1 + \frac{\lvert\boldsymbol{\theta}\rvert}{(\alpha\tau)}\right)$	0	0	Exp-Gamma

https://doi.org/10.1371/journal.pone.0208499.t001 developed an early version of the methodology presented in this paper, exploring the estimation of a sparse parameter vector utilizing the ℓ_q-(pseudo)norm, with $0 < q < 1$.

3 A systematic approach to construct surrogate functions for a class of inference problems

Here, we consider a general optimization cost defined as:

$$\mathcal{V}(\boldsymbol{\theta}) = \int_{\Omega(\boldsymbol{y})} K(\boldsymbol{z}, \boldsymbol{\theta}) d\mu(\boldsymbol{z}), \tag{9}$$

where $\boldsymbol{\theta}$ is a parameter vector, $\boldsymbol{y}$ is a given data (i.e. measurements), $\boldsymbol{z}$ is the *complete data* (comprised of the *observed data* $\boldsymbol{y}$ and the *hidden variables* (unobserved data), $\Omega(\boldsymbol{y})$ is a mapping from the sample space of $\boldsymbol{z}$ to the sample space of $\boldsymbol{y}$, $K(\cdot, \cdot)$ is a (positive) kernel function, and $\mu(\cdot)$ is a measure, see e.g [31]. The definition in (9) is based on the definition of the auxiliary function $\mathcal{Q}$ in the EM algorithm [5], where it is assumed throughout the paper that there is a mapping that relates the *not observed data* to the *observed data*, and that the *complete data* lies in $\Omega(\boldsymbol{y})$ [5]. Notice that in (9) the kernel function may not be a pdf. However, several functions can be expressed in terms of a pdf. The most common cases are Gaussian kernels (yielding VMGMs) [23] and Laplace kernels (yielding Laplace mixtures) [39].

Remark 2. *Notice that, as explained in Section 1, once the hidden data has been selected, the data augmentation procedure comes with the definition of* $\mathcal{V}(\boldsymbol{\theta})$ *in* (9). *From here, we follow the systematic nature of the EM and MM algorithms in terms of the iterative nature of the technique.*

3.1 Constructing the surrogate function

Since we are considering the optimization of the function $\mathcal{V}(\boldsymbol{\theta})$, we can also consider the optimization of the function

$$\mathcal{J}(\boldsymbol{\theta}) = \log \mathcal{V}(\boldsymbol{\theta}). \tag{10}$$

Without modifying the cost function in (10), we can multiply and divide by the logarithm of the kernel function, obtaining:

$$\mathcal{J}(\boldsymbol{\theta}) = \log \mathcal{V}(\boldsymbol{\theta}) = \log \mathcal{V}(\boldsymbol{\theta}) \frac{\log K(\boldsymbol{z}, \boldsymbol{\theta})}{\log K(\boldsymbol{z}, \boldsymbol{\theta})} = \log K(\boldsymbol{z}, \boldsymbol{\theta}) - \log \frac{K(\boldsymbol{z}, \boldsymbol{\theta})}{\mathcal{V}(\boldsymbol{\theta})}. \tag{11}$$

Let us assume that at the ith iteration we have the estimate $\hat{\boldsymbol{\theta}}^{(i)}$. Then, we can multiply by $\frac{K(\boldsymbol{z}, \hat{\boldsymbol{\theta}}^{(i)})}{\mathcal{V}(\hat{\boldsymbol{\theta}}^{(i)}}$ and integrate on both sides of (11) with respect to $d\mu(\boldsymbol{z})$, obtaining:

$$\begin{aligned} \mathcal{J}(\boldsymbol{\theta}) &= \int_{\Omega(\boldsymbol{y})} \log \mathcal{V}(\boldsymbol{\theta}) \frac{K(\boldsymbol{z}, \hat{\boldsymbol{\theta}}^{(i)})}{\mathcal{V}(\hat{\boldsymbol{\theta}}^{(i)})} d\mu(\boldsymbol{z}) = \log \mathcal{V}(\boldsymbol{\theta}) \\ &= \mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}) - \mathcal{H}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}). \end{aligned} \tag{12}$$

where:

$$\mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}) = \int_{\Omega(\boldsymbol{y})} \log[K(\boldsymbol{z}, \boldsymbol{\theta})] \frac{K(\boldsymbol{z}, \hat{\boldsymbol{\theta}}^{(i)})}{\mathcal{V}(\hat{\boldsymbol{\theta}}^{(i)})} d\mu(\boldsymbol{z}), \tag{13}$$

$$\mathcal{H}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(i)}) = \int\limits_{\Omega(\boldsymbol{y})} \log\left[\frac{K(\boldsymbol{z},\boldsymbol{\theta})}{\mathcal{V}(\boldsymbol{\theta})}\right]\frac{K(\boldsymbol{z},\hat{\boldsymbol{\theta}}^{(i)})}{\mathcal{V}(\hat{\boldsymbol{\theta}}^{(i)})}d\mu(\boldsymbol{z}), \tag{14}$$

are auxiliary functions. As in the EM algorithm, for any $\boldsymbol{\theta}$, and using Jensen's inequality [31], we have:

$$\begin{aligned}\mathcal{H}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(i)}) - \mathcal{H}(\hat{\boldsymbol{\theta}}^{(i)},\hat{\boldsymbol{\theta}}^{(i)}) &= \int\limits_{\Omega(\boldsymbol{y})} \log\left[\frac{K(\boldsymbol{z},\boldsymbol{\theta})}{\mathcal{V}(\boldsymbol{\theta})}\right]\frac{K(\boldsymbol{z},\hat{\boldsymbol{\theta}}^{(i)})}{\mathcal{V}(\hat{\boldsymbol{\theta}}^{(i)})}d\mu(\boldsymbol{z})\\ &\quad - \int\limits_{\Omega(\boldsymbol{y})} \log\left[\frac{K(\boldsymbol{z},\hat{\boldsymbol{\theta}}^{(i)})}{\mathcal{V}(\hat{\boldsymbol{\theta}}^{(i)})}\right]\frac{K(\boldsymbol{z},\hat{\boldsymbol{\theta}}^{(i)})}{\mathcal{V}(\hat{\boldsymbol{\theta}}^{(i)})}d\mu(\boldsymbol{z})\\ &= \int\limits_{\Omega(\boldsymbol{y})} \log\left[\frac{K(\boldsymbol{z},\boldsymbol{\theta})\mathcal{V}(\hat{\boldsymbol{\theta}}^{(i)})}{\mathcal{V}(\boldsymbol{\theta})K(\boldsymbol{z},\hat{\boldsymbol{\theta}}^{(i)})}\right]\frac{K(\boldsymbol{z},\hat{\boldsymbol{\theta}}^{(i)})}{\mathcal{V}(\hat{\boldsymbol{\theta}}^{(i)})}d\mu(\boldsymbol{z})\\ &\leq \log \int\limits_{\Omega(\boldsymbol{y})} \frac{K(\boldsymbol{z},\boldsymbol{\theta})}{\mathcal{V}(\boldsymbol{\theta})}d\mu(\boldsymbol{z})\\ &= 0.\end{aligned} \tag{15}$$

Hence, for any value of $\boldsymbol{\theta}$, the function $\mathcal{H}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(i)})$ in (14) is a decreasing function.

Remark 3. *The kernel function $K(\boldsymbol{z},\boldsymbol{\theta})$ satisfies the standing assumption $K(\boldsymbol{z},\boldsymbol{\theta}) > 0$ since the proposed scheme is built, among others, on the logarithm of the kernel function $K(\boldsymbol{z},\boldsymbol{\theta})$. The definition of the kernel function in (9) allows for kernels that are not pdf's. On the other hand, some kernels may correspond to a scaled version of a pdf. In that sense, for the cost function in (9) we can define a new kernel and a new measure as*

$$\bar{K}(\boldsymbol{z},\boldsymbol{\theta}) = \frac{K(\boldsymbol{z},\boldsymbol{\theta})}{\int K(\boldsymbol{z},\boldsymbol{\theta})d\boldsymbol{\theta}},\ d\bar{\mu}(\boldsymbol{z}) = \left(\int K(\boldsymbol{z},\boldsymbol{\theta})d\boldsymbol{\theta}\right)d\mu(\boldsymbol{z}),$$

$$\Rightarrow \mathcal{V}(\boldsymbol{\theta}) = \int\limits_{\Omega(\boldsymbol{y})} \bar{K}(\boldsymbol{z},\boldsymbol{\theta})d\bar{\mu}(\boldsymbol{z}).$$

Remark 4. *In the proposed methodology, it is possible to optimize the surrogate function defined by*

$$\bar{\mathcal{Q}}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(i)}) = \int\limits_{\Omega(\boldsymbol{y})} \log[K(\boldsymbol{z},\boldsymbol{\theta})]K(\boldsymbol{z},\hat{\boldsymbol{\theta}}^{(i)})d\mu(\boldsymbol{z}), \tag{16}$$

since $\mathcal{V}(\hat{\boldsymbol{\theta}}^{(i)})$ in (13) does not depend on the parameter $\boldsymbol{\theta}$. Thus, the proposed method corresponds to a variation of the EM algorithm that is not limited to probability density functions (e.g. the likelihood function) for solving ML and MAP estimation problems. Instead, our version considers general measures ($\mu(\boldsymbol{z})$), where the mapping over the measurement data $\Omega(\boldsymbol{y})$ is a given set.

