

# Optimal projection for parametric importance sampling in high dimensions

Maxime El Masri  ONERA/DTIS, ISAE-SUPAERO, Université de Toulouse

Jérôme Morio  ONERA/DTIS, Université de Toulouse

Florian Simatos ISAE-SUPAERO, Université de Toulouse

Date published: 2024-11-03 Last modified: 2024-03-26

## Abstract

We propose a dimension reduction strategy in order to improve the performance of importance sampling in high dimensions. The idea is to estimate variance terms in a small number of suitably chosen directions. We first prove that the optimal directions, i.e., the ones that minimize the Kullback–Leibler divergence with the optimal auxiliary density, are the eigenvectors associated with extreme (small or large) eigenvalues of the optimal covariance matrix. We then perform extensive numerical experiments showing that as dimension increases, these directions give estimations which are very close to optimal. Moreover, we demonstrate that the estimation remains accurate even when a simple empirical estimator of the covariance matrix is used to compute these directions. The theoretical and numerical results open the way for different generalizations, in particular the incorporation of such ideas in adaptive importance sampling schemes.

*Keywords:* Rare event simulation, Parameter estimation, Importance sampling, Dimension reduction, Kullback–Leibler divergence, Projection

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Importance Sampling</b>	<b>3</b>
<b>3</b>	<b>Efficient dimension reduction</b>	<b>4</b>
3.1	Projecting onto a low-dimensional subspace . . . . .	4
3.2	Definition of the function $\ell$ . . . . .	5
3.3	Main result of the paper . . . . .	5
3.4	Choice of the number of dimensions $k$ . . . . .	7
3.5	Theoretical result concerning the projection on $\mathbf{m}^*$ . . . . .	7
<b>4</b>	<b>Computational framework</b>	<b>8</b>
4.1	Numerical procedure for IS estimate comparison . . . . .	8
4.2	Choice of the auxiliary density $g'$ for the Gaussian model . . . . .	9

<b>5</b>	<b>Numerical results on five test cases</b>	<b>10</b>
5.1	Test case 1: one-dimensional optimal projection . . . . .	10
5.1.1	Evolution of the partial KL divergence and spectrum . . . . .	11
5.1.2	Numerical results . . . . .	11
5.2	Test case 2: projection in 2 directions . . . . .	13
5.2.1	Evolution of the partial KL divergence and spectrum . . . . .	13
<b>References</b>		<b>13</b>

## 1 Introduction

Importance Sampling (IS) is a stochastic method to estimate integrals of the form  $\mathcal{E} = \int \phi(\mathbf{x})f(\mathbf{x})d\mathbf{x}$  with a black-box function  $\phi$  and a probability density function (pdf)  $f$ . It rests upon the choice of an auxiliary density which can significantly improve the estimation compared to the naive Monte Carlo (MC) method (Agapiou et al. 2017), (Owen and Zhou 2000). The theoretical optimal IS density, also called zero-variance density, is defined by  $\phi f / \mathcal{E}$  when  $\phi$  is a positive function. This density is not available in practice as it involves the unknown integral  $\mathcal{E}$ , but a classical strategy consists in searching for an optimal approximation in a parametric family of densities. By minimising a “distance” to the optimal IS density, such as the Kullback–Leibler divergence, one can find optimal parameters in this family to get an efficient sampling pdf. Adaptive Importance Sampling (AIS) algorithms, such as the Mixture Population Monte Carlo method (Cappé et al. 2008), the Adaptive Multiple Importance Sampling method (Cornuet et al. 2012), or the Cross Entropy method (Rubinstein and Kroese 2011a), estimate the optimal parameters adaptively by updating at intermediate levels (Bugallo et al. 2017).

These techniques work very well, but only for moderate dimensions. In high dimensions, most of these techniques fail to give suitable parameters for two reasons:

1. the weight degeneracy problem, for which the self-normalized likelihood ratios (weights) in the IS densities degenerate in the sense that the largest one takes all the mass, while all other weights are negligible so that the final estimation essentially uses only one sample. See for instance (Bengtsson, Bickel, and Li 2008) for a theoretical analysis in the related context of particle filtering. The conditions under which importance sampling is applicable in high dimensions are notably investigated in a reliability context in (Au and Beck 2003): it is remarked that the optimal covariance matrix should not deviate significantly from the identity matrix. (El-Laham, Elvira, and Bugallo 2019) tackle the weight degeneracy problem by applying a recursive shrinkage of the covariance matrix, which is constructed iteratively with a weighted sum of the sample covariance estimator and a biased, but more stable, estimator;
2. the intricate estimation of distribution parameters in high dimensions and particularly covariance matrices, whose size increases quadratically in the dimension (Ashurbekova et al. 2020),(Ledoit and Wolf 2004). Empirical covariance matrix estimate has notably a slow convergence rate in high dimensions (Fan, Fan, and Lv 2008). For that purpose, dimension reduction techniques can be applied. The idea was recently put forth to reduce the effective dimension by only estimating these parameters (in particular the covariance matrix) in suitable directions (El Masri, Morio, and Simatos 2021), (Uribe et al. 2021). In this paper we delve deeper into this idea.

The main contribution of the present paper is to identify the optimal directions in the fundamental case when the parametric family is Gaussian, and perform numerical simulations in order to understand how they behave in practice. In particular, we propose directions which, in contrast to the recent paper (Uribe et al. 2021), do not require the objective function to be differentiable, and moreover

optimizes the Kullback–Leibler distance with the optimal density instead of simply an upper bound on it, as in (Uribe et al. 2021). In Section 3.1 we elaborate in more details on the differences between the two approaches.

The paper is organised as follows: in Section 2 we recall the foundations of IS. In Section 3, we state our main theoretical result and we compare it with the current state-of-the-art. The proof of our theoretical result are given in Appendix; Section 4 introduces the numerical framework that we have adopted, and Section 5 presents the numerical results obtained on five different test cases to assess the efficiency of the directions that we propose. We conclude in ?@sec-Ccl with a summary and research perspectives.

## 2 Importance Sampling

We consider the problem of estimating the following integral:

$$\mathcal{E} = \mathbb{E}_f(\phi(\mathbf{X})) = \int \phi(\mathbf{x})f(\mathbf{x})d\mathbf{x},$$

where  $\mathbf{X}$  is a random vector in  $\mathbb{R}^n$  with standard Gaussian pdf  $f$ , and  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}_+$  is a real-valued, non-negative function. The function  $\phi$  is considered as a black-box function which is potentially expensive to evaluate, and this means that the number of calls to  $\phi$  should be limited.

IS is an approach used to reduce the variance of the classical Monte Carlo estimator of  $\mathcal{E}$ . The idea of IS is to generate a random sample  $\mathbf{X}_1, \dots, \mathbf{X}_N$  from an auxiliary density  $g$ , instead of  $f$ , and to compute the following estimator:

$$\widehat{\mathcal{E}}_N = \frac{1}{N} \sum_{i=1}^N \phi(\mathbf{X}_i)L(\mathbf{X}_i), \quad (1)$$

with  $L = f/g$  the likelihood ratio, or importance weight, and the auxiliary density  $g$ , also called importance sampling density, is such that  $g(\mathbf{x}) = 0$  implies  $\phi(\mathbf{x})f(\mathbf{x}) = 0$  for every  $\mathbf{x}$  (which makes the product  $\phi L$  well-defined). This estimator is consistent and unbiased but its accuracy strongly depends on the choice of the auxiliary density  $g$ . It is well known that the optimal choice for  $g$  is (Bucklew 2013)

$$g^*(\mathbf{x}) = \frac{\phi(\mathbf{x})f(\mathbf{x})}{\mathcal{E}}, \quad \mathbf{x} \in \mathbb{R}^n.$$

Indeed, for this choice we have  $\phi L = \mathcal{E}$  and so  $\widehat{\mathcal{E}}_N$  is actually the deterministic estimator  $\mathcal{E}$ . For this reason,  $g^*$  is sometimes called zero-variance density, a terminology that we will adopt here. Of course,  $g^*$  is only of theoretical interest as it depends on the unknown integral  $\mathcal{E}$ . However, it gives an idea of good choices for the auxiliary density  $g$ , and we will seek to approximate  $g^*$  by an auxiliary density that minimizes a distance between  $g^*$  and a given parametric family of densities.

