

**UNIVERSIDADE FEDERAL DE SANTA CATARINA  
DEPARTAMENTO DE ENGENHARIA MECÂNICA**

Fábio Felipe dos Santos Nascentes

**CONTRIBUIÇÕES À EFICIÊNCIA DA OTIMIZAÇÃO  
GLOBAL ESTOCÁSTICA ADAPTATIVA**

Florianópolis

2019



Fábio Felipe dos Santos Nascentes

**CONTRIBUIÇÕES À EFICIÊNCIA DA OTIMIZAÇÃO  
GLOBAL ESTOCÁSTICA ADAPTATIVA**

Dissertação submetida ao Programa de Pós-Graduação em Engenharia Mecânica para a obtenção do Grau de Mestre em Engenharia Mecânica.

Orientador: Eduardo Alberto Fancello, Ph.D.  
Coorientador: Rafael Holdorf Lopez, Dr. Eng.

Florianópolis

2019

Ficha de identificação da obra elaborada pelo autor,  
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Nascentes, Fábio Felipe dos Santos  
Contribuições à eficiência da otimização global  
estocástica adaptativa / Fábio Felipe dos Santos  
Nascentes ; orientador, Eduardo Alberto Fancelllo,  
coorientador, Rafael Holdorf Lopez, 2019.  
228 p.

Dissertação (mestrado) - Universidade Federal de  
Santa Catarina, Centro Tecnológico, Programa de Pós  
Graduação em Engenharia Mecânica, Florianópolis, 2019.

Inclui referências.

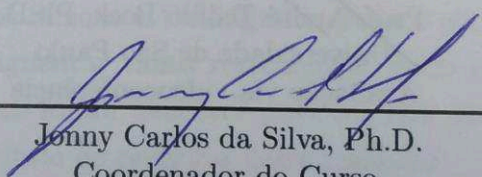
1. Engenharia Mecânica. 2. Otimização global  
estocástica adaptativa. 3. Tunelamento estocástico.  
4. Metamodelos. 5. Funções integrais. I. Fancelllo,  
Eduardo Alberto. II. Lopez, Rafael Holdorf. III.  
Universidade Federal de Santa Catarina. Programa de  
Pós-Graduação em Engenharia Mecânica. IV. Título.

Fábio Felipe dos Santos Nascentes

**CONTRIBUIÇÕES À EFICIÊNCIA DA OTIMIZAÇÃO  
GLOBAL ESTOCÁSTICA ADAPTATIVA**

Esta Dissertação foi julgada aprovada para a obtenção do Título de “Mestre em Engenharia Mecânica”, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia Mecânica.

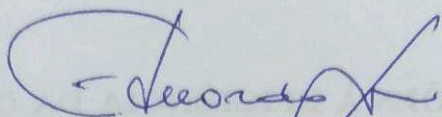
Florianópolis, 30 de Abril de 2019.



---

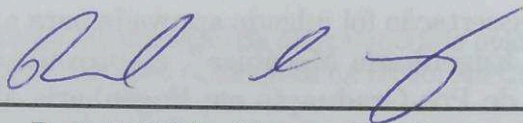
Jonny Carlos da Silva, Ph.D.  
Coordenador do Curso

Banca Examinadora:



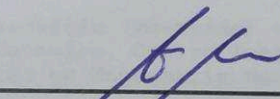
---

Eduardo Alberto Fancello, Ph.D.  
Presidente - Orientador  
Universidade Federal de Santa Catarina



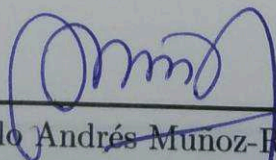
---

Rafael Holdorf Lopez, Dr. Eng.  
Coorientador  
Universidade Federal de Santa Catarina



---

Prof. André Teófilo Beck, Ph.D.  
Universidade de São Paulo  
Avaliação via videoconferência



---

Prof. Pablo Andrés Muñoz-Rojas, Dr. Eng.  
Universidade do Estado de Santa Catarina

## AGRADECIMENTOS

Agradeço primeiramente à minha esposa Natielly Nascentes que partilhou de toda a angústia e espera pela conclusão desse trabalho. Seu apoio e as suas várias interrupções foram de vital importância para a concretização dessa etapa. A ela também agradeço o imenso esforço e carinho ao realizar a primeira revisão deste trabalho.

