EasyChair Preprint

№ 13763

On Singular Bayesian Inference of Underdetermined Quantities

Fabrice Pautot

July 2, 2024

physical sciences forum

MDPI

Proceeding Paper

On Singular Bayesian Inference of Underdetermined Quantities†

Part I: Invariant discrete ill-posed inverse problems in small and large dimension

Fabrice Pautot, Independent Researcher, fabrice.pautot@proton.me

† Presented at the 43th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Ghent, Belgium, 1–5 July 2024

Abstract: When the quantities of interest remain underdetermined a posteriori, we would like to draw inferences for at least one particular solution. Can we do that in a Bayesian way? What is a probability distribution over an underdetermined quantity? How to get a posterior for one particular solution from a posterior for infinitely many underdetermined solutions? Guided by invariant underdetermined ill-posed inverse problems, we find that a probability distribution over an underdetermined quantity is non-absolutely continuous, partially improper wrt the initial reference measure but proper wrt its restriction to its support. Thus, it is necessary and sufficient to choose the prior restricted reference measure to assign partially improper priors by e.g. maximum entropy and the posterior restricted reference measure to obtain the proper posterior for one particular solution. We can then work with underdetermined models such as Hoeffding-Sobol expansions seamlessly, especially to effectively counter the curse of dimensionality within nonparametric inverse problems. We demonstrate Singular Bayesian Inference (SBI) at work in an advanced Bayesian Optimization application: dynamic pricing. Such a nice generalization of Bayesian-maxentropic inference could motivate many theoretical and practical developments.

Keywords: Underdetermined/indeterminate/non-identifiable/invariant quantities, partially improper/degenerate/singular/non absolutely continuous probability measures, reference measure, ill-posed inverse problems, inter/extrapolation, curse of dimensionality, MaxEnt, HDMR/Hoeffding-Sobol expansions/fANOVA/interactive splines.

1. Introduction

Many problems in science and engineering, especially inverse ones, involve quantities, parameters or solutions that are underdetermined and therefore non-identifiable a priori and sometimes remain so a posteriori. For example, if a statistical or physical model involves a sum resp. a product of several parameters, then these are (under)determined, invariant up to additive resp. multiplicative constants, a priori and a posteriori. For example, in medical dynamic contrast-enhanced imaging, the kinetic continuity equation for a contrast agent advected by the blood involves the ratios of the plasmatic volumetric flow rates to the plasma volume [1] (p. 20, eq. 47). Those parameters are therefore globally non-identifiable until we add further cogent information and that can become a challenge in medical research and clinical practice. Similarly, the solutions of a consistent underdetermined, i.e. indeterminate system of linear equations $\mathbf{Ax} = \mathbf{b}$ (like the cubic spline coefficients below) are determined up to $\ker(\mathbf{A})$.

When such quantities remain underdetermined a posteriori, we would like to estimate and draw inferences for at least one particular solution, from which we could, if necessary, estimate and draw inferences for all the solutions. If it is common to do this in a non-Bayesian way, e.g. by evaluating the particular solution $\mathbf{x}_0 = \mathbf{A}^+\mathbf{b}$ and the general

Citation: To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Published: date

solution $\mathbf{x} = \mathbf{A}^{+}\mathbf{b} + \left(\mathbf{I}_n - \mathbf{A}^{+}\mathbf{A}\right)\mathbf{w}, \mathbf{w} \in \mathbb{R}^n$ for an indeterminate system of linear equations, as far as we know this has not been done yet in a probabilistic way. That is unfortunate since such an approach should make it possible to obtain the credible intervals or the High Posterior Density Regions that must accompany any estimate or measurement according to e.g. the ISO *Guide to the expression of uncertainty in measurements* [2], well-determined or not. The purpose of Bayesian Numerical Linear Algebra is precisely to estimate the solutions of systems of linear equations together with their credible intervals, but to the best of our knowledge it is currently limited to well-determined systems with non-singular, positive definite matrices [3].

What is a probability distribution $p(\mathbf{x})$ when $\mathbf{x}$ is underdetermined? How to assign such distributions by standard means like the principle of maximum entropy? Given a posterior $p(\mathbf{x}|D)$ for some underdetermined $\mathbf{x}$, how to get a posterior $p(\mathbf{x}_0|D)$ for one particular solution $\mathbf{x}_0$? Where does $\mathbf{x}_0$ come from? While these questions may seem puzzling at first, the situation clears up considerably once we return to invariant ill-posed nonparametric inverse problems whose solutions are in general underdetermined at least a priori.

2. Invariant discretized nonparametric ill-posed inverse problems in small dimension

Without this belief [in the principle of continuity]…,
interpolation would be impossible…, science would not exist.
Henri Poincaré [4]

Functional, nonparametric inverse ill-posed problems like inter/extrapolation, also known as functional regression, deconvolution or reconstruction are ubiquitous in all experimental sciences. For Poincaré [4] and many others they are nothing but Bayesian problems. To make it concrete, and without loss of generality, let us nevertheless restart with the variational formulation of the classical (noisy) inter/extrapolation or functional regression problem

$$\hat{f} = \underset{f \in W^{2,2}([a,b])}{\arg\min} \quad \frac{1}{N}\sum_{i=1}^{N}\left(y_i - f\left(x_i\right)\right)^2 + \frac{\lambda}{b-a}\int_a^b f^{(k)}\left(x\right)^2 \mathrm{d}x \quad (1)$$

We want to reformulate and process it in in a purely Bayesian way. There are at least two main issues in this endeavor.

2.1 Infinite-dimensional function spaces: discretization

The first one is the need to define probability measures and to draw inferences over infinite-dimensional function spaces such as the Sobolev space $W^{2,2}\left([a,b]\right)$. Interestingly, this issue has been addressed in several ways that are deeply interconnected but nevertheless yield significantly different solutions, including:

- **Reproducing Kernel Hilbert Spaces** (RHKS) and **random/stochastic process** theories [5][6];
- In some extent, **Gaussian process (GP) functional regression**, also known as kriging or Wiener-Kolmogorov prediction. But GP regression is anything but Bayesian since the GP "prior" is updated via "GP conditioning" instead of Bayes rule [7][8];
- **Information Field Theory** whose purpose is to *directly* generalize classical, finite-dimensional Bayesian inference to countably infinite dimension using tools borrowed from Quantum Field Theory like path integrals [9];
- **Discretization or projection** [10][11], that is estimating only a finite subset $\mathbf{f} = \left\{f\left(x^1\right),...,f\left(x^n\right)\right\}$ of the function images on a grid $G = \left\{x^1,...,x^n\right\}$ by approxi-

mating the differential operator $\mathrm{d}^k / \mathrm{d}x^k$ by a finite differences scheme and the integral $\int_a^b f^{(k)}(x)^2 \, \mathrm{d}x$ by a numerical method like the trapezoidal rule [11]. In this way we get rid of any measure-theoretic issue, or rather we shall better control them, and we don't need any structure nor machinery beyond the original function space.