In this paper, the parametric family of densities is the Gaussian family  $\{g_{\mathbf{m}} : \mathbf{m} \in \mathbb{R}^n, \in \mathcal{S}_n^+\}$ , where  $g_{\mathbf{m}}$  denotes the Gaussian density with mean  $\mathbf{m} \in \mathbb{R}^n$  and covariance matrix  $\in \mathcal{S}_n^+$  with  $\mathcal{S}_n^+ \subset \mathbb{R}^{n \times n}$  the set of symmetric, positive-definite matrices:

$$g_{\mathbf{m}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} \|\cdot\|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{m})^\top \cdot^{-1}(\mathbf{x} - \mathbf{m})\right), \quad \mathbf{x} \in \mathbb{R}^n.$$

with  $\|\cdot\|$  the determinant of  $\cdot$ . Moreover, we will consider the Kullback–Leibler (KL) divergence to measure a “distance” between  $g^*$  and  $g_{\mathbf{m}}$ . Recall that for two densities  $f$  and  $h$ , with  $f$  absolutely continuous with respect to  $h$ , the KL divergence  $D(f, h)$  between  $f$  and  $h$  is defined by:

$$D(f, h) = \mathbb{E}_f\left[\log\left(\frac{f(\mathbf{X})}{h(\mathbf{X})}\right)\right] = \int \log\left(\frac{f(\mathbf{x})}{h(\mathbf{x})}\right) f(\mathbf{x})d\mathbf{x}.$$

Thus, our goal is to approximate  $g^*$  by  $g_{\mathbf{m}^*, \Sigma^*}$  with the optimal mean vector  $\mathbf{m}^*$  and the optimal covariance matrix  $\Sigma^*$  given by:

$$(\mathbf{m}^*, \Sigma^*) = \arg \min \{D(g^*, g_{\mathbf{m}}) : \mathbf{m} \in \mathbb{R}^n, \Sigma \in \mathcal{S}_n^+\}. \quad (2)$$

This optimization is in general convex and differentiable with respect to  $\mathbf{m}$  and  $\Sigma$ . Moreover, the solution of Equation 2 can be computed analytically by cancelling the gradient. In the Gaussian case, it is thus proved that  $\mathbf{m}^*$  and  $\Sigma^*$  are simply the mean and variance of the zero-variance density (Rubinstein and Kroese 2011b), (Rubinstein and Kroese 2017):

$$\mathbf{m}^* = \mathbb{E}_{g^*}(\mathbf{X}) \quad \text{and} \quad \Sigma^* = \text{Var}_{g^*}(\mathbf{X}). \quad (3)$$

### 3 Efficient dimension reduction

#### 3.1 Projecting onto a low-dimensional subspace

As  $g^*$  is unknown, the optimal parameters  $\mathbf{m}^*$  and  $\Sigma^*$  given by Equation 3 are not directly computable. However, we can sample from the optimal density as it is known up to a multiplicative constant. Therefore, usual estimation schemes start with estimating  $\mathbf{m}^*$  and  $\Sigma^*$ , say through  $\hat{\mathbf{m}}^*$  and  $\hat{\Sigma}^*$ , respectively, and then use these approximations to estimate  $\mathcal{E}$  through Equation 1 with the auxiliary density  $g_{\hat{\mathbf{m}}^*, \hat{\Sigma}^*}$ . Although the estimation of  $\mathcal{E}$  with the auxiliary density  $g_{\mathbf{m}^*, \Sigma^*}$  usually provides very good results, it is well-known that in high dimensions, the additional error induced by the estimations of  $\mathbf{m}^*$  and  $\Sigma^*$  severely degrades the accuracy of the final estimation (Papaioannou, Geyer, and Straub 2019), (Uribe et al. 2021). The main problem lies in the estimation of  $\Sigma^*$  which, in dimension  $n$ , involves the estimation of a quadratic (in the dimension) number of terms, namely  $n(n+1)/2$ . Recently, the idea to overcome this problem by only evaluating variance terms in a small number of influential directions was explored in (El Masri, Morio, and Simatos 2021) and (Uribe et al. 2021). In these two papers, the auxiliary covariance matrix is modeled in the form

$$= \sum_{i=1}^k (\nu_i - 1) \mathbf{d}_i \mathbf{d}_i^\top + I_n \quad (4)$$

where the  $\mathbf{d}_i$ 's are the  $k$  orthonormal directions which are deemed influential. It is easy to check that is the covariance matrix of the Gaussian vector

$$\nu_1^{1/2} Y_1 \mathbf{d}_1 + \dots + \nu_k^{1/2} Y_k \mathbf{d}_k + Y_{k+1} \mathbf{d}_{k+1} + \dots + Y_n \mathbf{d}_n$$

where the  $Y_i$ 's are i.i.d. standard normal random variables (one-dimensional), and the  $n-k$  vectors  $(\mathbf{d}_{k+1}, \dots, \mathbf{d}_n)$  complete  $(\mathbf{d}_1, \dots, \mathbf{d}_k)$  into an orthonormal basis. In particular,  $\nu_i$  is the variance in the direction of  $\mathbf{d}_i$ , i.e.,  $\nu_i = \mathbf{d}_i^\top \Sigma \mathbf{d}_i$ . In Equation 4,  $k$  can be considered as the effective dimension in which variance terms are estimated. In other words, in (El Masri, Morio, and Simatos 2021) and (Uribe et al. 2021), the optimal variance parameter is not sought in  $\mathcal{S}_n^+$  as in Equation 2, but rather in the subset of matrices of the form

$$\mathcal{L}_{n,k} = \left\{ \sum_{i=1}^k (\alpha_i - 1) \frac{\mathbf{d}_i \mathbf{d}_i^\top}{\|\mathbf{d}_i\|^2} + I_n : \alpha_1, \dots, \alpha_k > 0 \text{ and the } \mathbf{d}_i \text{'s are orthogonal} \right\}.$$

The relevant minimization problem thus becomes

$$(\mathbf{m}_{k,k}^*, \Sigma_{k,k}^*) = \arg \min \{D(g^*, g_{\mathbf{m}}) : \mathbf{m} \in \mathbb{R}^n, \Sigma \in \mathcal{L}_{n,k}\} \quad (5)$$

instead of Equation 2, with the effective dimension  $k$  being allowed to be adjusted dynamically. By restricting the space in which the variance is assessed, one seeks to limit the number of variance

terms to be estimated. The idea is that if the directions are suitably chosen, then the improvement of the accuracy due to the smaller error in estimating the variance terms will compensate the fact that we consider less candidates for the covariance matrix. In (El Masri, Morio, and Simatos 2021), the authors consider  $k = 1$  and  $\mathbf{d}_1 = \mathbf{m}^* / \|\mathbf{m}^*\|$ . When  $f$  is Gaussian, this choice is motivated by the fact that, due to the light tail of the Gaussian random variable and the reliability context, the variance should vary significantly in the direction of  $\mathbf{m}^*$  and so estimating the variance in this direction can bring information. In Section 3.5, we use the techniques of the present paper to provide a stronger theoretical justification of this choice, see Theorem 3.2 and the discussion following it. The method in (Uribe et al. 2021) is more involved:  $k$  is adjusted dynamically, while the directions  $\mathbf{d}_i$  are the eigenvectors associated to the largest eigenvalues of a certain matrix. They span a low-dimensional subspace called Failure-Informed Subspace, and the authors in (Uribe et al. 2021) prove that this choice minimizes an upper bound on the minimal KL divergence. In practice, this algorithm yields very accurate results. However, we will not consider it further in the present paper for two reasons. First, this algorithm is tailored for the reliability case where  $\phi = \mathbb{I}_{\{\varphi \geq 0\}}$ , with a function  $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$ , whereas our method is more general and applies to the general problem of estimating an integral (see for instance our test case of `?@sec-sub:payoff`). Second, the algorithm in (Uribe et al. 2021) requires the evaluation of the gradient of the function  $\varphi$ . However, this gradient is not always known and can be expensive to evaluate in high dimensions; in some cases, the function  $\varphi$  is even not differentiable, as will be the case in our numerical example in `?@sec-sub:portfolio`. In contrast, our method makes no assumption on the form or smoothness of  $\phi$ : it does not need to assume that it is of the form  $\mathbb{I}_{\{\varphi \geq 0\}}$ , or to assume that  $\nabla \varphi$  is tractable. For completeness, whenever the algorithm of (Uribe et al. 2021) was applicable and computing the gradient of  $\varphi$  did not require any additional simulation budget, we have run it on the test cases considered here and found that it outperformed our algorithm. In more realistic settings, computing  $\nabla \varphi$  would likely increase the simulation budget, and it would be interesting to compare the two algorithms in more details to understand when this extra computation cost is worthwhile. We reserve such a question for future research and will not consider the algorithm of (Uribe et al. 2021) further, as our aim in this paper is to establish benchmark results for a general algorithm which works for any function  $\phi$ .