The idea behind using a surrogate function is to obtain a simpler algorithm for the optimization of the objective function when compared to the original optimization problem. This can be achieved iteratively if the Fisher Identity for the surrogate function and the objective function is satisfied. That is,

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}} = \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}}. \tag{17}$$

Lemma 1. *For the class of objective functions in* (9), *the surrogate function* $\mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$ *in* (13) *satisfies the Fisher identity defined in* (17).

Proof. From (12) we have:

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}} = \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}} - \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{H}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}}.$$

Next, let us consider the gradient of the auxiliary function $\mathcal{H}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{H}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}} &= \int\limits_{\Omega(\boldsymbol{y})} \left[\frac{K(\boldsymbol{z}, \boldsymbol{\theta})}{\mathcal{V}(\boldsymbol{\theta})}\right]^{-1}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}} \frac{\partial}{\partial \boldsymbol{\theta}} \left[\frac{K(\boldsymbol{z}, \boldsymbol{\theta})}{\mathcal{V}(\boldsymbol{\theta})}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}} \frac{K(\boldsymbol{z}, \hat{\boldsymbol{\theta}}^{(i)})}{\mathcal{V}(\hat{\boldsymbol{\theta}}^{(i)})} d\mu(\boldsymbol{z}) \\
&= \int\limits_{\Omega(\boldsymbol{y})} \frac{\partial}{\partial \boldsymbol{\theta}} \left[\frac{K(\boldsymbol{z}, \boldsymbol{\theta})}{\mathcal{V}(\boldsymbol{\theta})}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}} d\mu(\boldsymbol{z}) = \frac{\partial}{\partial \boldsymbol{\theta}} \left[\int\limits_{\Omega(\boldsymbol{y})} \frac{K(\boldsymbol{z}, \boldsymbol{\theta})}{\mathcal{V}(\boldsymbol{\theta})} d\mu(\boldsymbol{z})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}} \\
&= 0.
\end{aligned}$$

Hence, (17) holds.

Remark 5. *Note that the Fisher identity in Lemma 1 is well known in the EM-framework. However, we have specialized this result for the problem in this paper (i.e. when* $K(\boldsymbol{z}, \boldsymbol{\theta})$ *is not necessarily a probability density function)*

Lemma 2. *The surrogate function* $\mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$ *in* (13) *can be utilized to obtain an adequate surrogate function that satisfies the properties in Mark's approach in* (38)–(40).

Proof. Notice that in Mark's approach the optimization problem corresponds to the minimization of the objective function. Hence, to maximize, we have $\mathcal{J}(\boldsymbol{\theta}) = -\mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}) + \mathcal{H}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$. From (12) we can construct the surrogate functions $\tilde{\mathcal{Q}}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$ and $\tilde{\mathcal{H}}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$ since

$$\mathcal{J}(\boldsymbol{\theta}) = -\underbrace{(\mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}) - \mathcal{Q}(\hat{\boldsymbol{\theta}}^{(i)}, \hat{\boldsymbol{\theta}}^{(i)}) + \mathcal{J}(\hat{\boldsymbol{\theta}}^{(i)})}_{\tilde{\mathcal{Q}}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})} + \underbrace{(\mathcal{H}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}) - \mathcal{Q}(\hat{\boldsymbol{\theta}}^{(i)}, \hat{\boldsymbol{\theta}}^{(i)}) + \mathcal{J}(\hat{\boldsymbol{\theta}}^{(i)})}_{\tilde{\mathcal{H}}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})}. \tag{18}$$

The function $\tilde{\mathcal{H}}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$ satisfies $\tilde{\mathcal{H}}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}) - \tilde{\mathcal{H}}(\hat{\boldsymbol{\theta}}^{(i)}, \hat{\boldsymbol{\theta}}^{(i)}) \geq 0$, which implies that $\mathcal{J}(\boldsymbol{\theta}) \leq \tilde{\mathcal{Q}}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$, satisfying (38). From $\tilde{\mathcal{Q}}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}) = \mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}) - \mathcal{Q}(\hat{\boldsymbol{\theta}}^{(i)}, \hat{\boldsymbol{\theta}}^{(i)}) + \mathcal{J}(\hat{\boldsymbol{\theta}}^{(i)})$ we can obtain $\tilde{\mathcal{Q}}(\hat{\boldsymbol{\theta}}^{(i)}, \hat{\boldsymbol{\theta}}^{(i)}) = \mathcal{J}(\hat{\boldsymbol{\theta}}^{(i)})$, satisfying (39). Finally, given that the auxiliary function $\tilde{\mathcal{H}}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$ satisfies (17), $\tilde{\mathcal{Q}}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$ satisfies (40).

Remark 6. *Since* $\frac{d}{d\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}) = \frac{d}{d\boldsymbol{\theta}} \tilde{\mathcal{Q}}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$, *it is simpler to consider the function* $\mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$ *instead of* $\tilde{\mathcal{Q}}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$ *in penalized (regularized) and MAP estimation problems, as shown in* [26] *and* [27].

We summarize our proposed algorithm in Table 2.

3.2 Surrogate functions for inverse problems

The objective function in (9) can be understood as an integral equation [40], where the unknown function of the integral equation corresponds to the kernel $K(\boldsymbol{z}, \boldsymbol{\theta})$. In this kind of

Table 2. Proposed algorithm.

Step 1: Find a kernel that satisfies (9).
Step 2: $i = 0$.
Step 3: Obtain an initial guess $\hat{\boldsymbol{\theta}}^{(i)}$.
Step 4: Compute $\mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$ as in (13).
Step 5: Compute $\tilde{\mathcal{Q}}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$.
Step 6: Incorporate $\tilde{\mathcal{Q}}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$ in the optimization problem and solve.
Step 7: $i = i + 1$ and back to Step 4 until convergence.

https://doi.org/10.1371/journal.pone.0208499.t002

problems, samples from $\mathcal{V}(\boldsymbol{\theta})$ are available, whilst $d\mu(\boldsymbol{z})$ is assumed known. Several problems can be posed as this kind of problems. In the following, we explain how to use the approach presented in this paper to solve different inverse problems that arise in the integral equation form.

3.2.1 Stellar rotational velocity estimation. One of the many problems in Astronomy deals with is the estimation of rotational velocities of stars. This particular problem is of great importance, since it allows astronomers to describe and model the stars formation, their internal structure and evolution, as well as how they interact with other stars, see e.g. [19, 41, 42].

Modern telescopes allow for the measurement of the rotational velocities from the telescope point of view, that is, a projection of the true rotational velocity. This is modelled (spatially) as the convolution of the true rotational velocity pdf and a uniform distribution over the sphere (for more details see e.g. [19]):

$$p(y|\sigma) = \int_y^\infty \frac{y}{x\sqrt{x^2 - y^2}} p(x|\sigma) dx, \tag{19}$$

where $p(y|\sigma)$ is the uniform projected rotational velocity pdf and $p(x|\sigma)$ is the true rotational velocity pdf to be estimated, and σ a hyperparameter. Thus, we can define the kernel function as $K(x, \sigma) = p(x|\sigma)$ and $d\mu(x) = \frac{y}{x\sqrt{x^2-y^2}} dx$. This definition allows for the direct utilization of the expressions in (13) in order to estimate the parameter σ that defines the unknown rotational velocity pdf. We illustrate with one example in Section 6.1.

3.2.2 Channel estimation in wireless communications. It is well known in the Communications community that the wireless channel corresponds to the superposition of different copies of the transmitted signal that have been reflected, refracted and scattered. Thus, those copies arrive at the receiver with different phase. Those components are referred to as *multipath components* [43]. On the other hand, it has been shown empirically that a good model for the multipath channel corresponds to either Rayleigh or Rice distributions [44]. However, there are cases when those distributions do not provide a good model for the channel. One of those cases corresponds to the presense of different channel models in the vicinity of the transmitter and the vicinity of the receiver. This situation ocurrs particularly in the so called *urban scenarios*, where the channel exhibits different behaviouirs in different places. For this scenario, an adequate model that takes into consideration different models and a transition from a *local distribution* and a *global* distribution is a continuous mixture of the form [18]

$$p(\boldsymbol{x}) = \int_{\Omega(\boldsymbol{x})} K(\boldsymbol{x}, \boldsymbol{\sigma}) d\mu(\boldsymbol{\sigma}), \tag{20}$$

where $K(\boldsymbol{x}, \boldsymbol{\sigma})$ is a pdf in a local area, $\mu(\boldsymbol{\sigma})$ is a distribution of $\boldsymbol{\sigma}$ which depends on the path from transmitter to the local cluster, and $\boldsymbol{\sigma}$ is a vector in the parameter space. Our approach can be directly used in order to estimate the true nature of the channel when expressed as a mixture. This can be done since the four most common chanel distributions are Rayleigh, Rice, log-normal and Nakagami-m [45]. Hence, assuming that the local and global distribution families are known, the attainment of the auxiliary function $\mathcal{Q}(\boldsymbol{\sigma}, \hat{\boldsymbol{\sigma}}^{(i)})$ is straightforward and thus the ML estimate of $\boldsymbol{\sigma}$.