Generally speaking, the last approach is the "cheapest", the safest and the right one as soon as we consider that it is better to take the limit $n \to +\infty$ only at the very end of the calculations, not at the very beginning (Gauss, Poincaré, Jaynes).

But in functional problems, there is a special reason for doing so. We are supposed to propagate the uncertainties on all parameters of interest by computing their marginal posteriors and by taking Bayes estimators such as marginal posterior expectations under quadratic loss function together with credible intervals such as marginal posterior standard deviations. Having the right posterior credible intervals or HPDR is crucial, especially if the function estimates are to be used as meta-/surrogate models within Bayesian Optimization or Design of Experiments that entirely rely on uncertainty quantification. In other words, we dismiss the Maximum a posteriori estimator (MAP) because it does not propagate uncertainty, so that there is no direct mapping between variational minimization problems and their truly Bayesian counterparts.

By completeness, a continuous function is uniquely given by the set of its images on a countable but dense subset of its domain. Therefore, estimating a continuous function boils down to estimating a countably infinite number of parameters. Given that estimating n parameters with uncertainty propagation requires the calculation of at least n $n-$dimensional integrals (e.g. the marginal posterior expectations), from a purely Bayesian standpoint, estimating a continuous function boils down to evaluating a countably infinite number of countably infinite-dimensional integrals. From this standpoint, estimating only a finite set of the function images and computing finite-dimensional integrals in a first step definitely appears to be a reasonable choice.

2.2 Invariance up to polynomials and partially improper priors

The second difficulty can occur for any $k > 0$. Poincaré principle of continuity above corresponds to $k > 1$. The null space of the differential operator $\mathrm{d}^k / \mathrm{d}x^k$ is the $k-$dimensional vector space of polynomials of degree at most $k-1$. Hence, the function $f(x)$ is a priori (under)determined, invariant up to those polynomials unless we add sufficiently many boundary conditions to break this invariance a priori and, subsequently, a posteriori [8] (p. 6).

As an example, in the variational setting with $k = 2$, recall that the solution of (1) is given by underdetermined cubic splines with $4n-4$ unknown coefficients and $4n-6$ conditions. Hence, we typically add two extra boundary conditions $f^{(2)}(a) = f^{(2)}(b) = 0$ to make the coefficients well-determined and to finally get *natural cubic splines* [6] (pp. xii-xiii). But in most circumstances such boundary conditions do not exist, especially when we need to extrapolate the function outside the range of past observations, which is impossible with cubic splines.

As observed by many authors [5][6][12][13][14][15][17], from the Bayesian standpoint, the prior invariance modulo polynomials of a regularization penalty with a differential operator implies that the corresponding prior is "partially improper" [6] (and non-informative) or, conversely, "partially informative" [16]. In measure theory, such non-absolutely continuous measures that concentrate their mass on a Lebesgue-negligible subset/subspace are also known as degenerate measures or singular probability distributions.

Precisely, upon discretizing the problem, we find that the quadratic form or precision matrix $\mathbf{R}$ for the regularization penalty with differential operator $\mathrm{d}^k / \mathrm{d}x^k$ has rank deficiency k without extra boundary conditions. For sake of simplicity, we skip all technicali-

ties and we simply assume that the discretization grid $G=\left\{x^1,...,x^n\right\}$ is regular $G=\left\{x^1=a,x^2=a+\Delta x,...,x^n=b\right\}$ with discretization step Δx and that $\left\{x_i,i=1,N\right\}\subset G$. G may be larger than the range of $\left\{x_i,i=1,N\right\}$, e.g. in extrapolation problems, and G may be finer than the natural grid, e.g. if we want to oversample a periodically sampled signal.

As an example, we numerically approximate $f^{(2)}(x)$ over G by $\mathbf{Lf}$ with a second-order accuracy centered finite differences scheme on the interiors points $x^i, i=2,n-1$, a second-order forward scheme on the left boundary $x^1=a$ and a second-order backward scheme for the right boundary $x^n=b$ [17]:

$$\mathbf{L}=\frac{1}{\Delta x^2}\begin{pmatrix}2&-5&4&-1&0&\dots&0\\1&-2&1&0&&&\vdots\\0&\ddots&\ddots&\ddots&\ddots&&\vdots\\\vdots&\ddots&\ddots&\ddots&\ddots&\ddots&\vdots\\\vdots&&\ddots&\ddots&\ddots&\ddots&0\\\vdots&&&0&1&-2&1\\0&\dots&0&-1&4&-5&2\end{pmatrix}$$

To implement e.g. the trapezoidal rule for numerical integration, we just divide the first and last rows of $\mathbf{L}$ by $\sqrt{2}$. Finally $\frac{1}{b-a}\int_a^b f^{(2)}(x)^2\,\mathrm{d}x\simeq\frac{\Delta x}{b-a}\left|\mathbf{Lf}\right|_2^2=\frac{\Delta x}{b-a}\mathbf{f}^{\mathbf{T}}\mathbf{L}^{\mathbf{T}}\mathbf{Lf}\triangleq\mathbf{f}^{\mathbf{T}}\mathbf{Rf}$

$\mathbf{L}$ and $\mathbf{R}$ have rank deficiency 2 as expected. Now, if we add for instance two Dirichlet boundary conditions $f(a)=f(b)=0$, we can use a centered finite differences scheme at the new boundary points x^2 and x^{n-1} and $\mathbf{L}$ becomes the $(n-2)\times(n-2)$ full rank matrix

$$\mathbf{L}=\frac{1}{\Delta x^2}\begin{pmatrix}-2/\sqrt{2}&1/\sqrt{2}&0&\dots&\dots&0\\1&-2&1&0&&\vdots\\0&\ddots&\ddots&\ddots&\ddots&\vdots\\\vdots&\ddots&\ddots&\ddots&\ddots&0\\\vdots&&0&1&-2&1\\0&\dots&\dots&0&1/\sqrt{2}&-2/\sqrt{2}\end{pmatrix}$$