### 3.2 Definition of the function $\ell$

The statement of our result involves the following function  $\ell$ , which is represented in Figure 1:

$$\ell : x \in (0, \infty) \mapsto -\log(x) + x - 1. \quad (6)$$

In the following,  $(\lambda, \mathbf{d}) \in \mathbb{R} \times \mathbb{R}^n$  is an eigenpair of a matrix  $A$  if  $A\mathbf{d} = \lambda\mathbf{d}$  and  $\|\mathbf{d}\| = 1$ . A diagonalizable matrix has  $n$  distinct eigenpairs, say  $((\lambda_i, \mathbf{d}_i), i = 1, \dots, n)$ , and we say that these eigenpairs are ranked in decreasing  $\ell$ -order if  $\ell(\lambda_1) \geq \dots \geq \ell(\lambda_n)$ . In the rest of the article, we denote as  $(\lambda_i^*, \mathbf{d}_i^*)$  the eigenpairs of  $^*$  ranked in decreasing  $\ell$ -order and as  $(\hat{\lambda}_i^*, \hat{\mathbf{d}}_i^*)$  the eigenpairs of  $^*$  ranked in decreasing  $\ell$ -order.

### 3.3 Main result of the paper

The main result of the present paper is to compute the exact value for  $_k^*$  in Equation 5, which therefore paves the way for efficient high-dimensional estimation schemes.

**Theorem 3.1.** *Let  $(\lambda_i^*, \mathbf{d}_i^*)$  be the eigenpairs of  $^*$  ranked in decreasing  $\ell$ -order. Then for  $1 \leq k \leq n$ , the solution  $(\mathbf{m}_{k,^*}^*, _k^*)$  to Equation 5 is given by*

$$\mathbf{m}_k^* = \mathbf{m}^* \text{ and } _k^* = I_n + \sum_{i=1}^k (\lambda_i^* - 1) \mathbf{d}_i^* (\mathbf{d}_i^*)^\top. \quad (7)$$

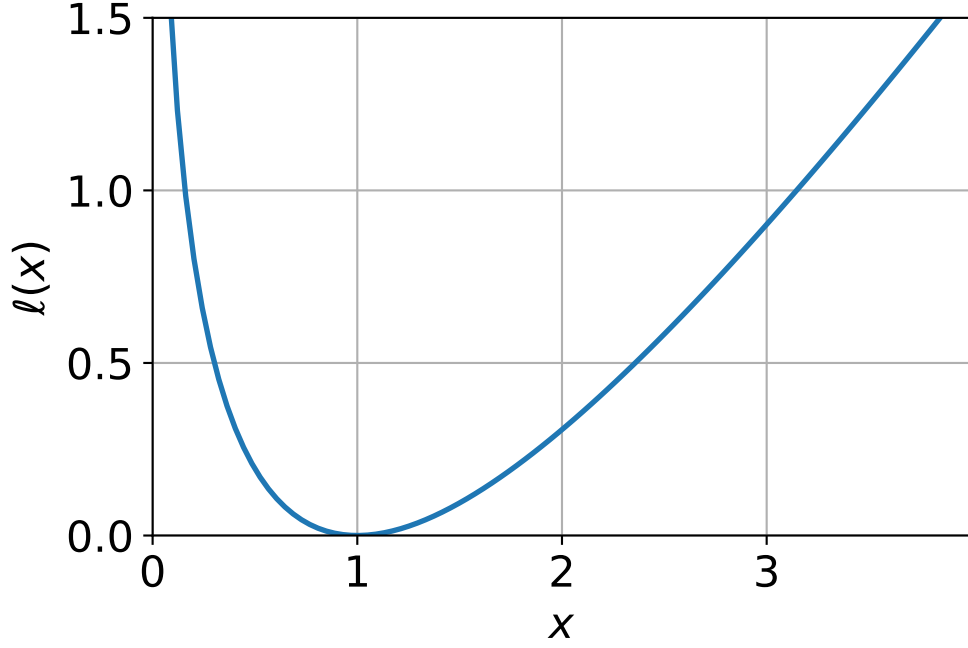


Figure 1: Plot of the function  $\ell$  given by Equation 6.

The proof of Theorem 3.1 is detailed in Appendix A. For  $k = 1$  for instance, the matrix  $\hat{\Sigma}_1^* = I_n + (\lambda_1^* - 1)\mathbf{d}_1^*(\mathbf{d}_1^*)^\top$  with  $(\lambda_1^*, \mathbf{d}_1^*)$  the eigenpair of  $\Sigma^*$  such as  $\lambda_1^*$  is either the largest or the smallest eigenvalue of  $\Sigma^*$ , depending on which one maximizes  $\ell$ .

This theoretical result therefore suggests to reduce dimension by computing the covariance matrix  $\hat{\Sigma}^*$  and its eigenpairs, rank them in decreasing  $\ell$ -order and then use the  $k$  first eigenpairs  $((\hat{\lambda}_i^*, \hat{\mathbf{d}}_i^*), i = 1, \dots, k)$  to build the covariance matrix  $\hat{\Sigma}_k^* = \sum_{i=1}^k (\hat{\lambda}_i^* - 1)\hat{\mathbf{d}}_i^*(\hat{\mathbf{d}}_i^*)^\top + I_n$  and the corresponding auxiliary density. This scheme is summarized in Algorithm 1. The effective dimension  $k$  is obtained by Algorithm 2, see Section 3.4 below. The proof of the theorem is shown in Appendix A.

---

**Algorithm 1** Algorithm suggested by Theorem 1.

---

- 1: **Data:** Sample sizes  $N$  and  $M$
  - 2: **Result:** Estimation  $\hat{\mathcal{E}}_N$  of integral  $\mathcal{E}$
  - 3: - Generate a sample  $\mathbf{X}_1^*, \dots, \mathbf{X}_M^*$  on  $\mathbb{R}^n$  independently according to  $g^*$
  - 4: - Estimate  $\hat{\mathbf{m}}^*$  and  $\hat{\Sigma}^*$  defined in Equation 8 and Equation 9 with this sample
  - 5: - Compute the eigenpairs  $(\hat{\lambda}_i^*, \hat{\mathbf{d}}_i^*)$  of  $\hat{\Sigma}^*$  ranked in decreasing  $\ell$ -order
  - 6: - Compute the matrix  $\hat{\Sigma}_k^* = \sum_{i=1}^k (\hat{\lambda}_i^* - 1)\hat{\mathbf{d}}_i^*(\hat{\mathbf{d}}_i^*)^\top + I_n$  with  $k$  obtained by applying Algorithm 2 with input  $(\hat{\lambda}_1^*, \dots, \hat{\lambda}_n^*)$
  - 7: - Generate a new sample  $\mathbf{X}_1, \dots, \mathbf{X}_N$  independently from  $g' = g_{\hat{\Sigma}_k^*}^*$
  - 8: - Return  $\hat{\mathcal{E}}_N = \frac{1}{N} \sum_{i=1}^N \phi(\mathbf{X}_i) \frac{f(\mathbf{X}_i)}{g'(\mathbf{X}_i)}$
- 