3.2.3 Neutrino mass search in particle physics. In the Particle Physics community there is a plethora of works dealing with the estimation of masses of neutrinos, see e.g. [20] and the references therein. Among the methods that are generally used to determine the absolute masses of neutrinos we find the β-decay and the direct determination of neutrino mass, see e.g [21]. In β-decay methods, the neutrinos are analized based on their energy spectrums, where the measurements, corresponding to the observed β-spectra, are associated with an integral equation of the form [20, 21]

$$F(U) = \int R(E)T'(E, U)dE + b, \tag{21}$$

where b is a constant that represents the measurement noise, $R(E)$ is the emitted β-spectra, and $T'(E, U)$ is the impulse response of the equipment. Again, by regarding $T'(E, U)$ as the kernel function and $R(E)dE$ as a measure, the attainment of the auxiliary function $\mathcal{Q}$ is straightforward.

3.2.4 Estimation of mixture distributions. Mixture distributions have been widely studied in the literature, particularly finite mixtures, see e.g. [46, 47]. Their representation can be expressed in a general fashion by the notation [48]

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \int_{\Omega(\boldsymbol{y})} K(\boldsymbol{z}, \boldsymbol{\theta}) d\boldsymbol{\mu}(\boldsymbol{z}), \tag{22}$$

where $K(\boldsymbol{z}, \boldsymbol{\theta})$ is a suitable function that may be either a pdf (for continuous random variables) or a probability function (for discrete random variables). The expression in (22) represents both the sum

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{\boldsymbol{z}_j \in \Omega(\boldsymbol{y})} K(\boldsymbol{z}_j, \boldsymbol{\theta}), \tag{23}$$

for a finite mixture, or the integral

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \int_{\Omega(\boldsymbol{y})} K(\boldsymbol{z}, \boldsymbol{\theta}) d\boldsymbol{z}, \tag{24}$$

for an infinite mixture.

Our approach can be tailored for the estimation of the parameters of a finite mixture of the form

$$p(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{k=1}^{N} \sum_{j=1}^{M} \lambda_j \phi_j(y_k, \theta_j), \tag{25}$$

where we have assumed that $p(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{k=1}^{N} p(y_k|\boldsymbol{\beta})$, $p(y|\boldsymbol{\beta}) = \sum_{j=1}^{M} \lambda_j \phi_j(y_k, \theta_j)$, and $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_M]$ are the mixing weights, $\phi_j(\boldsymbol{y}, \theta_j)$ are the components densities parametrised by θ_j. The kernel function defined in our approach can be utilized to represent jth component in the discrete mixture in (25) as

$$K_j(\boldsymbol{z}_j, \boldsymbol{\beta}_j) = \lambda_j \phi_j(\boldsymbol{y}, \boldsymbol{\theta}_j), \tag{26}$$

where the $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M]$ is the vector of parameters to be estimated, with $\boldsymbol{\beta}_j = [\lambda_j, \boldsymbol{\theta}_j]$. Notice that the dependence with respect to the variable $\boldsymbol{z}$ is implicit. Utilizing the expression derived in (13) we obtain the following E-step

$$\mathcal{Q}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^{(i)}) = \sum_{j=1}^{M} \log K_j(\boldsymbol{z}_j, \boldsymbol{\beta}_j) \frac{K_j(\boldsymbol{z}_j, \hat{\boldsymbol{\beta}}_j^{(i)})}{\sum_{j=1}^{M} K_j(\boldsymbol{z}_j, \hat{\boldsymbol{\beta}}_j^{(i)})}. \tag{27}$$

Notice that, as shown in [49], the auxiliary function $\mathcal{Q}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^{(i)})$ in (27) is the same that we obtain when we consider that the data are fully categorized, i.e. each y_k, $k = 1, \ldots, N$, is assumed to be drawn from only one distribution of the mixture. This assumption yields a data augmentation problem that is solved using the EM algorithm [46].

Remark 7. *If we consider a combination of infinite mixtures and finite mixtures of the form*

$$p(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{k=1}^{N} \int p(y_k|x_k) p(x_k|\boldsymbol{\beta}) dx_k, \tag{28}$$

with

$$p(x_k|\boldsymbol{\beta}) = \sum_{j=1}^{M} \lambda_j \phi_j(x_k, \boldsymbol{\theta}_j), \tag{29}$$

we can utlize the same approach described here to solve the problem of estimating the parameters in (28), $\boldsymbol{\beta}$. *In this case, the jth kernel is defined as* [50]

$$K_j(x_k, \boldsymbol{\beta}_j) = \lambda_j \phi(x_k; \theta_j), \tag{30}$$

and measure

$$d\mu(x_k) = p(y_k|x_k) dx_k. \tag{31}$$

Then, the log-likelihood function can be expressed as

$$\ell_N(\boldsymbol{\beta}) = \sum_{k=1}^{N} \log [\mathcal{V}_k(\boldsymbol{\beta})], \tag{32}$$

with

$$\mathcal{V}_k(\beta) = \sum_{j=1}^{M} \int_{-\infty}^{\infty} K(x_k, \boldsymbol{\beta}_j) d\mu(x_k). \tag{33}$$

This choice of functions leads to the direct implementation of our proposal, from which it is obtained

$$\mathcal{Q}_k(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^{(i)}) = \sum_{j=1}^{M} \int_{-\infty}^{\infty} \log \left[K\left(x_k, \boldsymbol{\beta}_j\right)\right] \frac{K(x_k, \hat{\boldsymbol{\beta}}_j^{(i)})}{\mathcal{V}_k(\hat{\boldsymbol{\beta}}^{(i)})} d\mu(x_k), \tag{34}$$

and the ML estimator can be locally obtained from

$$\bar{\mathcal{Q}}(\boldsymbol{\beta},\hat{\boldsymbol{\beta}}^{(i)}) = \sum_{k=1}^{N} \mathcal{Q}_k(\beta,\hat{\boldsymbol{\beta}}^{(i)}), \tag{35}$$

$$\hat{\boldsymbol{\beta}}^{(i+1)} = \arg\max_{\boldsymbol{\beta}} \bar{\mathcal{Q}}(\boldsymbol{\beta},\hat{\boldsymbol{\beta}}^{(i)}). \tag{36}$$

4 Marks' approach for constrained optimization

4.1 Constrained problems in statistical inference

Statistical Inference and System Identification techniques include a variety of methods that can be used in order to obtain a model of a system from data. Classical methods, such as *Least Squares*, ML, MAP [51], *Prediction Error Method*, *Instrumental Variables* [52], and *Stochastic Embedding* [53] have been considered in the literature for such task. However, the increasing complexity of modern system models has motivated researchers to revisit and reconsider those techniques for some problems. This has resulted in the incorporation of constraints and penalties, yielding an often more complicated optimization problem. For instance, it has been shown that the incorporation of linear equality constraints may improve the accuracy of the parameter estimates, see e.g [54]. On the other hand, the incorporation of regularization terms (or penalties) also improves the accuracy of the estimates, reducing the effect of noise and eliminating spurious local minima [55]. Regularization can be mainly incorporated in two ways: by adding regularizing constraints (a penalty function) or by including a probability density function (pdf) as a prior distribution for the parameters, see e.g. [27]. Another way to improve the estimation is by incorporating inequality constraints, where certain functions of the parameters may be required, for physical reasons amongst others, to lie between certain bounds [56]. From this point of view, it is possible to consider the classical methods with constraints or penalties, as in [53, 55–57].

Perhaps one of the most utilized approaches for penalized estimation (with complicated non-linear expressions) is the MM algorithm—for details on the MM algorithm see Section 2.2. This technique allows for the utilization of a surrogate function that is simple to handle, in terms of derivatives and optimization techniques, and that is, in turn, iteratively solved. However, its inequality constraint counterpart, here referred to as Marks' approach [10], is not as well known as the MM algorithm. Moreover, there is no straightforward manner to obtain such surrogate function. In this paper we focus on a systematic way to obtain the corresponding surrogate function using Marks' approach for a class of constraints.

4.2 Mark's approach

The approach in [10] deals with inequality constraints by using a similar approach to the EM and MM algorithms. The basic idea is, again, to generate a surrogate function that allows for an iterative procedure whose optimum value is the optimum value of the original optimization problem.

Let us consider the following constrained optimization problem:

$$\begin{aligned} \boldsymbol{\theta}^{*} = {} & \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \\ \text{s. t.} \quad & g(\boldsymbol{\theta}) \leq 0, \end{aligned} \tag{37}$$

where $f(\boldsymbol{\theta})$ is the objective function and $g(\boldsymbol{\theta})$ encodes the constraint of the optimization problem. In particular, let us focus on the case where $g(\boldsymbol{\theta})$ is not a convex function. This implies that the optimization problem cannot be solved directly using standard techniques, such as quadratic programming or fractional programming. This difficulty can be overcome by utilizing a surrogate function $\mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$ at a given estimate $\hat{\boldsymbol{\theta}}^{(i)}$, such that

$$g(\boldsymbol{\theta}) \leq \mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}) \tag{38}$$

$$g(\hat{\boldsymbol{\theta}}^{(i)}) = \mathcal{Q}(\hat{\boldsymbol{\theta}}^{(i)}, \hat{\boldsymbol{\theta}}^{(i)}) \tag{39}$$

$$\frac{d}{d\boldsymbol{\theta}} g(\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}} = \frac{d}{d\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}} \tag{40}$$

Provided the above properties are satisfied, then the following approximation of (37):

$$\begin{aligned} \boldsymbol{\theta}^{(i+1)} &= \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \\ \text{s. t.} \quad & \mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}) \leq 0, \end{aligned} \tag{41}$$

iteratively converges to the solution of the optimization problem (37). As shown in [10], the optimization problem in (41) is equivalent to the original problem (37), since the solution of (41) converges to a point that satisfies the Karush-Kuhn-Tucker conditions of the original optimization problem.