Let $\mathbf{X}\triangleq(x_1,...,x_N)^{\mathbf{T}}$, $\mathbf{Y}\triangleq(y_1,...,y_N)^{\mathbf{T}}$ and $\varepsilon\triangleq\sigma\sqrt{\lambda}$. Suppose that the likelihood is i.i.d. maxentropic Gaussian with standard deviation σ. There exists a $n\times n$ diagonal precision matrix/quadratic form $\mathbf{D}$ and a $n\times1$ column vector $\mathbf{J}$ such as

$$p(\mathbf{Y}|\mathbf{f},\mathbf{X},\sigma)\propto\sigma^{-N}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-f(x_i))^2}\propto\sigma^{-N}e^{-\frac{1}{2\sigma^2}\left(\mathbf{f}^{\mathbf{T}}\mathbf{Df}-2\mathbf{J}^{\mathbf{T}}\mathbf{f}+\mathbf{Y}^{\mathbf{T}}\mathbf{Y}\right)}$$

2.3 The regular case

If $\mathbf{R}$ is positive definite (e.g. Tikhonov regularization $k=0$ or k linearly independent extra boundary conditions [8]), we apply the principle of maximum entropy (MaxEnt) by constraining the first two moments $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}=\mathbf{R}^{-1}$ to convert the regularization quadratic form $\mathbf{f}^{\mathbf{T}}\mathbf{Rf}$ into the proper multivariate regularization/smoothing Gaussian prior $p(\mathbf{f}|\boldsymbol{\mu},\mathbf{R},\varepsilon,\sigma)\propto(\varepsilon/\sigma)^n e^{-\frac{\varepsilon^2}{\sigma^2}\frac{(\mathbf{f}-\boldsymbol{\mu})^{\mathbf{T}}\mathbf{R}(\mathbf{f}-\boldsymbol{\mu})}{2}}$.

With e.g. $p(\boldsymbol{\mu})\propto1$, $p(\sigma)\propto\sigma^{-1}$ and $p(\varepsilon)\propto\varepsilon^{-1}$, the joint posterior writes

$$p(\mathbf{f},\sigma,\varepsilon,\boldsymbol{\mu}|\mathbf{X},\mathbf{Y},\mathbf{R})\propto p(\mathbf{Y}|\mathbf{f},\mathbf{X},\sigma)\,p(\mathbf{f}|\boldsymbol{\mu},\mathbf{R},\varepsilon,\sigma)\,p(\boldsymbol{\mu})\,p(\sigma)\,p(\varepsilon)\propto$$
$$\sigma^{-(N+n+1)}\varepsilon^{n-1}e^{-\frac{1}{2\sigma^2}\left(\mathbf{f}^{\mathbf{T}}\left(\mathbf{D}+\varepsilon^2\mathbf{R}\right)\mathbf{f}-2\left(\mathbf{J}+\varepsilon^2\mathbf{R}\boldsymbol{\mu}\right)^{\mathbf{T}}\mathbf{f}+\varepsilon^2\boldsymbol{\mu}^{\mathbf{T}}\mathbf{R}\boldsymbol{\mu}+\mathbf{Y}^{\mathbf{T}}\mathbf{Y}\right)}$$

For comparison with the singular case below, in particular we have

$$\mathbb{E}\,\mathbf{f}\left|\mathbf{X},\mathbf{Y},\boldsymbol{\mu}=0,\mathbf{R},\varepsilon,\sigma=\left(\mathbf{D}+\varepsilon^2\mathbf{R}\right)^{-1}\mathbf{J}\right.$$

$$\mathbb{V}\,\mathbf{f}\left|\mathbf{X},\mathbf{Y},\boldsymbol{\mu}=0,\mathbf{R},\varepsilon,\sigma=\sigma^2\mathrm{diag}\left(\mathbf{D}+\varepsilon^2\mathbf{R}\right)^{-1}\right.$$

$$p\left(\varepsilon\left|\mathbf{X},\mathbf{Y},\boldsymbol{\mu}=0,\mathbf{R},\sigma\right.\right)\propto\varepsilon^{n-1}/\sqrt{\left|\mathbf{D}+\varepsilon^2\mathbf{R}\right|}e^{\frac{1}{2}\mathbf{J}^{\mathbf{T}}\left(\mathbf{D}+\varepsilon^2\mathbf{R}\right)^{-1}\mathbf{J}}$$

From the linear algebra standpoint, the posterior precision matrix/quadratic form $\mathbf{D}+\varepsilon^2\mathbf{R}$ is a symmetric linear matrix pencil $(\mathbf{D},\mathbf{R})$ [18].

2.4 The singular case

When $\mathbf{R}$ is singular positive semi-definite, which is the most common situation, we face two issues.

First, we cannot apply MaxEnt *directly* anymore to get a partially improper Gaussian regularization prior like $p\left(\mathbf{f}\left|\mathbf{R},\varepsilon,\sigma\right.\right)\propto\left(\varepsilon/\sigma\right)^{\mathrm{rank}(\mathbf{R})}e^{-\frac{\varepsilon^2}{2\sigma^2}\mathbf{f}^{\mathbf{T}}\mathbf{R}\mathbf{f}}$ since it has no differential entropy. For the time being, let us nevertheless assume that we can assign such prior. By contrast to the regular case where the proper prior is not location-invariant because it has a first moment so that we must introduce a location parameter $\boldsymbol{\mu}$ even if it does not change the differential entropy, it makes no sense to introduce such a location parameter in a partially improper prior because it has no first moment. It follows that the posterior expectations will be entirely determined by the data. That's exactly what we want: the problem being a priori location-invariant, we should not say anything at all about location: a location-invariant prior must not have a first moment. A non-location-invariant proper prior with a non-informative location hyperprior e.g. $p(\boldsymbol{\mu})\propto 1$ has nothing to do with a location-invariant prior that is necessarily partially improper.