Em especial, agradeço à minha mãe Graça, minha irmã Marina e meus sogros Valdirene e Marcelo pelo apoio incondicional durante essa etapa intensa.

Ao Professor Eduardo Alberto Fancello o meu muito obrigado por todos os ensinamentos, conselhos e confiança, que de uma forma excepcional possibilitou a realização deste trabalho.

Ao Professor Rafael Holdorf Lopez agradeço de forma mais do que especial pois ele foi o maior responsável pela conclusão deste trabalho. Seus ensinamentos, os momentos de descontração, os milhões de e-mails respondidos e os magníficos conselhos foram capazes de mudar o rumo da minha vida profissional.

Aos amigos Simone e Wellington agradeço por terem paciência em me aturar e aguentar tanta reclamação. Todo o apoio, suporte e compartilhamento foram decisivos para superar esta etapa. A eles também dedico o título de revisores desse trabalho.

Agradeço também ao apoio do Instituto Federal de Educação, Ciência e Tecnologia de Goiás que me concedeu a licença para capacitação para poder realizar o mestrado.

Por fim agradeço aos demais colegas de curso, professores e amigos que de alguma forma me influenciaram positivamente e torceram pela conclusão do trabalho.





*“A tragédia de uma vida não está em não  
alcançar seu objetivo. A tragédia está em  
não ter objetivo para alcançar”*

*Benjamin Mays*



## Resumo

O recente avanço computacional vem proporcionando análises cada vez mais sofisticadas de sistemas estruturais. Nesse contexto, problemas de otimização estrutural também ganharam mais notoriedade, principalmente quando aliados ao estudo da confiabilidade estrutural e à análise de incertezas. É natural desses problemas que suas funções sejam de alto custo computacional e extremamente multimodais. Portanto, neste trabalho foi desenvolvido um ferramental algorítmico para a solução de tais problemas utilizando os metamodelos e a Otimização Global Estocástica Eficiente (*Stochastic Efficient Global Optimization* - sEGO). Foi levada em consideração a abordagem Adaptativa, que é utilizada no manejo inteligente do orçamento computacional durante a execução do sEGO. Para diminuir a variabilidade que os valores da função objetivo podem assumir e mantermos uma distribuição aceitável de seu valor, aplicamos a técnica Tunelamento Estocástico como método de normalização, cuja finalidade é a de diminuir a extensão do contradomínio das funções objetivo. Para testarmos a robustez e eficiência do sEGO Adaptativo, foram implementadas cinco técnicas de adição de pontos de preenchimento. Entre elas, a abordagem *Expected Improvement with Reinterpolation* (EIR) é pela primeira vez demonstrada e analisada no contexto de ruídos heterogêneos. Foram realizadas análises numéricas em dezoito problemas de otimização estocásticos, sem restrições, testando as cinco métricas para o sEGO Adaptativo. Os resultados mostram que a normalização possibilita ao sEGO realizar uma maior busca do domínio, mesmo a orçamentos computacionais baixos, obtendo excelentes resultados e fornecendo uma sensibilidade maior à forma como aproximamos a função objetivo. Foi possível identificar que, com uma maior quantidade de pontos adicionados, a robustez do sEGO aumentou para problemas de dimensões elevadas, se comparado aos resultados presentes na literatura. Identificamos também, qual o comportamento obtido pelo sEGO em cada uma das cinco técnicas de adição de pontos. Obtemos desde técnicas que realizam somente uma busca local a técnicas que prezam somente pela busca exploratória do espaço. De posse desses resultados, realizamos uma análise estatística das melhores soluções de modo a encontrar a técnica que faz com que a robustez do sEGO seja a maior, apresentando o melhor valor mínimo e a menor variabilidade nos resultados.

**Palavras-chave:** Funções integrais, Otimização global, *Stochastic Efficient Global Optimization*, *Stochastic Kriging*, Tunelamento Estocástico.