Remark 8. *Mark's approach can be considered as a generalization of the MM algorithm, since the latter can be derived (for a broad class of problems) from the former. Let us consider the following problem*:

$$\boldsymbol{\theta}^* = arg \min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}). \tag{42}$$

Using the epigraph representation of (42) [58], *we obtain the equivalent problem*

$$\begin{aligned} \boldsymbol{\theta}^* &= arg \min_{\boldsymbol{\theta}} t \\ s.\,t. \quad & f(\boldsymbol{\theta}) \leq t, \end{aligned} \tag{43}$$

Using Mark's approach (41), *we can iteratively find a local optimum of* (42) *via*

$$\begin{aligned} \boldsymbol{\theta}^{(i+1)} &= arg \min_{\boldsymbol{\theta}} t \\ s.\,t. \quad & \mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}) \leq t, \end{aligned} \tag{44}$$

where $\mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$ *in* (44) *is a surrogate function for* $f(\boldsymbol{\theta})$ *in* (42). *From the epigraph representation we then obtain*

$$\boldsymbol{\theta}^{(i+1)} = arg \min_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}), \tag{45}$$

which is the definition of the MM algorithm (see 2.2) for more details.

5 A quadratic surrogate function for a class of kernels

In this section we focus on a special class of the kernel functions $K(\boldsymbol{z}, \boldsymbol{\theta})$. For this particular class, the following is satisfied:

$$\frac{\partial}{\partial\boldsymbol{\theta}} \log\left[K(\boldsymbol{z}, \boldsymbol{\theta})\right] = \mathbf{A}(\boldsymbol{z})\boldsymbol{\theta} + \mathbf{b}, \tag{46}$$

where $\mathbf{A}(\boldsymbol{z})$ is a matrix and $\mathbf{b}$ is a vector, both of adequate dimensions. Then, we have that

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}}\mathcal{Q}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(i)}) &= \int\limits_{\Omega(\boldsymbol{y})}[\mathbf{A}(\boldsymbol{z})\boldsymbol{\theta}+\mathbf{b}]\frac{K(\boldsymbol{z},\hat{\boldsymbol{\theta}}^{(i)})}{\mathcal{V}(\hat{\boldsymbol{\theta}}^{(i)})}d\mu(\boldsymbol{z})\\ &= \left[\int\limits_{\Omega(\boldsymbol{y})}\mathbf{A}(\boldsymbol{z})\frac{K(\boldsymbol{z},\hat{\boldsymbol{\theta}}^{(i)})}{\mathcal{V}(\hat{\boldsymbol{\theta}}^{(i)})}d\mu(\boldsymbol{z})\right]\boldsymbol{\theta}+\mathbf{b}\int\limits_{\Omega(\boldsymbol{y})}\frac{K(\boldsymbol{z},\hat{\boldsymbol{\theta}}^{(i)})}{\mathcal{V}(\hat{\boldsymbol{\theta}}^{(i)})}d\mu(\boldsymbol{z})\\ &= \boldsymbol{R}\boldsymbol{\theta}+\boldsymbol{b}.\end{aligned} \tag{47}$$

Remark 9. *Notice that the previous expression is linear with respect to* $\boldsymbol{\theta}$. *This implies that the function* $\mathcal{Q}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(i)})$ *is quadratic with respect to the parameter vector* $\boldsymbol{\theta}$.

From the Fisher Identity in (17) we have that

$$\frac{\partial}{\partial \boldsymbol{\theta}}\mathcal{J}(\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}} = \mathbf{R}\hat{\boldsymbol{\theta}}^{(i)}+\mathbf{b}, \tag{48}$$

from which we can solve for $\mathbf{R}$ in some cases. In other cases, the matrix $\mathbf{R}$ can also be computed using Monte Carlo algorithms. In particular, if $\mathbf{A}(\boldsymbol{z})$ is a diagonal matrix, then $\mathbf{R}$ is also a diagonal matrix defined by $\mathbf{R} = \text{diag}[r_1, r_2,\ldots]$. Thus, we have

$$\frac{\partial}{\partial \theta_k}\mathcal{J}(\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}} = r_k\hat{\theta}_i^{(i)}+b_k \Rightarrow r_k = \frac{\frac{\partial}{\partial \theta_k}\mathcal{J}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}}-b_k}{\hat{\theta}_k^{(i)}},$$

where θ_i is the ith component of the parameter vector $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_i^{(i)}$ is the ith component of the estimate $\hat{\boldsymbol{\theta}}^{(i)}$, r_i is the ith element of the diagonal of $\mathbf{R}$, and b_i is the ith element of the vector $\mathbf{b}$. Hence, when optimizing the auxiliary function $\mathcal{Q}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(i)})$ we obtain

$$\frac{\partial}{\partial \boldsymbol{\theta}}\mathcal{Q}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(i)}) = \begin{bmatrix} r_1 & & \\ & r_2 & \\ & & \ddots \end{bmatrix}\boldsymbol{\theta}+\mathbf{b}=0 \Rightarrow \hat{\theta}_k^{(i+1)} = -\frac{b_k}{r_k}.$$

Equivalently,

$$\hat{\boldsymbol{\theta}}^{(i+1)} = \mathbf{R}^{-1}\mathbf{b}. \tag{49}$$

This implies that in our approach, it is not necessary to obtain the auxiliary function $\mathcal{Q}$ and optimize it. Instead, by computing $\mathbf{R}$ and $\mathbf{b}$ at every iteration, the new estimate can be obtained.

Remark 10. *The computation of the matrix* $\boldsymbol{R}$ *can be cumbersome when the matrix* $\boldsymbol{A}(\boldsymbol{z})$ *is not diagonal. In those cases, the integral that defines* $\boldsymbol{R}$ *can be computed utilizing Markov Chain Monte Carlo, quasi-Monte Carlo* [59] *or quadrature methods* [60].

The class defined in (46) arises naturally when dealing with VMGM, because the corresponding kernel function is normal and, thus, its logarithm is a quadratic function.

In particular, the utilization of VMGM encompasses different expressions commonly used for parameter estimation. Indeed, we have:

(i) Lasso: The Lasso [61], expressed as a Laplace pdf, is represented by a VMGM [38], with $p(\boldsymbol{\theta}|\boldsymbol{z})$ a zero-mean Gaussian distribution (of iid terms) as the kernel and $p(\boldsymbol{z})$ an exponential distribution with parameter $\gamma^2/2$. That is,

$$p(\boldsymbol{\theta}) = \prod_{j=1}^{\mathrm{p}} \frac{\gamma}{2} e^{-\gamma|\theta_j|} = \prod_{j=1}^{\mathrm{p}} \int \mathcal{N}_{\theta_j}(0, z_j) \left(\frac{\gamma^2}{2} e^{-\frac{\gamma^2}{2} \boldsymbol{z}_j} \right) dz_j. \tag{50}$$

(ii) Elastic-Net: The Elastic-Net penalty [62] is interpreted as a pdf if it corresponds to the product of two pdf's, a Laplacian (as in the Lasso case) and a Gaussian pdf. In this sense, by grouping those pdf's, we obtain [63]

$$p(\boldsymbol{\theta}) = k_{EN} \prod_{j=1}^{\mathrm{p}} \int_1^{\infty} \mathcal{N}_{\theta_j}\left(0, \frac{(\lambda_j - 1)}{\lambda_j \kappa}\right) p(\lambda_j) d\lambda_j, \tag{51}$$

with $p(\lambda_j) \propto \sqrt{\frac{1}{\lambda_j - 1}} e^{-\frac{1}{8}\frac{(1-\kappa)^2}{\kappa}\lambda_j}$.

(iii) Group-Lasso: The Group-Lasso penalty is obtained via VMGM-representation as

$$p(\boldsymbol{\theta}) = k_{GL} \prod_{j=1}^{\mathrm{p}} \int_0^{\infty} \mathcal{N}_{\boldsymbol{\theta}_{\mathrm{g}}}\left(\mathbf{0}, \frac{\lambda_{\mathrm{g}}}{\gamma} \mathbf{I}_{\mathrm{G_g}}\right) \chi^2_{\mathrm{G_g}+1}(\lambda_{\mathrm{g}}) d\lambda_{\mathrm{g}}, \tag{52}$$

where χ^2_l is the chi-squared distribution with l degrees of freedom.

For the class of kernels here described, the proposed method for constructing surrogate functions can also be understood as part of sequential quadratic programming (SQP) methods [64] when, for instance, the above penalties are utilized as constraints in a constrained ML estimation problem. Indeed, the general case of equality and inequality-constrained minimization problems is defined as [65]:

$$\begin{aligned} \boldsymbol{\theta}^* &= \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \\ \text{s.t.} \quad & h(\boldsymbol{\theta}) = 0, g(\boldsymbol{\theta}) \leq 0, \end{aligned} \tag{53}$$

which is solved by iteratively defining quadratic functions that approximate the objective function and the inequality constraint around a current iterate $\hat{\boldsymbol{\theta}}^{(i)}$. In the same way, our proposal generates an algorithm with quadratic surrogate functions, where an auxiliary function $\tilde{\mathcal{Q}}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$ in (18) must be constructed for $f(\boldsymbol{\theta})$ and/or $g(\boldsymbol{\theta})$ in (53).