Second, the posterior precision matrix and pencil may be singular and positive semi-definite too. If $\mathbf{R}$ or $\mathbf{D}$ is positive definite, then $(\mathbf{D},\mathbf{R})$ is regular, i.e. $\exists\varepsilon>0,\left|\mathbf{D}+\varepsilon^2\mathbf{R}\right|>0$. $\mathbf{D}$ has rank equal to the number of different values in $\{x_i, i=1,N\}$ and is positive definite iff $G\subset\{x_i, i=1,N\}$. If $\mathbf{R}$ has rank deficiency k, $(\mathbf{D},\mathbf{R})$ has rank deficiency at least $\max(0,k-N)$. Therefore, $(\mathbf{D},\mathbf{R})$ is singular, i.e. $\forall\varepsilon>0,\left|\mathbf{D}+\varepsilon^2\mathbf{R}\right|\equiv 0$, and the joint posterior is singular as well (in the sense of probability theory, i.e. still non absolutely continuous, partially improper, degenerate) for at least all $N<k$.

In many applications, the sample size N is very large compared to k so that the posterior pencil is non-singular with very high probability. But thats not true in other important, "small data" applications like Bayesian optimization or Design of Experiments: starting from only two samples (x_1,y_1) and (x_2,y_2) chosen at random, the goal is to find the next sample (x_3,y_3) that optimizes some criterion, e.g. minimizes the predictive Shannon entropy of the $\arg\max$ of an expensive black-box function. $\mathbf{D}$ has rank at most 2 and, without extra boundary conditions, the joint posterior is singular for any $k>2$: $\mathbf{f}$ remains underdetermined a posteriori but we nevertheless need to estimate at least one next optimal sample.

Hence, it appears that probability distributions over underdetermined quantities are partially improper. It remains to explain how to assign them and how to estimate at least one particular solution from a partially improper posterior when the solutions remain underdetermined a posteriori.

3. Partially improper-proper measures or how to estimate underdetermined quantities

Let $X\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ be a $n-$dimensional partially improper Gaussian measure with singular positive semi-definite, $k-$rank deficient covariance matrix $\boldsymbol{\Sigma}=\mathbf{R}^+$ or precision matrix $\boldsymbol{\Sigma}^+=\mathbf{R}$. X concentrates its mass on its $(n-k)-$ dimensional support

$\mathrm{Supp}(X)=\left\{\mathbf{x}\in\mathbb{R}^n,(\mathbf{x}-\boldsymbol{\mu})^{\mathbf{T}}\mathbf{R}(\mathbf{x}-\boldsymbol{\mu})>0\right\}=\left\{\boldsymbol{\mu}+\boldsymbol{\Sigma}^{1/2}\mathbf{v},\mathbf{v}\in\mathbb{R}^n\right\}$ that has Lebesgue measure 0. It follows that $\forall E\subset\mathrm{Supp}(X),\ p(X\in E)=0\Rightarrow\forall E\subset\mathbb{R}^n,\ p(X\in E)=0$.

Thus, such distribution is not absolutely continuous with respect to the reference measure, here the Lebesgue measure $\lambda(\mathbb{R}^n)=\mathrm{d}x_1...\mathrm{d}x_n$. Subsequently, it has no probability density function, i.e. no Radon-Nikodym derivative $\mathrm{d}X/\mathrm{d}\lambda(\mathbb{R}^n)$, no moments, no differential entropy, nothing.

However, by the disintegration theorem, X is proper wrt the restriction $\lambda(\mathrm{Supp}(X))$ of the Lebesgue reference measure $\lambda(\mathbb{R}^n)$ to its support at the same time, with probability density function $p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})=\mathrm{d}X/\mathrm{d}\lambda\left[\mathrm{Supp}(X)\right]\propto 1/\sqrt{|\boldsymbol{\Sigma}|}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathbf{T}}\boldsymbol{\Sigma}^{+}(\mathbf{x}-\boldsymbol{\mu})}$ where $\boldsymbol{\Sigma}^{+}$ stands for the Moore-Penrose pseudo-inverse and $|\boldsymbol{\Sigma}|^{+}$ for the pseudo-determinant [19][20].

We are free to choose any reference measure we like. So, replacing the unrestricted original reference measure by the dominating restricted one is how we go back and forth from the world of infinitely many underdetermined quantities to the world of one particular solution. We proceed as follows.

1) Select the r*estricted prior reference measure* $\lambda_{\text{prior}}\triangleq\lambda\left[\mathrm{Supp}(\mathbf{R})\right]$ and apply MaxEnt without constraining the first moment to get the proper Gaussian regularization prior $p(\mathbf{f}|\mathbf{R},\varepsilon,\sigma)=\mathrm{d}(\mathbf{f}|\mathbf{R},\varepsilon,\sigma)/d\lambda_{\text{prior}}\propto(\sigma/\varepsilon)^{\mathrm{rank}(\mathbf{R})}e^{-\frac{\varepsilon^2}{2\sigma^2}\mathbf{f}^{\mathbf{T}}\mathbf{R}^{+}\mathbf{f}}$ whose differential entropy is $\frac{1}{2}\ln\left|2\pi e\mathbf{R}^{+}\right|^{+}$.

2) Select the initial reference measure $\lambda(\mathbb{R}^n)$ seen as the product measure $\lambda_{\text{prior}}\times\lambda\left[\mathrm{Supp}(\mathbf{R})^{\perp}\right]$. The prior becomes partially improper with pseudo, unnormalizable probability density function $p(\mathbf{f}|\mathbf{R},\varepsilon,\sigma)=\mathrm{d}(\mathbf{f}|\mathbf{R},\varepsilon,\sigma)/\mathrm{d}\lambda(\mathbb{R}^n)\propto(\varepsilon/\sigma)^{\mathrm{rank}(\mathbf{R})}e^{-\frac{\varepsilon^2}{2\sigma^2}\mathbf{f}^{\mathbf{T}}\mathbf{R}\mathbf{f}}$.

3) Select the *restricted posterior reference measure* $\lambda_{\text{posterior}}\triangleq\lambda\left[\mathrm{Supp}(\mathbf{D},\mathbf{R})\right]$ and write Bayes rule directly wrt it for one particular solution X_0

$$\frac{\mathrm{d}(X_0|D)}{\mathrm{d}\lambda_{\text{posterior}}}\propto\frac{\mathrm{d}(D|X)}{\mathrm{d}\lambda_{\text{posterior}}}\frac{\mathrm{d}(X)}{\mathrm{d}\lambda_{\text{posterior}}}$$

In this way, after all, partially improper posteriors disappear. Generally speaking, both $\mathrm{d}(X)/\mathrm{d}\lambda_{\text{posterior}}$ and $\mathrm{d}(D|X)/\mathrm{d}\lambda_{\text{posterior}}$ still are partially improper wrt to $\lambda_{\text{posterior}}$ because $\mathrm{Supp}(\mathbf{D})\subset\mathrm{Supp}(\mathbf{D},\mathbf{R})$ and $\mathrm{Supp}(\mathbf{R})\subset\mathrm{Supp}(\mathbf{D},\mathbf{R})$ but that does not matter. The posterior $\mathrm{d}(X_0|D)/\mathrm{d}\lambda_{\text{posterior}}$ e.g. $p(\mathbf{f}_0,\sigma,\varepsilon|\mathbf{X},\mathbf{Y},\mathbf{R})$ for the particular solution $\mathbf{f}_0$ is proper Gaussian with posterior precision matrix $\mathbf{D}+\varepsilon^2\mathbf{R}$.