*Remark.* Since the function  $\ell$  is minimized at 1, eigenpairs with  $\lambda_i^* = 1$  are selected in the sum of Equation 7 once all other eigenpairs have been picked as the eigenpairs are  $\ell$ -ordered: in other words, if  $\lambda_i^* = 1$  then  $\lambda_j^* = 1$  for all  $j \geq i$ . Note also that the minimizer 1 plays a special role as we are interested in covariance matrices of  $\mathcal{L}_{n,k}$  which, once diagonalized, have mostly ones in the main diagonal (except for  $k$  values associated with the  $\alpha_i$ ). As  $k$  will be small (See Section 3.4), typically

$k = 1$  or  $2$ , this amounts to finding covariance matrices that are perturbations of the identity (this is relevant as we assume  $f$  is standard Gaussian). Therefore, when approximating  $g^*$  by such matrices, we should first consider eigenvalues as different as possible from 1 (with the discrepancy from 1 being measured by  $\ell$ ).

In the first step of Algorithm 1, we assume  $g^*$  can be sampled independently. This is a reasonable assumption as classical techniques such as importance sampling with self-normalized weights or Markov Chain Monte Carlo (MCMC) can be applied in this case (see for instance (Chan and Kroese 2012), (Grace, Kroese, and Sandmann 2014)). In this paper, we choose to apply a basic rejection method that yields perfect independent samples from  $g^*$ , possibly at the price of a high computational cost. As the primary goal of this paper is to understand whether the  $\mathbf{d}_i^*$ 's are indeed good projection directions, this cost will not be taken into account. Possible improvements to relax this assumption are discussed in the conclusion of the paper and in Appendix C.

### 3.4 Choice of the number of dimensions $k$

The choice of the effective dimension  $k$ , i.e., the number of projection directions considered, is important. If it is close to  $n$ , then the matrix  $\hat{\Sigma}_k^*$  will be close to  $\hat{\Sigma}^*$  which is the situation we want to avoid in the first place. On the other hand, setting  $k = 1$  in all cases may be too simple and lead to suboptimal results. In practice, however this is often a good choice. In order to adapt  $k$  dynamically, we consider a simple method based on the value of the KL divergence. Given the eigenvalues  $\lambda_1, \dots, \lambda_n$  ranked in decreasing  $\ell$ -order, we look for the maximal gap between two consecutive eigenvalues of the sequence  $(\ell(\lambda_1), \dots, \ell(\lambda_n))$ . This allows to choose  $k$  such that  $\sum_{i=1}^k \ell(\lambda_i)$  is close to  $\sum_{i=1}^n \ell(\lambda_i)$  which is equal, up to an additive constant, to the minimal KL divergence (shown in [@lem-D](#)). The precise method is described in Algorithm 2.

---

#### Algorithm 2 Choice of the number of dimensions

---

- 1: **Data:** Sequence of positive numbers  $\lambda_1, \dots, \lambda_n$  in decreasing  $\ell$ -order
  - 2: **Result:** Number of selected dimensions  $k$
  - 3: - Compute the increments  $\delta_i = \ell(\lambda_{i+1}) - \ell(\lambda_i)$  for  $i = 1 \dots n - 1$
  - 4: - Return  $k = \arg \max \delta_i$ , the index of the maximum of the differences.
- 

### 3.5 Theoretical result concerning the projection on $\mathbf{m}^*$

In (El Masri, Morio, and Simatos 2021), the authors propose to project on the mean  $\mathbf{m}^*$  of the optimal auxiliary density  $g^*$ . Numerically, this algorithm is shown to perform well, but only a very heuristic explanation based on the light tail of the Gaussian distribution is provided to motivate this choice. It turns out that the techniques used in the proof of Theorem 3.1 can shed light on why projecting on  $\mathbf{m}^*$  may indeed be a good idea. Let us first state our theoretical result, and then explain why it justifies the idea of projecting on  $\mathbf{m}^*$ .

**Theorem 3.2.** Consider  $\Sigma \in \mathcal{L}_{n,1}$  of the form  $\Sigma = I_n + (\alpha - 1)\mathbf{d}\mathbf{d}^\top$  with  $\alpha > 0$  and  $\|\mathbf{d}\| = 1$ . Then the minimizer in  $(\alpha, \mathbf{d})$  of the KL divergence between  $f$  and  $g_{\Sigma^*}$ , is  $(1 + \|\mathbf{m}^*\|^2, \mathbf{m}^*/\|\mathbf{m}^*\|)$ :

$$(1 + \|\mathbf{m}^*\|^2, \mathbf{m}^*/\|\mathbf{m}^*\|) = \arg \min_{\alpha, \mathbf{d}} \left\{ D(f, g_{\Sigma^*, I_n + (\alpha - 1)\mathbf{d}\mathbf{d}^\top}) : \alpha > 0, \|\mathbf{d}\| = 1 \right\}.$$

The proof of Theorem 3.2 is detailed in Appendix A. In other words,  $\mathbf{m}^*$  appears as an optimal projection direction when one seeks to minimize the KL divergence between  $f$  and the Gaussian density with mean  $\mathbf{m}^*$  and covariance of the form  $I_n + (\alpha - 1)\mathbf{d}\mathbf{d}^\top$ . Let us now explain why this minimization problem is indeed relevant, and why choosing an auxiliary density which minimizes



this KL divergence may indeed lead to an accurate estimation. The justification deeply relies on the recent results by (Chatterjee and Diaconis 2018).

As mentioned above, in a reliability context where one seeks to estimate a small probability  $p = \mathbb{P}(\mathbf{X} \in A)$ , Theorem 1.3 in (Chatterjee and Diaconis 2018) shows that  $D(g^*, g)$  governs the sample size required for an accurate estimation of  $p$ : more precisely, the estimation is accurate if the sample size is larger than  $e^{D(g^*, g)}$ , and inaccurate otherwise. This motivates the rationale for minimizing the KL divergence with  $g^*$ .

However, in high dimensions, importance sampling is known to fail because of the weight degeneracy problem whereby  $\max_i L_i / \sum_i L_i \approx 1$ , with the  $L_i$ 's the unnormalized importance weights, or likelihood ratios:  $L_i = f(\mathbf{X}_i)/g(\mathbf{X}_i)$  with the  $\mathbf{X}_i$ 's i.i.d. drawn according to  $g$ . Theorem 2.3 in (Chatterjee and Diaconis 2018) shows that the weight degeneracy problem is avoided if the empirical mean of the likelihood ratios is close to 1, and for this, Theorem 1.1 in (Chatterjee and Diaconis 2018) shows that the sample size should be larger than  $e^{D(f, g)}$ . In other words, these results suggest that the KL divergence with  $g^*$  governs the sample size for an accurate estimation of  $p$ , while the KL divergence with  $f$  governs the weight degeneracy problem.

In light of these results, it becomes natural to consider the KL divergence with  $f$  and not only  $g^*$  (Owen and Zhou 2000). Of course, minimizing  $D(f, g_{\mathbf{m}})$  without constraints on  $\mathbf{m}$  and is trivial since  $g_{\mathbf{m}} = f$  for  $\mathbf{m} = 0$  and  $= I_n$ . However, these choices are the ones we want to avoid in the first place, and so it makes sense to impose some constraints on  $\mathbf{m}$  and . If one keeps in mind the other objective of getting close to  $g^*$ , then the choice  $\mathbf{m} = \mathbf{m}^*$  becomes very natural, and we are led to considering the optimization problem of Theorem 3.2 (when  $\in \mathcal{L}_{n,1}$  is a rank-1 perturbation of the identity).