6 Numerical examples

In this section, we illustrate our proposed algorithm with two numerical examples.

6.1 Example 1: Estimation of the distribution of stellar rotational velocities

A common model for $p(x|\sigma)$ found in the Astronomy literature is the Maxwellian distribution (see e.g. [19, 66])

$$p(x|\sigma) = \sqrt{\frac{2}{\pi}\frac{1}{\sigma^3}} x^2 e^{-x^2/(2\sigma^2)}. \tag{54}$$

In practice, the measurements correspond to realizations of $p(y|\sigma)$ [19], from which the likelihood function can be defined as:

$$p(\mathbf{y}|\sigma) = \prod_{k=1}^{N} p(y_k|\sigma), \tag{55}$$

where $\mathbf{y} = [y_1,\ldots, y_N]^T$,

$$p(y_k|\sigma) = \int_{y_k}^{\infty} \frac{y_k}{x_k\sqrt{x_k^2 - y_k^2}} p(x_k|\sigma) dx_k,$$

x_k is Maxwellian distributed, and N is the number of measurement points. Hence, the log-likelihood function can be expressed as:

$$\ell(\sigma) = \sum_{k=1}^{N} \log\left[\int_{y_k}^{\infty} \frac{y_k}{x_k\sqrt{x_k^2 - y_k^2}} p(x_k|\sigma) dx_k\right]. \tag{56}$$

If we define the complete data $\boldsymbol{z} = (\mathbf{x}, \mathbf{y})$, the kernel function $K(\cdot, \cdot)$ and the measure $\mu(\cdot)$ in (9) can be defined as

$$K(x_k, \sigma) = p(x_k|\sigma) = \sqrt{\frac{2}{\pi}} \frac{x_k^2}{\sigma^3} e^{-x_k^2/(2\sigma^2)}, \tag{57}$$

and

$$d\mu(x_k, y_k) = \frac{y_k}{x_k\sqrt{x_k^2 - y_k^2}} dx_k. \tag{58}$$

Then, the log-likelihood function in (56) can be written as:

$$\ell(\sigma) = \sum_{k=1}^{N} \log\,[\mathcal{V}_k(\sigma)], \tag{59}$$

with

$$\mathcal{V}_k(\sigma) = \int_{y_k}^{\infty} K(x_k, \sigma) d\mu(x_k, y_k), \tag{60}$$

Thus, the ML estimator is obtained from:

$$\hat{\sigma}_{\mathrm{ML}} = \arg\max_{\sigma} \sum_{k=1}^{N} \log \mathcal{V}_k(\sigma). \tag{61}$$

Since the parameter that is needed to be estimated is part of the convolution in (19), the optimization problem in (61) cannot be solved in a straightforward manner. Instead, we utilize the re-interpretation of the EM algorithm that we propose for solving (61).

First, notice that from the surrogate function $\mathcal{Q}(\sigma, \hat{\sigma}^{(i)})$ can be expressed as:

$$\mathcal{Q}(\sigma, \hat{\sigma}^{(i)}) = \sum_{k=1}^{N} \mathcal{Q}_k(\sigma, \hat{\sigma}^{(i)}), \tag{62}$$

with

$$\mathcal{Q}_k(\sigma, \hat{\sigma}^{(i)}) = \int_{y_k}^{\infty} \log\left(K(x_k, \sigma)\right) \frac{K(x_k, \hat{\sigma}^{(i)})}{\mathcal{V}_k(\hat{\sigma}^{(i)})} d\mu(x_k, y_k). \tag{63}$$

For convenience, we can differentiate the auxiliary function $\mathcal{Q}(\sigma, \hat{\sigma}^{(i)})$ in (62) with respect to $1/\sigma$ obtaining:

$$\frac{\partial \mathcal{Q}(\sigma, \hat{\sigma}^{(i)})}{\partial (1/\sigma)} = \sum_{k=1}^{N} \int_{y_k}^{\infty} \left[3\sigma - \frac{x_k^2}{\sigma}\right] \frac{K(x_k, \hat{\sigma}^{(i)})}{\mathcal{V}_k(\hat{\sigma}^{(i)})} d\mu(x_k, y_k). \tag{64}$$

Then, equating to zero and solving for σ we finally obtain

$$\hat{\sigma}^{(i+1)} = \sqrt{\frac{\mathcal{S}(\mathbf{y}, \hat{\sigma}^{(i)})}{3N}}, \tag{65}$$

where

$$\mathcal{S}(\mathbf{y}, \hat{\sigma}^{(i)}) = \sum_{t=1}^{N} \int_{y_k}^{\infty} x_k^2 \frac{K(x_k, \hat{\sigma}^{(i)})}{\mathcal{V}_k(\hat{\sigma}^{(i)})} d\mu(x_k, y_k). \tag{66}$$

In Table 3 we summarized the specialisation of our proposed algorithm for this example.

For the numerical simulation, we have considered the problem solved in [19], with the true dispersion parameter $\sigma_0 = 8$. The measurement data $\mathbf{y} = [y_1, \ldots, y_N]$ was generated using the *Slice Sampler* (see e.g. [67]) applied to (19). The simulation setup is as follows:

- The data length is given by $N = 10000$.
- The number of Monte Carlo (MC) simulations is 50.
- The stopping criterion is given by:

$$\|\hat{\sigma}^{(i)} - \hat{\sigma}^{(i-1)}\| / \|\hat{\sigma}^{(i)}\| < 10^{-6},$$

or the maximum number of iterations of 100 has been reached.

The results are shown in Fig 1, were the estimated $p(x|\sigma)$ for each MC simulation is shown. It is clear that the estimated Maxwellian distributions are very similar to the *true* density distribution. The mean value of the estimated parameter was $\hat{\sigma} = 7.9920$. The estimation from each MC simulation is shown in Fig 2. It can be clearly seen that the estimated parameter $\hat{\sigma}$ is close to the *true* value.

6.2 Example 2: Channel estimation in wireless communications

When modelling the wireless channel, a popular technique that is commonly used corresponds to the transmission of a *sine tone* at a given frequency, see e.g. [68]. The received power is then modelled as a random variable. The corresponding distribution has been widely studied in the literature from measurements, and the empirical data have shown that the two most common distributions are Rayleigh an Rice [44, 45]. Hence, using similar ideas as in [50], in this example we formulate the channel distribution as a discrete sum based on a Rayleigh and a Rice component to determine the nature of the wireless channel.

Table 3. Proposed algorithm for Maxwellian distribution estimation in Example 1.

Step 1: $i = 0$.
Step 2: Obtain an initial guess $\hat{\boldsymbol{\sigma}}^{(i)}$.
Step 3: Compute the integral given by (66).
Step 4: Compute $\hat{\sigma}^{(i+1)}$ (65)
Step 5: $i = i + 1$ and back to Step 3 until convergence.

https://doi.org/10.1371/journal.pone.0208499.t003

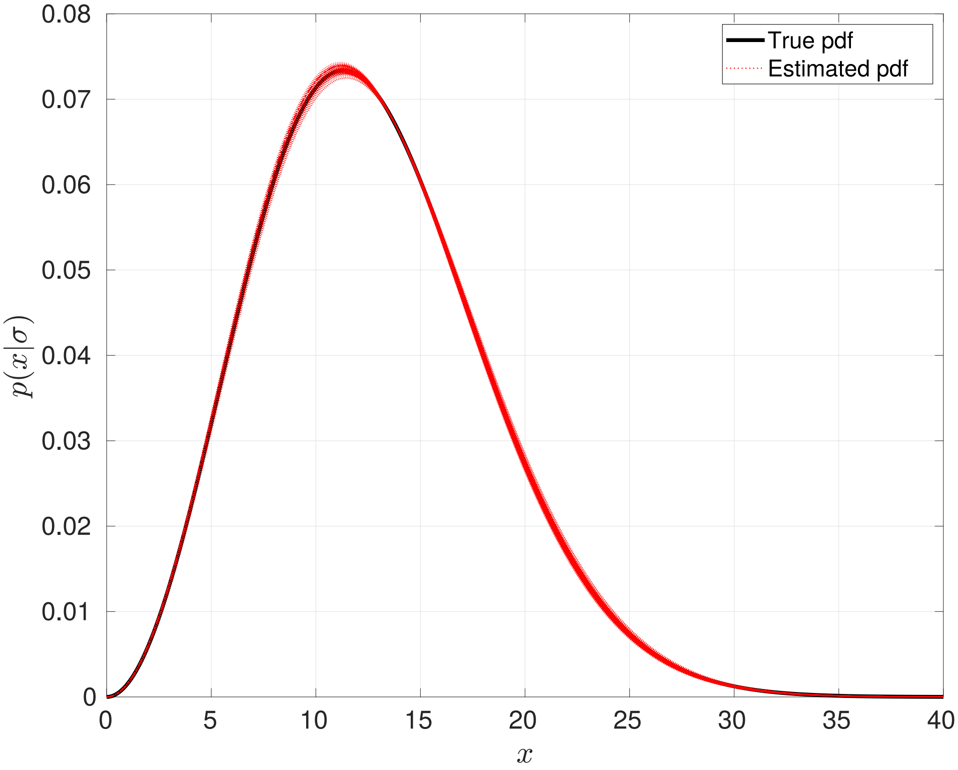

Fig 1. Estimated distribution for the stellar rotational velocity.