At the end, we just need to replace the posterior inverse by the pseudo-inverse and the determinant by the pseudo-determinant to estimate one particular solution $\mathbf{f}_0$

$$\begin{aligned}
&\mathbb{E}\,\mathbf{f}_0|\mathbf{X},\mathbf{Y},\mathbf{R},\varepsilon,\sigma=\left(\mathbf{D}+\varepsilon^2\mathbf{R}\right)^{+}\mathbf{J}\\
&\mathbb{V}\,\mathbf{f}_0|\mathbf{X},\mathbf{Y},\mathbf{R},\varepsilon,\sigma=\sigma^2\mathrm{diag}\left(\mathbf{D}+\varepsilon^2\mathbf{R}\right)^{+}\\
&p(\varepsilon|\mathbf{X},\mathbf{Y},\mathbf{R})\propto\varepsilon^{\mathrm{rank}(\mathbf{R})-1}/\sqrt{\left|\mathbf{D}+\varepsilon^2\mathbf{R}\right|^{+}}\,e^{\frac{1}{2}\mathbf{J}^{\mathbf{T}}\left(\mathbf{D}+\varepsilon^2\mathbf{R}\right)^{+}\mathbf{J}}
\end{aligned}$$

At this point we should provide some experimental results but let us first deal with the intra/extrapolation problem in arbitrary dimension.

4. Invariant discrete ill-posed inverse problems in large dimension

4.1. Tackling the curse of dimensionality

Now, we consider the problem of inter/extrapolating a scalar function $f(v_1,...,v_d)$ of d variables. The $d-$dimensional, $k-$order regularization penalty becomes e.g.

$$R^{(k)}(f)=\sum_{\alpha_1+...+\alpha_d=k}\int\left(\frac{\partial^k f}{\partial v_1^{\alpha_1}...\partial v_d^{\alpha_d}}\right)^2 \mathrm{d}v_1...\mathrm{d}v_d \qquad (2)$$

Clearly, the main issue is the curse of dimensionality (COD): the number of function images to be estimated on the $d-$dimensional hypergrid G is exponential in d. We can think about many approaches to fight the COD such as sparse grids [21] but SBI provides an extremely simple and powerful, fully probabilistic, semi-nonparametric way to go polynomial in d.

Basically, we want to approximate function f by some functions whose number of variables is smaller than d and bounded. To remain Gaussian when both the likelihood and the prior are Gaussians, we must remain quadratic and, subsequently, linear, additive. Therefore, we are led to approximate function f by a sum of functions.

A first possibility, motivated by the Kolmogorov-Arnold representation theorem, is to introduce a *generalized additive model* (GAM) [22] that approximates f as a sum of univariate functional components $f(v_1,...,v_d)\simeq f_0+\sum_{i=1}^{d} f(v_i)$. Clearly, those components are underdetermined up to additive constants but that identifiability issue is easily fixed by requiring all components to have e.g. zero mean, which are linear constraints. However, GAM are coarse and certainly not the best approximation when we are precisely interested in the interactions between tuples of variables v_i.

4.2. Classical constrained well-determined HDMR

A much more powerful model, which includes GAM at first order, is High Dimensional Model Representation (HDMR) [23][24][25][26], also known as (generalized) Hoeffding-Sobol expansions [25][26], (generalized) functional ANOVA decomposition [23][25][27] or interactive spline models [6][28]. We can write any function of d variables as

$$f(v_1,...,v_d)=f_0+\sum_{k=1}^{d} f_k(v_k)+\sum_{k=1}^{d}\sum_{l=k+1}^{d} f_{k,l}(v_k,v_l)+\sum_{k=1}^{d}\sum_{l=k+1}^{d}\sum_{m=l+1}^{d} f_{k,l,m}(v_k,v_l,v_m)+...+f_r(v_1,...,v_d)$$

Therefore, we truncate this expansion at e.g. second order

$$f(v_1,...,v_d)\simeq f_0+\sum_{k=1}^{d} f_k(v_k)+\sum_{k=1}^{d}\sum_{l=k+1}^{d} f_{k,l}(v_k,v_l)$$

Now, each $p-$variate functional component is determined up to functions of at most $p-1$ variables. Making this representation well-determined and unique again is less easy. We typically add hierarchical orthogonality constraints a priori that decompose the variance when the input variables v_i are mutually independent. That property makes constrained Hoeffding-Sobol expansions the basis of global factorial sensitivity analysis (e.g. Sobol indices) [23][25] and constrained HDMR a glass box of Machine Learning allowing to explain black-box ML algorithms such as kernel methods or decision trees [25].

Unfortunately, said scalar products are multiple integrals of dimension up to d [23][24][25][26]: the COD is back. To overcome it again, those integrals are approximated by Monte-Carlo methods like Random Sampling-HDMR [23][24]. But, since the required number of samples for those Monte-Carlo approximations to be sufficiently accurate may be huge, we finally introduce some functional basis to reduce it by going parametric [23][24]. So, starting from essentially nonparametric models to fight the COD, we end up with essentially parametric methods because unconstrained models are underdetermined and fitting constrained ones still suffers the COD. That's just one example and we

find plenty of algorithms and methods whose purpose is, after all, only to mitigate the COD due to said prior constraints.

4.3. Unconstrained underdetermined HDMR

Fortunately, thanks to SBI, those prior constraints become not only unnecessary but also undesirable, given that they can be added only a posteriori if ever required, for instance to compare several HDMR expansions each other.