## 4 Computational framework

### 4.1 Numerical procedure for IS estimate comparison

The objective of the numerical simulations is to evaluate the impact of the choice of the covariance matrix on the estimation accuracy of a high dimensional integral  $\mathcal{E}$ . We thus want to compare the IS estimation results for different auxiliary densities and more particularly for different choices of the auxiliary covariance matrix when the IS auxiliary density is Gaussian. The details of the considered covariance matrices is given in Section 4.2. To extend this comparison, we also compute the results when the IS auxiliary density is chosen with the von Mises–Fisher–Nakagami (vMFN) model recently proposed in (Papaioannou, Geyer, and Straub 2019) for high dimensional probability estimation (See Appendix B).

In Section 5 we test these different models of auxiliary densities on five test cases, where  $f$  is a standard Gaussian density. This choice is not a theoretical limitation as we can in principle always come back to this case by transforming the vector  $\mathbf{X}$  with isoprobabilistic transformations (see for instance (Hohenbichler and Rackwitz 1981), (Liu and Der Kiureghian 1986)).

The precise numerical framework that we will consider to assess the efficiency of the different auxiliary models is as follows. We assume first that  $M$  i.i.d. random samples  $\mathbf{X}_1^*, \dots, \mathbf{X}_M^*$  distributed from  $g^*$  are available from rejection sampling (unless in Appendix C where we consider MCMC). From these samples, the parameters of the Gaussian and of the vMFN auxiliary density are computed to get an auxiliary density  $g'$ . Finally,  $N$  samples are generated from  $g'$  to provide an estimation of  $\mathcal{E}$  with IS. This procedure is summarized by the following stages:

1. Generate a sample  $\mathbf{X}_1^*, \dots, \mathbf{X}_M^*$  independently according to  $g^*$ ;
2. From  $\mathbf{X}_1^*, \dots, \mathbf{X}_M^*$ , compute the parameters of the auxiliary parametric density  $g'$ ;



3. Generate a new sample  $\mathbf{X}_1, \dots, \mathbf{X}_N$  independently from  $g'$ ;
4. Estimate  $\mathcal{E}$  with  $\widehat{\mathcal{E}}_N = \frac{1}{N} \sum_{i=1}^N \phi(\mathbf{X}_i) \frac{f(\mathbf{X}_i)}{g'(\mathbf{X}_i)}$ .

The number of samples  $M$  and  $N$  are respectively set to  $M = 500$  and  $N = 2000$ . The computational cost to generate  $M = 500$  samples distributed from  $g^*$  with rejection sampling is often unaffordable in practice; if  $\mathcal{E}$  is a probability of order  $10^{-p}$ , then approximately  $500 \times 10^p$  calls to  $\phi$  are necessary for the generation of  $\mathbf{X}_1^*, \dots, \mathbf{X}_M^*$ . Finally, whatever the auxiliary parametric density  $g'$  computed from  $\mathbf{X}_1^*, \dots, \mathbf{X}_M^*$ , the number of calls to  $\phi$  for the estimation step stays constant and equal to  $N$ . The number of calls to  $\phi$  for the whole procedure on a  $10^{-p}$  probability estimation is about  $500 \times 10^p + N$ . A more realistic situation is considered in Appendix C where MCMC is applied to generate samples from  $g^*$ . The resulting samples are dependent but the computational cost is significantly reduced. The number of calls to  $\phi$  with MCMC is then equal to  $M$  which leads to a total computational cost of  $M + N$  for the whole procedure.

This procedure is then repeated 500 times to provide a mean estimation  $\widehat{\mathcal{E}}$  of  $\mathcal{E}$ . In the result tables, for each auxiliary density  $g'$  we report the corresponding value for the relative error  $\widehat{\mathcal{E}}/\mathcal{E} - 1$  and the coefficient of variation of the 500 iterations (the empirical standard deviation divided by  $\widehat{\mathcal{E}}$ ). As was established in the proof of Theorem 3.1, the KL divergence is, up to an additive constant, equal to  $D'() = \log\| + \text{tr}(*^{-1})$  which we will refer to as partial KL divergence. In the result tables, we also report thus the mean value of  $D'()$  to analyse the relevance of the auxiliary density  $g_{\widehat{\mathbf{m}}^*}$ , for six choices of covariance matrix. The next sections specify the different parameters of  $g'$  for the Gaussian model and for the vMFN model we have considered in the simulations.

## 4.2 Choice of the auxiliary density $g'$ for the Gaussian model

The goal is to get benchmark results to assess whether one can improve estimations of Gaussian IS auxiliary density by projecting the covariance matrix  $*$  in the proposed directions  $\mathbf{d}_i^*$ . The algorithm that we study here (Algorithms 1+2) aims more precisely at understanding whether:

- projecting can improve the situation with respect to the empirical covariance matrix;
- the  $\mathbf{d}_i^*$ 's are good candidates, in particular compared to the choice  $\mathbf{m}^*$  suggested in (El Masri, Morio, and Simatos 2021);
- what is the impact in making errors in estimating the eigenpairs  $(\lambda_i^*, \mathbf{d}_i^*)$ .

Let us define the estimate  $\widehat{\mathbf{m}}^*$  of  $\mathbf{m}^*$  from the  $M$  i.i.d. random samples  $\mathbf{X}_1^*, \dots, \mathbf{X}_M^*$  distributed from  $g^*$  with

$$\widehat{\mathbf{m}}^* = \frac{1}{M} \sum_{i=1}^M \mathbf{X}_i^*. \quad (8)$$

In our numerical test cases, we will compare six different choices of Gaussian auxiliary distributions  $g'$  with mean  $\widehat{\mathbf{m}}^*$  and the following covariance matrices summarized in Table 1:

1.  $*$ : the optimal covariance matrix given by Equation 3;
2.  $\widehat{*}$ : the empirical estimation of  $*$  given by

$$\widehat{*} = \frac{1}{M} \sum_{i=1}^M (\mathbf{X}_i^* - \widehat{\mathbf{m}}^*)(\mathbf{X}_i^* - \widehat{\mathbf{m}}^*)^\top. \quad (9)$$

The four other covariance matrices considered in the numerical simulations are of the form  $\sum_{i=1}^k (v_i - 1)\mathbf{d}_i \mathbf{d}_i^\top + I_n$  where  $v_i$  is the variance of  $*$  in the direction  $\mathbf{d}_i$ ,  $v_i = \mathbf{d}_i^\top * \mathbf{d}_i$ . The considered choice of  $k$  and  $\mathbf{d}_i$  gives the following covariance matrices:

3.  $\hat{\Sigma}_{\text{opt}}^*$  is obtained by choosing  $\mathbf{d}_i = \mathbf{d}_i^*$  of Theorem 3.1, which is supposed to be perfectly known from  $\Sigma^*$  and  $k$  is computed with Algorithm 2;
4.  $\hat{\Sigma}_{\text{opt}}^{+d}$  is obtained by choosing  $\mathbf{d}_i = \hat{\mathbf{d}}_i^*$  the  $i$ -th eigenvector of  $\Sigma^*$  (in  $\ell$ -order), which is an estimation of  $\mathbf{d}_i^*$ , and  $k$  is computed with Algorithm 2;
5.  $\hat{\Sigma}_{\text{mean}}^*$  is obtained by choosing  $k = 1$  and  $\mathbf{d}_1 = \mathbf{m}^* / \|\mathbf{m}^*\|$ ;
6.  $\hat{\Sigma}_{\text{mean}}^{+d}$  is obtained by choosing  $k = 1$  and  $\mathbf{d}_1 = \hat{\mathbf{m}}^* / \|\hat{\mathbf{m}}^*\|$ , where  $\hat{\mathbf{m}}^*$  given by Equation 8.