https://doi.org/10.1371/journal.pone.0208499.g001

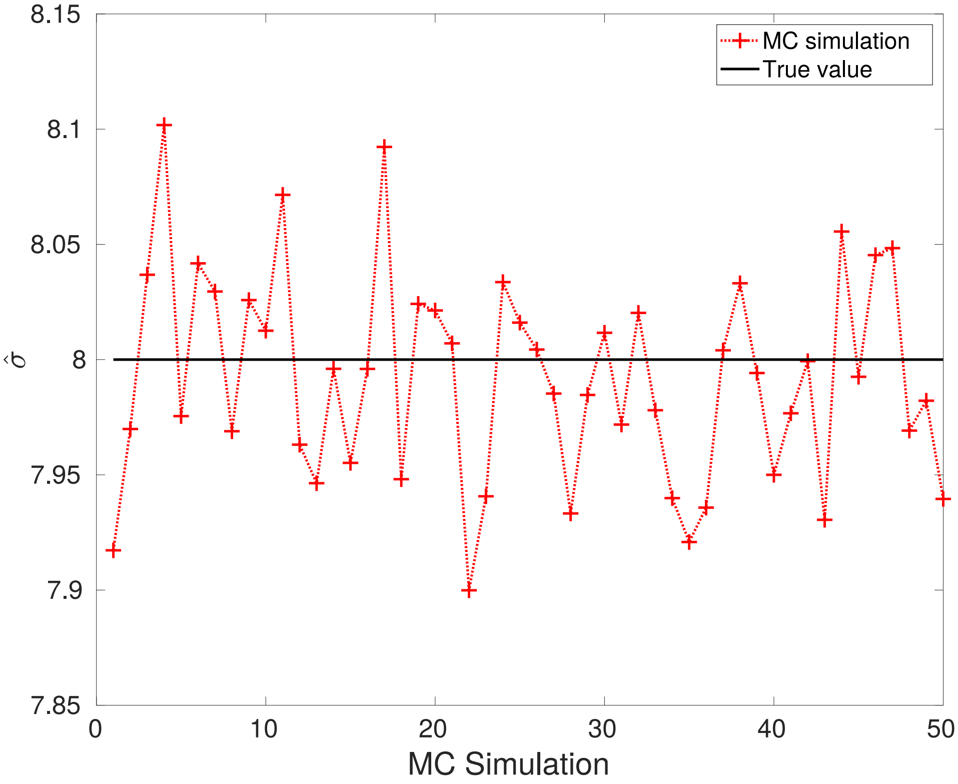

Fig 2. Convergence of the proposed approach to the global optimum.

https://doi.org/10.1371/journal.pone.0208499.g002

First, the discrete mixture that we want to adjust from data is given by

$$p(x|\boldsymbol{\theta}) = \lambda_1 p_{\text{Rayleigh}}(x|\sigma_1^2) + \lambda_2 p_{\text{Rice}}(x|\nu, \sigma_2^2), \tag{67}$$

where

$$p_{\text{Rayleigh}}(x|\sigma_1^2) \quad = \frac{x}{\sigma_1^2} e^{-\frac{x^2}{2\sigma_1^2}}, \tag{68}$$

$$p_{\text{Rice}}(x|\nu, \sigma_2^2) \quad = \frac{x}{\sigma^2} e^{-\frac{(x^2+\nu^2)}{2\sigma_2^2}} I_0\left(\frac{x\nu}{\sigma_2^2}\right), \tag{69}$$

$$\boldsymbol{\theta} = [\sigma_1^2, \lambda_1, \nu, \sigma_2^2, \lambda_2], \tag{70}$$

and $I_0(\cdot)$ is the modified Bessel function of zeroth order. In addition, we must also include the constraint $\lambda_1 + \lambda_2 = 1$ so $p(x)$ is a pdf. Thus, we can directly apply our proposed approach by assuming that each measurement point can be associated with a hidden variable that describes if such data point comes from the Rayleigh component or the Rice component, as it is traditionally formulated when dealing with discrete mixtures [46]. Hence, the auxiliary function $\mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)})$ is given by

$$\mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}) = \sum_{t=1}^{2}\sum_{j=1}^{2} \zeta_{tj}^{(i)} \log \lambda_j + \sum_{t=1}^{N}\sum_{j=1}^{2} \zeta_{tj}^{(i)} \log f_j(x_t, \boldsymbol{\theta}_j), \tag{71}$$

where $\hat{\boldsymbol{\theta}}^{(i)}$ is the current estimate, ζ is the unobserved (*hidden*) data and ζ_{tj} is an indicator parameter such that $\zeta_{tj} = 1$ if the t-th observation comes from component j and is zero otherwise. It is given by:

$$\zeta_{tj}^{(i)} = \frac{\lambda_j^{(i)} f_j(x_t, \theta_j^{(i)})}{\sum_{l=1}^{2} \lambda_l^{(i)} f_l(x_t, \boldsymbol{\theta}_l^{(i)})}. \tag{72}$$

The estimate $\hat{\boldsymbol{\theta}}^{(i+1)}$ at the next iteration is then given by:

$$\hat{\boldsymbol{\theta}}^{(i+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(i)}), \tag{73}$$

from which we obtain the following expressions:

$$\hat{\nu}_j^{(i+1)} = \frac{\sum_{t=1}^{N} \zeta_{tj}^{(i)} x_t \frac{I_1(\rho_{tj}^{(i)})}{I_0(\rho_{tj}^{(i)})}}{\mathcal{P}_j(\hat{\beta}_j^{(i)})} \tag{74}$$

$$[\hat{\sigma}_j^2]^{(i+1)} = \frac{\sum_{t=1}^{N} \zeta_{tj}^{(i)} \left[x_t^2 + \left[\hat{\nu}_j^2\right]^{(i)} - 2x_t \hat{\nu}_j^{(i)} \frac{I_1(\rho_{tj}^{(i)})}{I_0(\rho_{tj}^{(i)})}\right]}{2\mathcal{P}_j(\hat{\beta}_j^{(i)})} \tag{75}$$

$$\hat{\lambda}_j^{(i+1)} = \frac{\mathcal{P}_j(\hat{\beta}_j^{(i)})}{\sum_{l=1}^{2} \mathcal{P}_l(\hat{\beta}_l^{(i)})} \tag{76}$$

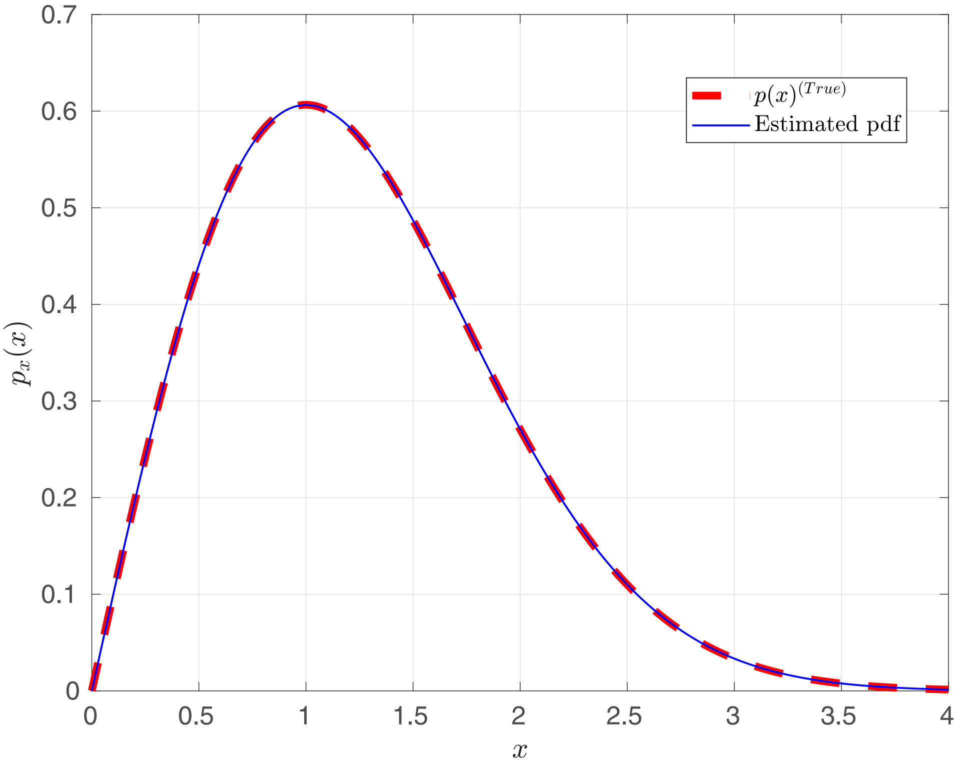

Fig 3. Rayleigh distribution estimation using a Rayleigh-Rice mixture.

https://doi.org/10.1371/journal.pone.0208499.g003

with

$$\rho_{tj}^{(i)} = \frac{x_t \hat{v}_j^{(i)}}{[\hat{\sigma}_j^2]^{(i)}} \tag{77}$$

$$\mathcal{P}_j(\hat{\beta}_j^{(i)}) = \sum_{t=1}^{N} \zeta_{tj}^{(i)} \tag{78}$$

We also consider the utilization of Akaike's Information Criterion (AIC) in order to obtain an accurate yet simple model and, thus, discriminating from a Rayleigh channel, a Rice channel, and a mixture of both.