We can work seamlessly with unconstrained, underdetermined models that always fit the data better than constrained ones. We just need to plug the HDMR expansion into the regularization penalty (2) to get the "maxentropic" partially improper Gaussian prior $\mathcal{N}\left(0,\mathbf{R}^{+}\right)$ for the stacked vector $\mathbf{f}$ of all HDMR components unknowns. $\mathbf{R}$ still is band diagonal. The likelihood quadratic form $\mathbf{D}$ becomes a block matrix with huge structural rank deficiency. Of course, the posterior matrix pencil $\left(\mathbf{D},\mathbf{R}\right)$ is always singular since the HDMR components remain underdetermined a posteriori. Then, we estimate one particular HDMR representation together with its posterior credible intervals by computing the marginal posterior moments of the proper posterior wrt the restricted posterior reference measure as described above, from which we can estimate any other particular representation we like.

5. Application and results: multi-product dynamic pricing

5.1. Multi-product dynamic pricing as singular Bayesian Optimization

SBI has been directly implemented, tested and validated in an application that is too sophisticated to be described in detail here: multi-product dynamic pricing.

Starting with past observed sales data over time t for a set of P potentially mutually dependent (i.e. complementary/halo or substitutable/cannibal) products or goods, including selling price vectors $\mathbf{p}_t$, sales volumes Q_t^i and numbers of potential and actual anonymous customers N_t^E and N_t^i, the goal is to maximize a black-box, expensive financial criterion like the total revenue or the total profit margin by Bayesian Optimization. We need to compute the probability distributions of two demand functions of the $P-$dimensional price vector $\mathbf{p}$ per product, the potential-to-actual customer conversion rate g_i and the sales volume per customer f_i, which are the mathematical expectations of a $\Gamma-$Poisson likelihood $\prod_{i=1}^{P}\prod_{t\in\Omega}\Gamma\left(Q_t^i\middle|N_t^i,f_i\left(\mathbf{p}_t,t\right)\right)\mathcal{P}\left(N_t^i\middle|N_t^E g_i\left(\mathbf{p}_t,t\right)\right)$.

From the posterior distributions for those demand functions, we get the posterior distributions for the sales volume Q, for the criterion per potential customer for each product and finally for the total criterion to be optimized. We use second-order $k=2$ partially improper Gaussian smoothing priors without any boundary condition because we need to extrapolate the functions on the whole price search intervals and, because the demand functions should go to 0^{+} when the selling prices go to $+\infty$, second-order underdetermined multiplicative, factorized HDMR instead of additive ones

$$h\left(\mathbf{p}\right)\simeq\prod_{i=1}^{P}h_i\left(p_i\right)\prod_{i=1}^{P}\prod_{j=i+1}^{P}h_{i,j}\left(p_i,p_j\right)$$

Hereafter, the demand functions are stationary, but they may depend on time. In this case, we can use combined partially improper spatiotemporal priors with total variation regularization $k=1$ over time for the stacked vector $\mathbf{f_t}$ of all unknowns over time

$$p\left(\mathbf{f}_t\middle|\varepsilon_p,\varepsilon_t\right)\propto\sqrt{\left|\varepsilon_p^2\mathbf{R}_p+\varepsilon_t^2\mathbf{R}_t\right|^{+}}\,e^{-\frac{1}{2}\mathbf{f}_t^{\mathrm{T}}\left(\varepsilon_p{}^2\mathbf{R}_p+\varepsilon_t{}^2\mathbf{R}_t\right)\mathbf{f}_t}$$

that yield extremely low-rank posterior precision matrices $\mathbf{D}_t+\varepsilon_p^2\mathbf{R}_p+\varepsilon_t^2\mathbf{R}_t$.

To validate the estimation of the HDMR functional components, they are just rescaled a posteriori by setting all their means to 1 but the univariate component depend-

ing on the product own price. Bayesian optimization is done by exhaustive search on an acquisition function like Predictive Entropy Search [29] to avoid local minima and to validate the highly singular estimation stage. All posterior calculations are done analytically using suitable transformations and approximations but the marginalization of the regularization hyperparameters that is done by simple one-dimension numerical vector integration.

Functional validation is done via computer simulations with ground truth: given demand functions with known optimal prices, we generate random past sales data. We compute the marginal posterior moments of the criterion to be optimized and we estimate the next optimal price vector $\mathbf{p}_1$ by Bayesian optimization of the acquisition function. Then, we generate new random sales data according to prices $\mathbf{p}_1$, estimate the next price vector $\mathbf{p}_2$ and we repeat the process until convergence or not towards the optimal prices.

5.2. Exprimental results

Figure 1 shows the results of a $7-$product dynamic pricing problem. 10 different selling prices per product. Starting from scarce (compared to the cardinal of the sampling space) and bad sales data with selling prices far away from the optimal prices set to 60€ (red), the total revenue per potential customer (bottom center) is maximized very quickly after a few iterations. The 10^7 parameters to be estimated for each of the 14 demand functions reduce to 2618 thanks to second-order multiplicative underdetermined HDMR. For instance, we have $\text{rank}(\mathbf{R}) = 2541$, $\text{rank}(\mathbf{D}) = 24$ and $\text{rank}(\mathbf{D},\mathbf{R}) = 2549$.

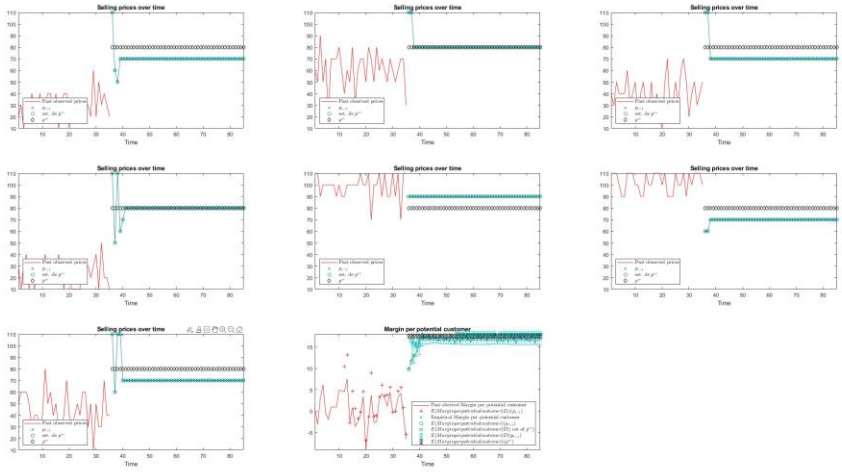

Figure 1. 7 -product dynamic pricing results. Subplots 1-7 show the past observed selling prices (red), the optimized prices (cyan) and the optimal prices (black) for each product over time. Subplot 8 (bottom center) shows the corresponding empirical past observed (red), optimized (cyan crosses) and theoretical optimal (black) total profit margin per potential customer over time.