The matrices  $\hat{\Sigma}_{\text{opt}}^*$  and  $\hat{\Sigma}_{\text{mean}}^*$  use the estimation  $\Sigma^*$  with the optimal directions  $\mathbf{d}_i^*$  or  $\mathbf{m}^*$ , while the matrices  $\hat{\Sigma}_{\text{opt}}^{+d}$  and  $\hat{\Sigma}_{\text{mean}}^{+d}$  involve an estimation of these directions from  $\Sigma^*$ . By definition,  $\Sigma^*$  will give optimal results, while results for  $\Sigma^*$  will deteriorate as the dimension increases, which is the well-known behavior which we try to improve. Moreover,  $\Sigma^*$  and the projection directions  $\mathbf{d}_i^*$  or  $\mathbf{m}^*$ , are of course unknown in practice. For simulation comparison purpose, they could be determined analytically in simple test cases and otherwise we obtained them by a brute force Monte Carlo scheme with a very high simulation budget. Finally, we emphasize that Algorithm 1 corresponds to estimating and projecting on the  $\mathbf{d}_i^*$ 's, and so the matrix  $\hat{\Sigma}_k^*$  of Algorithm 1 is equal to the matrix  $\hat{\Sigma}_{\text{opt}}^{+d}$ .

Table 1: Presentation of the six covariance matrices considered in the numerical examples.

	$\Sigma^*$	$\Sigma^{+d}$	$\hat{\Sigma}_{\text{opt}}^*$	$\hat{\Sigma}_{\text{mean}}^*$	$\hat{\Sigma}_{\text{opt}}^{+d}$	$\hat{\Sigma}_{\text{mean}}^{+d}$
Initial covariance matrix	$\Sigma^*$	$\Sigma^{+d}$	$\Sigma^{+d}$	$\Sigma^{+d}$	$\Sigma^{+d}$	$\Sigma^{+d}$
Projection directions (exact or estimated)	-	-	Exact	Exact	Estimated	Estimated
Choice for the projection direction	None	None	Opt	Mean	Opt	Mean

## 5 Numerical results on five test cases

The proposed numerical framework is applied on three examples that are often considered to assess the performance of importance sampling algorithms and also two test cases from the area of financial mathematics.

### 5.1 Test case 1: one-dimensional optimal projection

We consider a test case where all computations can be made exactly. This is a classical example of rare event probability estimation, often used to test the robustness of a method in high dimensions. It is given by  $\phi(\mathbf{x}) = \mathbb{I}_{\{\varphi(\mathbf{x}) \geq 0\}}$  with  $\varphi$  the following affine function:

$$\varphi : \mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n \mapsto \sum_{j=1}^n x_j - 3\sqrt{n}. \quad (10)$$

The quantity of interest  $\mathcal{E}$  is defined as  $\mathcal{E} = \int_{\mathbb{R}^n} \phi(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \mathbb{P}(\varphi(\mathbf{X}) \geq 0) \simeq 1.35 \cdot 10^{-3}$  for all  $n$  where the density  $f$  is the standard  $n$ -dimensional Gaussian distribution. Here, the zero-variance density is  $g^*(\mathbf{x}) = \frac{f(\mathbf{x}) \mathbb{I}_{\{\varphi(\mathbf{x}) \geq 0\}}}{\mathcal{E}}$ , and the optimal parameters  $\mathbf{m}^*$  and  $\Sigma^*$  in Equation 3 can be computed exactly, namely  $\mathbf{m}^* = \alpha \mathbf{1}$  with  $\alpha = e^{-9/2} / (\mathcal{E} (2\pi)^{1/2})$  and  $\mathbf{1} = \frac{1}{\sqrt{n}}(1, \dots, 1) \in \mathbb{R}^n$  the normalized constant vector, and  $\Sigma^* = (v - 1) \mathbf{1} \mathbf{1}^\top + I_n$  with  $v = 3\alpha - \alpha^2 + 1$ .

### 5.1.1 Evolution of the partial KL divergence and spectrum

Figure 2a represents the evolution as the dimension varies between 5 and 100 of the partial KL divergence  $D'$  for three different choices of covariance matrix: the optimal matrix  $^*$ , its empirical estimation  $^{\wedge*}$  and the estimation  $^{\wedge*}_k$  of the optimal lower-dimensional covariance matrix. We can notice that the partial KL divergence for  $^{\wedge*}_k$  grows much faster than the other two, and that the partial KL divergence for  $^{\wedge*}_k$  remains very close to the optimal value  $D'(^*)$ . As the KL divergence is a proxy for the efficiency of the auxiliary density (it is for instance closely related to the number of samples required for a given precision (Chatterjee and Diaconis 2018)), this suggests that using  $^{\wedge*}_k$  will provide results close to optimal.

We now check this claim. As  $^* = (\nu - 1)\mathbf{1}\mathbf{1}^\top + I_n$ , its eigenpairs are  $(\nu, \mathbf{1})$  and  $(1, \mathbf{d}_i)$  where the  $\mathbf{d}_i$ 's form an orthonormal basis of the space orthogonal to the space spanned by  $\mathbf{1}$ . In particular,  $(\nu, \mathbf{1})$  is the largest (in  $\ell$ -order) eigenpair of  $^*$  and  $^*_k = ^*$  for any  $k \geq 1$ .

In practice, we do not use this theoretical knowledge and  $^*, ^*_k$  and the eigenpairs are estimated. The six covariance matrices introduced in Section 4.2 and in which we are interested are as follows:

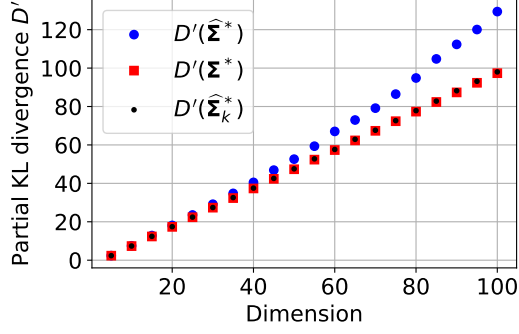
- $^* = (\nu - 1)\mathbf{1}\mathbf{1}^\top + I_n$ ;
- $^{\wedge*}$  given by Equation 9;
- $^{\wedge*}_{\text{opt}}$  and  $^{\wedge*}_{\text{mean}}$  are equal and given by  $(\hat{\lambda} - 1)\mathbf{1}\mathbf{1}^\top + I_n$  with  $\hat{\lambda} = \mathbf{1}^\top ^{\wedge*} \mathbf{1}$ . This amounts to assuming that the projection direction  $\mathbf{1}$  is perfectly known, whereas the variance in this direction is estimated;
- $^{\wedge*}_{\text{opt}} = (\hat{\lambda} - 1)\hat{\mathbf{d}}\hat{\mathbf{d}}^\top + I_n$  with  $(\hat{\lambda}, \hat{\mathbf{d}})$  the smallest eigenpair of  $^{\wedge*}$ . The difference with the previous case is that we do not assume anymore that the optimal projection direction  $\mathbf{1}$  is known, and so it needs to be estimated;
- $^{\wedge*}_{\text{mean}} = (\hat{\lambda} - 1)\frac{\hat{\mathbf{m}}^*(\hat{\mathbf{m}}^*)^\top}{\|\hat{\mathbf{m}}^*\|^2} + I_n$  with  $\hat{\mathbf{m}}^*$  given by Equation 8 and  $\hat{\lambda} = \frac{(\hat{\mathbf{m}}^*)^\top \hat{\mathbf{m}}^*}{\|\hat{\mathbf{m}}^*\|^2}$ . Here we assume that  $\mathbf{m}^*$  is a good projection direction, but is unknown and therefore needs to be estimated.

Note that in the particularly simple case considered here, both  $\hat{\mathbf{m}}^*/\|\hat{\mathbf{m}}^*\|$  and  $\hat{\mathbf{d}}$  are estimators of  $\mathbf{1}$  but they are obtained by different methods. In the next example we will consider a case where  $\mathbf{m}^*$  is not an optimal projection direction as given by Theorem 3.1.