With the above formulation, we consider two cases: a Rayleigh distributed channel and a Rice distributed channel.

6.2.1 Rayleigh distributed channel. In this example, the random variable x is drawn from the Rayleigh distribution

$$p(x)^{(\text{True})} = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{79}$$

with $\sigma^2 = 1$, using the *Slice Sampler* [69]. The best estimated model corresponds to the single Rayleigh component in the mixture. The corresponding estimation of σ^2 yields $\hat{\sigma}_1^2 = 1.0007 \pm 9.003 \times 10^{-4}$. Fig 3 shows the *true* Rayleigh distribution and the mean estimated pdf from the 50 MC simulations. We observe an important agreement between the true pdf and the estimated model.

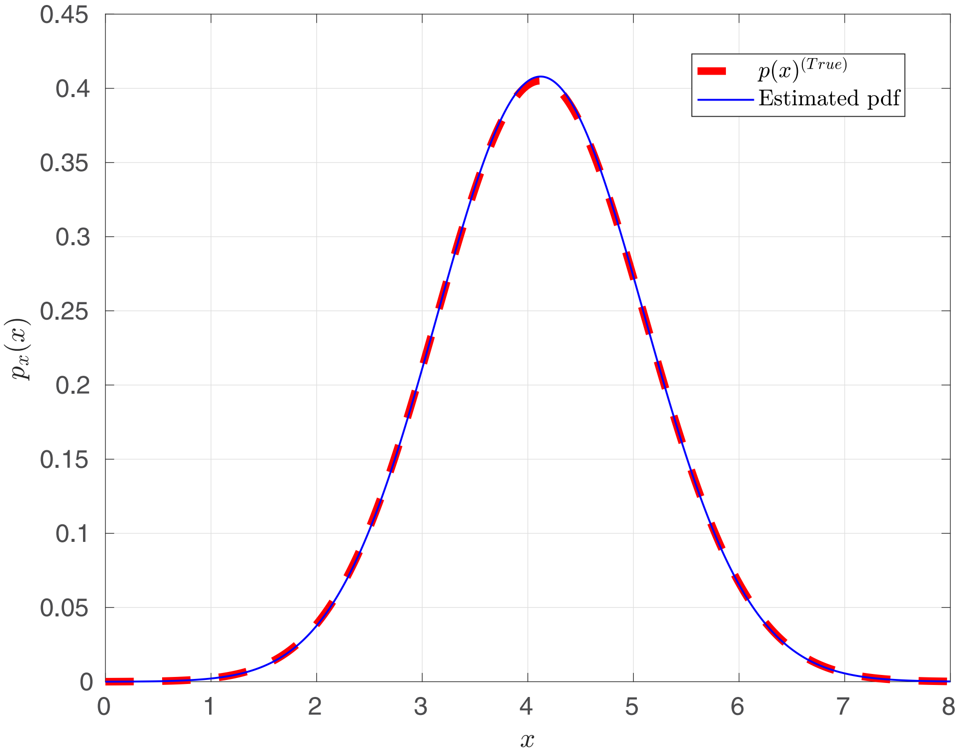

Fig 4. Rice distribution estimation using a Rayleigh-Rice mixture.

https://doi.org/10.1371/journal.pone.0208499.g004

6.2.2 Rice distributed channel. In this case, the data is drawn from the Rice distribution

$$p(x)^{(\mathrm{True})} = \frac{x}{\sigma^2} \exp\left(-\frac{x^2+v^2}{2\sigma^2}\right) I_0\left(\frac{xv}{\sigma^2}\right) \tag{80}$$

with $v = 4$ and $\sigma^2 = 1$, using the *Slice Sampler*. The best model is selected as a single Rician component. The corresponding estimated parameters are $\hat{v}_2 = 4.003 \pm 1.3 \times 10^{-3}$ and $\hat{\sigma}_2^2 = 0.9858 \pm 2.2 \times 10^{-3}$. In Fig 4 we show the *true* Rice distribution and the mean estimated pdf from 50 MC simulations. We can observe that the estimator exhibits a good performance for the estimation of a Rice distribution.

Conclusions and future work

In this paper we have presented a systematic approach for constructing surrogate functions in a wide range of inference problems. Our approach can be utilized for constructing surrogate functions for both the cost function and the constraints, generalizing the popular EM and MM algorithms. Our approach is based on the utilization of data augmentation and kernel functions, yielding simple optimization algorithms when the kernel can be expressed as VMGM. We have shown how our proposal can be utilized to solve inverse problems that are expressed as integral equations and mixture distributions.

In addition, we have shown that our approach can be utilised for constrained/penalized ML and MAP estimations problems. In particular, common problems in statistical inference can directly be solved using our proposal since they can be posed as Variance Mean Gaussian Mixtures (VMGM), yielding quadratic surrogate functions.

In the last two decades the problem of sparse estimation has attracted a lot of attention. Since our approach can be utilized in those problems, and since it is based on the principles of the MM algorithm, a detailed analysis can be done in terms of accuracy and convergence of our technique, and compared against other techniques, such as the ones in [28, 30], and [29], where different Lasso-type problems are compared, the MM algorithm is utilized in constrained problems for ML estimation in generalized linear model regression, and the MM algorithm is used for (unconstrained) sparse estimation under non-convex penalties, respectively.

Supporting information

S1 Data Set. Monte Carlo simulations for Examples 1 and 2.
(ZIP)

Acknowledgments

This work was partially supported by the Fondo Nacional de Desarrollo Científico y Tecnológico-Chile through grants No. 3140054 and 1181158. This work was also partially supported by the Comisión Nacional de Investigación Científica y Tecnológica, Advanced Center for Electrical and Electronic Engineering (AC3E, Proyecto Basal FB0008), Chile. The work of R. Orellana was partially supported by the PIIC program of the Universidad Técnica Federico Santa María, scholarship 015/2018. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Conceptualization: Rodrigo Carvajal, Juan Carlos Agüero.

Formal analysis: Rodrigo Carvajal, Juan Carlos Agüero.

Funding acquisition: Rodrigo Carvajal, Rafael Orellana, Juan Carlos Agüero.

Investigation: Rodrigo Carvajal, Dimitrios Katselis, Juan Carlos Agüero.

Methodology: Rodrigo Carvajal, Juan Carlos Agüero.

Resources: Juan Carlos Agüero.

Validation: Pedro Escárate.

Visualization: Pedro Escárate.

Writing – original draft: Rodrigo Carvajal, Rafael Orellana, Dimitrios Katselis, Pedro Escárate, Juan Carlos Agüero.

Writing – review & editing: Rodrigo Carvajal, Rafael Orellana, Dimitrios Katselis, Pedro Escárate, Juan Carlos Agüero.

References

1. Carvajal R, Agüero JC, Godoy BI, Goodwin GC. EM-Based Maximum-Likelihood Channel Estimation in Multicarrier Systems With Phase Distortion. IEEE Trans Vehicular Technol. 2013; 62(1):152–160. https://doi.org/10.1109/TVT.2012.2217361
2. Goldsmith A. Wireless Communications. New York, NY, USA: Cambridge University Press; 2005.
3. Godoy BI, Goodwin GC, Agüero JC, Marelli D, Wigren T. On identification of FIR systems having quantized output data. Automatica. 2011; 47(9):1905–1915. https://doi.org/10.1016/j.automatica.2011.06.008

4. Wang L, Zhang J, Yin GG. System identification using binary sensors. IEEE Trans Autom Control. 2003; 48(11):1892–1907. https://doi.org/10.1109/TAC.2003.819073

5. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from imcomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B. 1977; 39(1):1–38.

6. Van Dyk DA, Meng XL. The Art of Data Augmentation. Journal of Computational and Graphical Statistics. 2001; 10(1):1–50. https://doi.org/10.1198/10618600152418584

7. Hunter DR, Lange K. A tutorial on MM algorithms. The American Statistician. 2004; 58(1):30–37. https://doi.org/10.1198/0003130042836

8. Pollakis E, Cavalcante RLG, Stańczak S. Base Station Selection for Energy Efficient Network Operation with the Majorization- Minimization Algorithm. In: Proc. 13th IEEE Int. Workshop on Signal Process. Adv. Wireless Commun. (SPAWC 2012). Çeşme, Turkey; 2012.

9. Figueiredo MAT, Bioucas-Dias JM, Nowak RD. Majorization–Minimization Algorithms for Wavelet-Based Image Restoration. IEEE Trans Image Process. 2007; 16(12):2980–2991. https://doi.org/10.1109/TIP.2007.909318 PMID: 18092597

10. Marks BR, Wright GP. A General Inner Approximation Algorithm for Nonconvex Mathematical Programs. Operations Research. 1978; 26(4):681–683. https://doi.org/10.1287/opre.26.4.681

11. Agüero JC, Tang W, Yuz JI, Delgado R, Goodwin GC. Dual time-frequency domain system identification. Automatica. 2012; 48(12):3031–3041. https://doi.org/10.1016/j.automatica.2012.08.033

12. Beal MJ, Ghahramani Z. The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structures. In: Proc. of the 7th Valencia International Meeting. Valencia, Spain; 2003.