Figure 2 shows the marginal posterior expectations and standard deviations for the univariate components of a second-order underdetermined multiplicative HDMR for both demand functions f_i and g_i for one product in a $3-$product dynamic pricing problem. All functions off the diagonal should be identically equal to 1 after rescaling because each demand function depends only on its own price. Probability matching looks satisfactory (i.e. $\simeq 68\%$ coverage at 1σ).

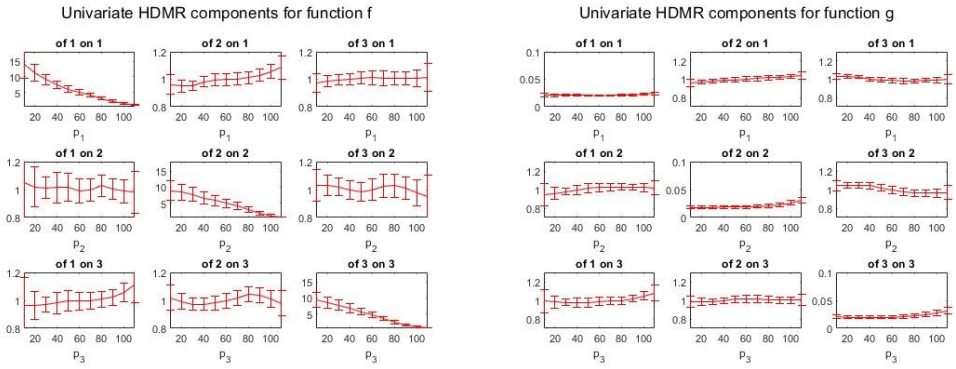

Figure 2. Univariate second-order underdetermined multiplicative HDMR components marginal posterior expectations (red) and standard deviations (error bars) for both demand functions of one product in a $3-$product dynamic pricing problem.

Figure 3 shows the marginal posterior expectations and standard deviations for the bivariate components of second-order underdetermined multiplicative HDMR for both demand functions for a given product in a $3-$product dynamic pricing problem. All functions should be identically equal to 1 after rescaling because each demand function depends only on its own price. Probability matching looks satisfactory as well.

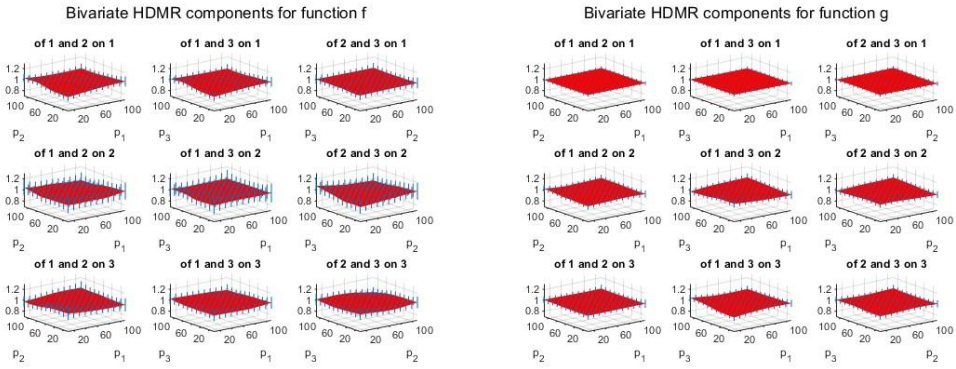

Figure 3. Bivariate second-order underdetermined multiplicative HDMR components marginal posterior expectations (red) and standard deviations (blue error bars) for both demand functions of one product in a $3-$product dynamic pricing problem.

6. Discussion and conclusion

While the need to estimate underdetermined quantities is not a matter of discussion, to the best of our knowledge it has never been done probabilistically. This requires a counterintuitive conceptual leap, that of working with probability distributions that do not possess any of the usual and indispensable properties. Fortunately, ill-posed inverse problems guide us by imposing such distributions, a priori and a posteriori. Probability distributions for underdetermined quantities are found to be partially improper and we only need to restrict the initial reference measure to their support to go back and forth from the world of infinitely many underdetermined solutions to that of one particular solution. In the end, the procedure is essentially the same as in the non-probabilistic case, i.e. all we have to do is replace the inverse by the Moore-Penrose pseudo-inverse and the determinant by the pseudo-determinant of the posterior precision matrix, at least in the Gaussian setting.

But SBI proves to be useful, if not indispensable, for perfectly well-determined problems as well. Until now, we used to fight the COD and to analyze high-dimensional functions and methods with additive or multiplicative, parametric or nonparametric representations. Although essentially underdetermined, they were made well-determined by adding prior constraints. Unfortunately, fitting well-determined models

with hierarchical orthogonality constrained models suffers the COD again. Thanks to SBI, we can work seamlessly with unconstrained, underdetermined models that always fit the data better than constrained ones with exponential speedup. We finally get much more simple, general, invariant and efficient $O(N)$ and down to $O(d^3)$ or even $O(d^2)$ algorithms with recursive low-rank updates, with user-supplied tradeoff between accuracy and computational complexity. Despite their heaviness, constrained HDMR/Hoeffding-Sobol/fANOVA/interactive spline models were already of considerable interest in inverse problems, global factorial sensitivity analysis and explainable Machine Learning. We can expect their unconstrained counterparts to be even more so.

Much remains to be done. First, we should formalize SBI rigorously as it deserves and generalize it as far as possible. Certain theoretical consequences are immediate: there must exist a singular theory of underdetermined information fields. Others are less straightforward and requires some investigations. What is the impact of SBI on Information Geometry? Is the Fisher-Rao metric in the space of multivariate Gaussian measures with positive semi-definite covariance or precision matrices well-defined? How to cope with the reference measure juggling on the way?

On the practical side, today the most popular meta-/surrogate model for Bayesian/parsimonious Optimization is GP functional regression. But it is not Bayesian, not invariant a priori modulo polynomials with positive definite kernels, does not propagate uncertainty and has $O(N^3)$ generic computational complexity instead of $O(N)$ for truly Bayesian nonparametric functional regression as soon as the likelihood factorizes. Despite these flaws, it is popular thanks to its intrinsic immunity to the COD (why?) and perhaps due to the lack of efficient and user-friendly nonparametric Bayesian algorithms in large dimension. The situation could change with truly Bayesian functional regression combined with underdetermined models to fight the COD that is basically expected to outperform GP functional regression on all quality and efficiency evaluation criteria but perhaps the scaling in the number of variables. Subsequently, truly (singular) Bayesian functional estimation-based Bayesian Optimization is expected to outperform GP functional regression-based "Bayesian" Optimization, especially if it eases the evaluation and optimization of acquisition functions like Predictive Entropy Search.