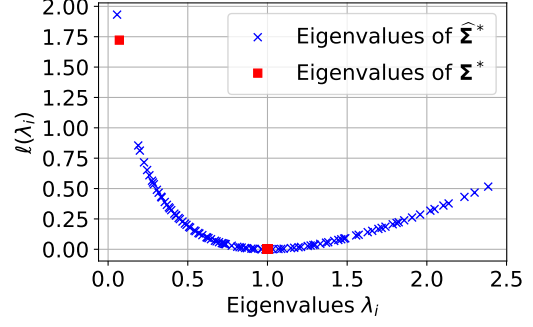
Figure 2b represents the images by  $\ell$  of the eigenvalues of  $^*$  and  $^{\wedge*}$ . This picture carries a very important insight. We notice that the estimation of most eigenvalues is poor: indeed, all the blue crosses except the leftmost one are meant to be estimator of 1, whereas we see that they are more or less uniformly spread around 1. This means that the variance terms in the corresponding directions are poorly estimated, which could be the explanation on why the use of  $^{\wedge*}$  gives an inaccurate estimation. But what we remark also is that the function  $\ell$  is quite flat around one: as a consequence, although the eigenvalues offer significant variability, this variability is smoothed by the action of  $\ell$ . Indeed, the images of the eigenvalues by  $\ell$  take values between 0 and 0.8 and have smaller variability. Moreover,  $\ell(x)$  increases sharply as  $x$  approaches 0 and thus efficiently distinguishes between the two leftmost estimated eigenvalues and is able to separate them.

### 5.1.2 Numerical results

We report in Table 2 the numerical results for the six different matrices and the vMFN model for the dimension  $n = 100$ . The column  $^*$  gives the optimal results, while the column  $^{\wedge*}$  corresponds to the results that we are trying to improve. Comparing these two columns, we notice as expected that the estimation of  $\mathcal{E}$  with  $^{\wedge*}$  is significantly degraded. Compared to the first column  $^*$ , the third and fourth columns with  $^{\wedge*}_{\text{opt}} = ^{\wedge*}_{\text{mean}}$  correspond to the best projection direction  $\mathbf{1}$  (as for  $^*$ ) but estimating the variance in this direction (instead of the true variance) with  $\mathbf{1}^\top ^{\wedge*} \mathbf{1}$ . This choice performs very well, with numerical results similar to the optimal ones. This can be understood since in this case, both



(a) Evolution of the partial KL divergence as the dimension increases, with the optimal covariance matrix  $\Sigma^*$  (red squares), the sample covariance  $\hat{\Sigma}^*$  (blue circles), and the projected covariance  $\hat{\Sigma}_k^*$  (black dots).



(b) Computation of  $\ell(\lambda_i)$  for the eigenvalues of  $\Sigma^*$  (red squares) and  $\hat{\Sigma}^*$  (blue crosses) in dimension  $n = 100$ .

Figure 2: Partial KL divergence and spectrum for the function  $\phi = \mathbb{I}_{\varphi \geq 0}$  with  $\varphi$  the linear function given by Equation 10.

$\hat{\Sigma}_{\text{opt}}^*$  and  $\Sigma^*$  are of the form  $\alpha \mathbf{1}\mathbf{1}^\top + I_n$  and so estimating  $\hat{\Sigma}_{\text{opt}}^*$  requires only a one-dimensional estimation (namely, the estimation of  $\alpha$ ). Next, the last two columns  $\hat{\Sigma}_{\text{opt}}^{+d}$  and  $\hat{\Sigma}_{\text{mean}}^{+d}$  highlight the impact of having to estimate the projection directions in addition to the variance since these two matrices are of the form  $\hat{\alpha} \hat{\mathbf{1}}\hat{\mathbf{1}}^\top + I_n$  with both  $\hat{\alpha}$  (the variance term) and  $\hat{\mathbf{1}}$  (the direction) being estimated. We observe that these matrices yield results which are close to optimal and greatly improve the estimation obtained using  $\hat{\Sigma}^*$ .

Moreover, we observe that  $\hat{\Sigma}_{\text{mean}}^{+d}$  gives better results than  $\hat{\Sigma}_{\text{opt}}^{+d}$ . We suggest that this is because  $\hat{\mathbf{m}}^* / \|\hat{\mathbf{m}}^*\|$  is a better estimator of  $\mathbf{1}$  than the eigenvector of  $\hat{\Sigma}^*$ . Indeed, evaluating  $\hat{\mathbf{m}}^*$  requires the estimation of  $n$  parameters, whereas  $\hat{\Sigma}^*$  needs around  $n^2/2$  parameters to estimate, so the eigenvector is finally more noisy than the mean vector. In the last column, we present the vMFN estimation that is slightly more efficient than the estimation obtained with  $\hat{\Sigma}_{\text{mean}}^{+d}$ .

Thus, the proposed idea improves significantly the probability estimation in high dimensions. But we see that the method taken in (El Masri, Morio, and Simatos 2021) with the projection  $\mathbf{m}^*$  is at least as much efficient in this example where we need only a one-dimensional projection. The next case shows that the projection on more than one direction can outperform the one-dimensional projection on  $\mathbf{m}^*$ .

Table 2: Numerical comparison of the estimation of  $\mathcal{E} \approx 1.35 \cdot 10^{-3}$  considering the Gaussian model with the six covariance matrices defined in Section 4.2 and the vFMN model, when  $\phi = \mathbb{I}_{\varphi \geq 0}$  with  $\varphi$  the linear function given by Equation 10. As explained in the text,  $\hat{\Sigma}_{\text{mean}}^*$  and  $\hat{\Sigma}_{\text{opt}}^*$  are actually equal in this case. The computational cost is  $N = 2000$ .

	$\Sigma^*$	$\hat{\Sigma}^*$	$\hat{\Sigma}_{\text{opt}}^*$	$\hat{\Sigma}_{\text{mean}}^*$	$\hat{\Sigma}_{\text{opt}}^{+d}$	$\hat{\Sigma}_{\text{mean}}^{+d}$	vMFN
D'	0	0	0	0	0	0	/
Relative error (%)	0	0	0	0	0	0	0.0

Coefficient of variation (%)	0	0	0	0	0	0	0	0.0
------------------------------	---	---	---	---	---	---	---	-----

---

## 5.2 Test case 2: projection in 2 directions

The second test case is again a probability estimation, i.e., it is of the form  $\phi = \mathbb{I}_{\{\phi \geq 0\}}$  with now the function  $\phi$  having some quadratic terms:

$$\phi : \mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n \mapsto x_1 - 25x_2^2 - 30x_3^2 - 1. \quad (11)$$

The quantity of interest  $\mathcal{E}$  is defined as  $\mathcal{E} = \int_{\mathbb{R}^n} \phi(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \mathbb{P}_f(\phi(\mathbf{X}) \geq 0)$  for all  $n$  where the density  $f$  is the standard  $n$ -dimensional Gaussian distribution. This function is motivated in part because  $\mathbf{m}^*$  and  $\mathbf{d}_1^*$  are different and also because Algorithm 2 chooses two projection directions. Thus, this is an example where  $\hat{\mathbf{m}}_{\text{mean}}$  and  $\hat{\mathbf{m}}_{\text{opt}}$  are significantly different.

### 5.2.1 Evolution of the partial KL divergence and spectrum

We check on ?@fig-inefficiency-parab-1 that the partial KL divergence obeys the same behavior as for the previous example, namely the one associated with  $\hat{\mathbf{m}}^*$  increases much faster than the ones associated with  $\mathbf{m}^*$  and  $\hat{\mathbf{d}}_k^*$ , which again suggests that projecting can improve the situation. Since the function  $\phi$  only depends on the first three variables and is even in  $x_2$  and  $x_3$ , one gets that  $\mathbf{m}^* = \alpha \mathbf{e}_1$  with  $\alpha = \mathbb{E}(X_1 \mid X_1 \geq 25X_2^2 + 30X_3^2 + 1) \approx 1.9$  (here and in the sequel,  $\mathbf{e}_i$  denotes the  $i$ th canonical vector of  $\mathbb{R}^n$ , i.e., all its coordinates are 0 except the  $i$ -th one which is equal to one), and that  $\hat{\mathbf{m}}^*$  is diagonal with

$$\hat{\mathbf{m}}^* = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & 0 & \dots & 0 \\ 0 & 0 & \lambda_3 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

Note that the off-diagonal elements of the submatrix  $(\hat{\mathbf{m}}^*)_{1 \leq i, j \leq 3}$  are indeed 0 since they result from integrating an odd function of an odd random variable with an even conditioning. For instance, if  $F(x) = \mathbb{P}(30X_3^2 + 1 \leq x)$ , then by conditioning on  $(X_1, X_3)$  we obtain

$$\hat{m}_{12}^* = \mathbb{E}((X_1 - \alpha)X_2 \mid X_1 - 25X_2^2 \geq 30X_3^2 + 1) = \frac{1}{\mathcal{E}} \mathbb{E}[(X_1 - \alpha) \mathbb{E}(X_2 F(X_1 - 25X_2^2) \mid X_1)]$$

which is 0 as  $x_2 F(x_1 - x_2^2)$  is an odd function of  $x_2$  for fixed  $x_1$ , and  $X_2$  has an even density.