13. Gopaluni RB. A particle filter approach to identification of nonlinear processes under missing observations. Can J Chem Eng. 2008; 86(6):1081–1092. https://doi.org/10.1002/cjce.20113

14. Hobolth A, Jensen JL. Statistical Inference in Evolutionary Models of DNA Sequences via the EM Algorithm. Statistical Applications in Genetics and Molecular Biology. 2005; 4(1). https://doi.org/10.2202/1544-6115.1127 PMID: 16646835

15. Schön TB, Wills A, Ninness B. System identification of nonlinear state-space models. Automatica. 2011; 47(1):39–49. https://doi.org/10.1016/j.automatica.2010.10.013

16. Yang S, Kim JK, Zhu Z. Parametric fractional imputation for mixed models with nonignorable missing data. Statistics and Its Interface. 2013; 6(3):339–347. https://doi.org/10.4310/SII.2013.v6.n3.a4

17. Meng XL. Thirty Years of EM and Much More. Statistica Sinica. 2007; 17(3):839–840.

18. Suzuki H. A Statistical Model for Urban Radio Propogation. IEEE Transactions on Communications. 1977; 25(7):673–680. https://doi.org/10.1109/TCOM.1977.1093888

19. Curé M, Rial DF, Christen A, Cassetti J. A method to deconvolve stellar rotational velocities. Astronomy & Astrophysics. 2014; 565.

20. Kraus C, Bornschein B, Bornschein L, Bonn J, Flatt B, Kovalik A, et al. Final results from phase II of the Mainz neutrino mass searchin tritium β decay. The European Physical Journal C—Particles and Fields. 2005; 40(4):447–468.

21. Semkow TM, Li X. Application of integral equations to neutrino mass searches in beta decay; 2018. ArXiv e-prints. Available from: http://adsabs.harvard.edu/abs/2018arXiv180105009S.

22. Polson NG, Scott JG. Data augmentation for non-Gaussian regression models using variance-mean mixtures. Biometrika. 2013; 100:459–471. https://doi.org/10.1093/biomet/ass081

23. Barndorff-Nielsen O, Kent J, Sorensen M. Normal variance-mean mixtures and z distributions. Int Stat Review. 1982; 50(2):145–159. https://doi.org/10.2307/1402598

24. West M. On scale mixtures of normal distributions. Biometrika. 1987; 74(3):646–648. https://doi.org/10.1093/biomet/74.3.646

25. Balakrishnam N, Leiva V, Sanhueza A, Vilca F. Estimation in the Birnbaum-Saunders distribution based on scale-mixture of normals and the EM algorithm. SORT. 2009; 33(2):171–192.

26. Carvajal R, Agüero JC, Godoy BI, Katselis D. A MAP approach for ℓ_q-norm regularized sparse parameter estimation using the EM algorithm. In: Proc. of the 25th IEEE Int. Workshop on Mach. Learning for Signal Process (MLSP 2015). Boston, USA; 2015.

27. Godoy BI, Agüero JC, Carvajal R, Goodwin GC, Yuz JI. Identification of sparse FIR systems using a general quantisation scheme. Int J Control. 2014; 87(4):874–886. https://doi.org/10.1080/00207179.2013.861611

28. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B. 2006; 68(1):49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

29. Fan J, Liu H, Sun Q, Zhang T. I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. Ann Statist. 2018; 46(2):814–841. https://doi.org/10.1214/17-AOS1568

30. Xu J, Chi E, Lange K. Generalized Linear Model Regression under Distance-to-set Penalties. In: Proc. of 30th Neural Information Processing Systems Conference. Barcelons, Spain; 2016. p. 1385–1395.

31. Durret R. Probability: Theory and examples. 4th ed. Cambridge University Press; 2010.

32. Akaike H. A new look at the statistical model identification. IEEE Trans Aut Control. 1974; 19(10):716–723. https://doi.org/10.1109/TAC.1974.1100705

33. McLachlan GJ, Krishnan T. The EM Algorithm and Extensions. Wiley; 1997.

34. Vaida F. Parameter convergence for EM and MM algorithms. Statistica Sinica. 2005; 15(3):831–840.

35. Sriperumbudur BK, Torres DA, Lanckriet GRG. A majorization-minimization approach to the sparse generalized eigenvalue problem. Machine Learning. 2011; 85(1–2):3–39. https://doi.org/10.1007/s10994-010-5226-3

36. Tanner MA. Tools for Statistical Inference: Observed Data and Data Augmentation Methods. Springer; 1991. https://doi.org/10.1007/978-1-4684-0510-1

37. Figueiredo MAT. Adaptive sparseness for supervised learning. IEEE Trans Pattern Anal Mach Intell. 2003; 25(9):1150–1159. https://doi.org/10.1109/TPAMI.2003.1227989

38. Park T, Casella G. The Bayesian Lasso. Journal of the American Statistical Association. 2008; 103 (482):681–686. https://doi.org/10.1198/016214508000000337

39. Garrigues P, Olshausen B. Group sparse coding with a laplacian scale mixture prior. Advances in Neural Information Processing Systems. 2010; 23:676–684.

40. Wazwaz AM. Linear and Nonlinear Integral Equations: Methods and Applications. Springer; 2011.

41. Christen A, Escarate P, Curé M, Rial DF, Cassetti J. A method to deconvolve stellar rotational velocities II. Astronomy & Astrophysics. 2016; 595(A50):1–8.

42. Chandrasekhar S, Münch G. On the Integral Equation Governing the Distribution of the True and the Apparent Rotational Velocities of Stars. Astrophysical Journal. 1950; 111:142–156. https://doi.org/10.1086/145245

43. Molisch A. Wireless Communications. Wiley-IEEE Press; 2005.

44. Salous S. Radio Propagation Measurement and Channel Modelling. Wiley; 2013.

45. Pätzold M. Mobile Radio Channels. Wiley; 2011.

46. Mengersen KL, Robert C, Titterington M. Mixtures: Estimation and Applications. Wiley; 2011.

47. McLachlan G, Peel D. Finite Mixture Models. Wiley; 2004.

48. DeGroot M. Optimal statistical decisions. New York, NY, USA: McGraw-Hill; 1970.

49. Redner RA, Walker HF. Mixture Densities, Maximum Likelihood and the EM Algorithm. SIAM Review. 1984; 26(2):195–239. https://doi.org/10.1137/1026034

50. Orellana R, Carvajal R, Agüero JC. Maximum Likelihood Infinite Mixture Distribution Estimation Utilizing Finite Gaussian Mixtures. In: 18th IFAC Symposium on System Identification (SYSID). Stockholm, Sweden; 2018.

51. Goodwin GC, Payne RL. Dynamic system identification: experiment design and data analysis. Academic Press New York; 1977.

52. Söderström T, Stoica P, editors. System Identification. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.; 1988.

53. Ljung L, Goodwin GC, Agüero JC. Stochastic Embedding revisited: A modern interpretation. In: Proc. of the 53rd IEEE Conf. Decision and Control). Los Angeles, CA, USA; 2014.

54. Mahata K, Söderström T. Improved estimation performance using known linear constraints. Automatica. 2004; 40(8):1307–1318. https://doi.org/10.1016/j.automatica.2004.03.001

55. Lange K. Optimization. 2nd ed. New York, USA: Springer; 2013.

56. Hanson MA. Inequality constrained maximum likelihood estimation. Ann Inst Stat Math. 1965; 17 (1):311–321. https://doi.org/10.1007/BF02868175

57. Hyder MM, Mahata K. A Robust Algorithm for Joint-Sparse Recovery. Signal Process Letters, IEEE. 2009; 16(12):1091–1094. https://doi.org/10.1109/LSP.2009.2028107

58. Grant MC, Boyd SP. Graph Implementations for Nonsmooth Convex Programs. In: Blondel VD, Boyd SP, Kimura H, editors. Recent Advances in Learning and Control. London: Springer; 2008. p. 95–110.

59. Lemieux C. Monte Carlo and Quasi-Monte Carlo Sampling. New York, NY, USA: Springer; 2009.

60. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes: The Art of Scientific Computing. 3rd ed. Cambridge University Press; 2007.

61. Tibshirani R. Regression shrinkage and selection via the lasso. J Royal Statist Soc B. 1996; 58(1):267–288.

62. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B. 2005; 67:301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

63. Li Q, Lin N. The Bayesian elastic net. Bayesian Anal. 2010; 5(1):151–170. https://doi.org/10.1214/10-BA506

64. Fletcher R. Practical methods of optimization. 2nd ed. Chichester, GB: John Wiley & Sons; 1987.

65. Izmailov AF, Solodov MV. Newton-type methods for optimization and variational problems. Cham, Switzerland: Springer; 2014.

66. Deutsch AJ. Maxwellian distributions for stellar rotation. In: Proc. of the IAU Colloquium on Stellar Rotation. Ohio, USA; 1969. p. 207–218.

67. Robert CP, Casella G. Monte Carlo statistical methods. Springer, New York; 1999.

68. Rodríguez M, Feick R, Carrasco H, Valenzuela R, Derpich M, Ahumada L. Wireless Access Channels with Near-Ground Level Antennas. IEEE Transactions on Wireless Communications. 2012; 11 (6):2204–2211. https://doi.org/10.1109/TWC.2012.041612.110735

69. Neal RM. Slice sampling. Ann Statist. 2003; 31(3):705–767. https://doi.org/10.1214/aos/1056562461