The next most obvious application of SBI would be to generalize Bayesian Numerical Linear Algebra to indeterminate systems of linear equations but we cannot help but think of the possibility of building AI deep networks with unconstrained underdetermined nonparametric additive models that are linear in the parameters, in the same vein as Kolmogorov-Arnold networks [30].

Anyway, there is a nice singular and underdetermined world to explore.

Funding: Not applicable.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not yet publicly available.

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

References

1. Pautot F., Theoretical basis of in vivo tomographic tracer kinetics, Part I: On tracers that conserve their mass, https://www.researchgate.net/publication/303020662_Theoretical_basis_of_in_vivo_tomographic_tracer_kinetics_Part_I_On_tracers_that_conserve_their_mass
2. JCGM 100:2008, Evaluation of measurement data – Guide to the expression of uncertainty in measurement, DOI: https://doi.org/10.59161/JCGM100-2008E
3. Hennig Ph., Probabilistic Interpretation of Linear Solvers, https://doi.org/10.48550/arXiv.1402.2058
4. Poincaré H., La science et l'hypothèse, Chapter XI, 1902, https://gallica.bnf.fr/ark:/12148/bpt6k26745q/f1.item
5. Wahba G., Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression, J. R. Statist. Soc. B (1978) **40**, No.3, pp. 364-372
6. Wahba G., Spline Models for Observational Data, 1990, ISBN 978-0898712445
7. Rasmussen C.E., Gaussian Processes for Machine Learning, 2006, ISBN 026218253X, https://gaussianprocess.org/gpml/
8. MacKay D. J. C., Introduction to Gaussian Processes, https://www.inference.org.uk/mackay/gpB.pdf
9. Enßlin T. A., Information Field Theory, 2013, arXiv:1301.2556
10. Bretthorst G. L., Bayesian Interpolation and Deconvolution, 1992, Technical Report CR-RD-AS-92-4
11. Boutelier T., Kudo K., Pautot F., Sasaki M., Bayesian hemodynamic parameter estimation by bolus tracking perfusion weighted imaging, IEEE Trans Med Imaging, 2012 Jul; 31(7):1381-95, doi: 10.1109/TMI.2012.2189890, Epub 2012 Mar 6.
12. Raghavan N., Cox D. D., Analysis of the Posterior for Spline Estimators in Logistic Regression, Journal of Statistical Planning and Inference, Vol. 71, Issues 1-2, August 1998, pp. 117-136, https://doi.org/10.1016/S0378-3758(98)00085-8
13. Green P. J., Silverman B.W., Nonparametric Regression and Generalized Linear Models, A roughness penalty approach, Chapman & Hall, 1993, ISBN 0-412-30040-0
14. Scheipl F., Kneib Th., Fahrmeir L., Penalized Likelihood and Bayesian Function Selection in Regression Models, Advances in Statistical Analysis, Vol. 97, pp. 349-385, 2013, https://doi.org/10.1007/s10182-013-0211-3
15. Yue Y. R., Speckman P. L., Sun D., Priors for Bayesian Adaptive Spline Smoothing, Ann Inst Stat Math (2012) 64:577–613
16. Speckman P. L., Sun D., Fully Bayesian Spline Smoothing and Intrinsic Autoregressive Priors, Biometrika (2003), Vol. 90, 2, pp. 289-302
17. Finite difference coefficients, https://en.wikipedia.org/wiki/Finite_difference_coefficient
18. Parlett B. N., Symmetrix matrix pencils, Journal of Computational and Applied Mathematics, 38 (1991) 373-385
19. Anderson T. W., An Introduction to Multivariate Statistical Analysis Third Edition, Wiley & Sons, 2003, ISBN 0-471-36091
20. Rao C. R., Linear Statistical Inference and Its Applications, second edition, John Wiley & Sons, 1973, ISBN 9780471708230
21. Garcke J., Sparse Grids in a Nutshell, In: Garcke, J., Griebel, M. (eds) Sparse Grids and Applications. Lecture Notes in Computational Science and Engineering, vol 88. Springer, Berlin, Heidelberg, 2012. https://doi.org/10.1007/978-3-642-31703-3_3
22. Hastie T. J., Tibshirani R. J., Generalized Additive Models, 1990, ISBN 0-412-34390-8
23. Rabitz H., Alis Ö. F. General foundations for high-dimensional model representations, Journal of Mathematical Chemistry, Vol. 25, 197-233, 1999
24. Alis Ö. F., Rabitz H., Efficient implementation of high dimensional model representations, Journal of Mathematical Chemistry, 2001, Volume 29, page 127-142
25. Bastian C. D., Rabitz H., High Dimensional Model Representation as a Glass Box in supervised Machine Learning, 2018, arXiv:1807.10320
26. Chastaing G., Gamboa F., Prieur C., Generalized Hoeffding-Sobol Decomposition for Dependent Variables - Application to Sensitivity Analysis, Electronic Journal of Statistics, Vol. 0 (0000), ISSN: 1935-7524, DOI 10.1214/154957804100000000, https://arxiv.org/pdf/1112.1788
27. Gu Ch., Smoothing Spline ANOVA models, Springer, 2nd edition, 2013, ISBN 1461453682
28. Wahba G., Partial and Interaction Splines Models for the Semiparametric Estimation of Functions of Several Variables, Colorado State Univ., Computer Science and Statistics. Proceedings of the 18th Symposium on the Interface, 1986, https://ntrs.nasa.gov/citations/19890004538
29. Hernandez-Lobato J. M., Hoffman M. W., Ghahramani Z., Predictive Entropy Search for Efficient Global Optimization of Black-box Functions, arXiv:1406.2541v1 [stat.ML] 10 June 2014, https://doi.org/10.48550/arXiv.1406.2541
30. Liu Z. et al, KAN : Kolmogorov-Arnold Networks, arXiv:2404.19756v1 [cs.LG] 30 Apr 2024, https://doi.org/10.48550/arXiv.2404.19756