We can numerically compute  $\lambda_1 \approx 0.28$ ,  $\lambda_2 \approx 0.009$  and  $\lambda_3 \approx 0.008$ . These values correspond to the red squares in ?@fig-inefficiency-parab-2 which shows that the smallest eigenvalues are properly estimated. Moreover, Algorithm 2 selects the two largest eigenvalues, which have the highest  $\ell$ -values. These two eigenvalues thus correspond to the eigenvectors  $\mathbf{e}_2$  and  $\mathbf{e}_3$ , and so we see that on this example, the optimal directions predicted by Theorem 3.1 are significantly different (actually, orthogonal) from  $\mathbf{m}^*$  which is proportional to  $\mathbf{e}_1$ .

## References

Agapiou, Sergios, Omiros Papaspiliopoulos, Daniel Sanz-Alonso, and Andrew M Stuart. 2017. “Importance Sampling : Intrinsic Dimension and Computational Cost.” *Statistical Science* 32 (3): 405–31. <https://doi.org/10.1214/17-STS611>.

- Ashurbekova, Karina, Antoine Usseglio-Carleve, Florence Forbes, and Sophie Achard. 2020. "Optimal Shrinkage for Robust Covariance Matrix Estimators in a Small Sample Size Setting."
- Au, S. K., and J. L. Beck. 2003. "Important Sampling in High Dimensions." *Structural Safety* 25 (2): 139–63. [https://doi.org/10.1016/S0167-4730\(02\)00047-4](https://doi.org/10.1016/S0167-4730(02)00047-4).
- Bengtsson, Thomas, Peter Bickel, and Bo Li. 2008. "Curse-of-Dimensionality Revisited: Collapse of the Particle Filter in Very Large Scale Systems." In *Institute of Mathematical Statistics Collections*, 316–34. Beachwood, Ohio, USA: Institute of Mathematical Statistics. <https://doi.org/10.1214/193940307000000518>.
- Bucklew, James. 2013. "Introduction to Rare Event Simulation." In, 58–61. Springer Science & Business Media. <https://doi.org/10.1007/978-1-4757-4078-3>.
- Bugallo, Monica F., Victor Elvira, Luca Martino, David Luengo, Joaquin Miguez, and Petar M. Djuric. 2017. "Adaptive Importance Sampling: The Past, the Present, and the Future." *IEEE Signal Processing Magazine* 34 (4): 60–79. <https://doi.org/10.1109/MSP.2017.2699226>.
- Cappé, Olivier, Randal Douc, Arnaud Guillin, Jean-Michel Marin, and Christian P. Robert. 2008. "Adaptive Importance Sampling in General Mixture Classes." *Statistics and Computing* 18 (4): 447–59. <https://doi.org/10.1007/s11222-008-9059-x>.
- Chan, Joshua C. C., and Dirk P. Kroese. 2012. "Improved Cross-Entropy Method for Estimation." *Statistics and Computing* 22 (5): 1031–40. <https://doi.org/10.1007/s11222-011-9275-7>.
- Chatterjee, Sourav, and Persi Diaconis. 2018. "The Sample Size Required in Importance Sampling." *The Annals of Applied Probability* 28 (2): 1099–1135. <https://doi.org/10.1214/17-AAP1326>.
- Cornuet, Jean-Marie, Jean-Michel Marin, Antonietta Mira, and Christian P. Robert. 2012. "Adaptive Multiple Importance Sampling." *Scandinavian Journal of Statistics* 39 (4): 798–812. <https://doi.org/10.1111/j.1467-9469.2011.00756.x>.
- El Masri, Maxime, Jérôme Morio, and Florian Simatos. 2021. "Improvement of the Cross-Entropy Method in High Dimension for Failure Probability Estimation Through a One-Dimensional Projection Without Gradient Estimation." *Reliability Engineering & System Safety* 216: 107991. <https://doi.org/10.1016/j.ress.2021.107991>.
- El-Laham, Yousef, Victor Elvira, and Mónica Bugallo. 2019. "Recursive Shrinkage Covariance Learning in Adaptive Importance Sampling." In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 624–28. IEEE. <https://doi.org/10.1109/CAMSAP45676.2019.9022450>.
- Fan, Jianqing, Yingying Fan, and Jinchi Lv. 2008. "High Dimensional Covariance Matrix Estimation Using a Factor Model." *Journal of Econometrics* 147 (1): 186–97.
- Grace, Adam W., Dirk P. Kroese, and Werner Sandmann. 2014. "Automated State-Dependent Importance Sampling for Markov Jump Processes via Sampling from the Zero-Variance Distribution." *Journal of Applied Probability* 51 (3): 741–55. <https://doi.org/10.1239/jap/1409932671>.
- Hohenbichler, Michael, and Rüdiger Rackwitz. 1981. "Non-Normal Dependent Vectors in Structural Safety." *Journal of the Engineering Mechanics Division* 107 (6): 1227–38. <https://doi.org/10.1061/JMCEA3.0002777>.
- Ledoit, Olivier, and Michael Wolf. 2004. "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices." *Journal of Multivariate Analysis* 88 (2): 365–411. [https://doi.org/10.1016/S0047-259X\(03\)00096-4](https://doi.org/10.1016/S0047-259X(03)00096-4).
- Liu, Pei-Ling, and Armen Der Kiureghian. 1986. "Multivariate Distribution Models with Prescribed Marginals and Covariances." *Probabilistic Engineering Mechanics* 1 (2): 105–12. [https://doi.org/10.1016/0266-8920\(86\)90033-0](https://doi.org/10.1016/0266-8920(86)90033-0).
- Owen, Art, and Yi Zhou. 2000. "Safe and Effective Importance Sampling." *Journal of the American Statistical Association* 95 (449): 135–43. <https://doi.org/10.1080/01621459.2000.10473909>.
- Papaioannou, Iason, Sebastian Geyer, and Daniel Straub. 2019. "Improved Cross Entropy-Based Importance Sampling with a Flexible Mixture Model." *Reliability Engineering & System Safety* 191 (November): 106564. <https://doi.org/10.1016/j.ress.2019.106564>.

- Rubinstein, Reuven Y., and Dirk P Kroese. 2011a. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. New York; London: Springer. <https://doi.org/10.1007/978-1-4757-4321-0>.
- . 2011b. “The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning.” In, 67–72. New York; London: Springer. <https://doi.org/10.1007/978-1-4757-4321-0>.
- Rubinstein, Reuven Y., and Dirk P. Kroese. 2017. “Simulation and the Monte Carlo Method.” In, Third edition, 149–58. Wiley Series in Probability and Statistics. Hoboken, New Jersey: Wiley. <https://doi.org/10.1002/9781118631980>.
- Uribe, Felipe, Iason Papaioannou, Youssef M. Marzouk, and Daniel Straub. 2021. “Cross-Entropy-Based Importance Sampling with Failure-Informed Dimension Reduction for Rare Event Simulation.” *SIAM/ASA Journal on Uncertainty Quantification* 9 (2): 818–47. <https://doi.org/10.1137/20M1344585>.