## Abstract

The recent computational advancement has been providing increasingly sophisticated analyzes of structural systems. In this context, structural optimization problems also gained more prominence, especially when allied to the study of structural reliability and the analysis of uncertainties. It is natural for these problems that their functions are of high computational cost and extremely multimodal. Therefore, in this work an algorithmic tool was developed using metamodels and Stochastic Efficient Global Optimization (sEGO) for the solution of such problems. Adaptivity has been taken into account. This approach is used for producing an intelligent handling of the computational budget during the execution of the sEGO. In order to reduce the variability that the objective function values can assume and maintain an acceptable distribution of its value, we apply the Stochastic Tunneling technique as a normalization method, whose purpose is to reduce the range of the objective function. In order to test the robustness and efficiency of the Adaptive sEGO, five infill points addition techniques were implemented. Among them, the Expected Improvement with Reinterpolation (EIR) approach is first demonstrated and analyzed in the context of heterogeneous noises. Numerical analyses were performed on eighteen stochastic optimization problems without constraints testing the five metrics for the Adaptive sEGO. The results show that normalization enables sEGO to perform a greater search of the domain even at low computational budgets obtaining excellent results and providing a greater sensitivity to the way we approach the objective function. It was possible to identify that, with a larger number of points added, the robustness of the sEGO increased in problems of high dimensions, when compared to the results present in the literature. We also identify the behavior of sEGO in each of the five techniques of infill points. We obtain from techniques that only perform a local search, to techniques that only make an exploratory search of the space. With these results, we performed a statistical analysis of the best solutions in order to find the technique that makes the largest robustness of sEGO, presenting the best minimum value and the lowest variability in the results.

**Keywords:** Integral functions, Global optimization, Stochastic Efficient Global Optimization, Stochastic Kriging, Stochastic tunneling.



## Lista de ilustrações

Figura 1 – Plataforma simplesmente apoiada sujeita a vibrações.	31
Figura 2 – Estrutura sujeita a variabilidade estocástica dos ventos e incerteza nos materiais. . . . .	32
Figura 3 – Elemento MEF em 3D com um alvo circular. . . . .	33
Figura 4 – Função estocástica bidimensional. . . . .	36
Figura 5 – Convexidade de uma função 1D. . . . .	44
Figura 6 – Tipos de metamodelos. . . . .	53
Figura 7 – Aproximação em metamodelos polinomiais. . . . .	56
Figura 8 – Análise dos erros na regressão. . . . .	59
Figura 9 – Influência dos parâmetros na correlação. . . . .	62
Figura 10 – Modelos substitutos da função (3.6). . . . .	72
Figura 11 – Interpolação com o Kriging e seu erro. . . . .	73
Figura 12 – Metamodelo em Kriging e seu RMSE. . . . .	76
Figura 13 – Incerteza sobre o valor predito em $d^a$ . . . . .	78
Figura 14 – Diferentes avaliações do EI durante o EGO. . . . .	82
Figura 15 – Fluxograma do EGO. . . . .	83
Figura 16 – Influência de $n_r$ para o MCI. . . . .	90
Figura 17 – Kriging determinístico para a função estocástica. . . . .	92
Figura 18 – Metamodelo considerando a nova matriz de correlação. . . . .	95
Figura 19 – Aproximação via SK. . . . .	101
Figura 20 – Aplicação da normalização. . . . .	123
Figura 21 – Modificação da função original via normalização. . . . .	125
Figura 22 – Função estocástica e sua normalização. . . . .	126
Figura 23 – Aproximação via SK para função 1D normalizada. . . . .	129
Figura 24 – sEGO utilizando a métrica MQ. . . . .	130
Figura 25 – sEGO utilizando a métrica AEI. . . . .	132
Figura 26 – sEGO utilizando a métrica EQI. . . . .	133
Figura 27 – sEGO utilizando a métrica TSSO. . . . .	135
Figura 28 – sEGO utilizando a métrica EIR. . . . .	136
Figura 29 – Gráficos determinístico e estocásticos para a função 1D-3. . . . .	142
Figura 30 – BP para análise estatística. . . . .	155

Figura 31 – Comparação entre o sEGO adaptativo com normalização e sem normalização. Os retângulos laranja e vermelho representam, respectivamente, os resultados com e sem normalização. . . . . 160

Figura 32 – Influência de  $\bar{\sigma}_0^2$  sobre o sEGO adaptativo com normalização. . . . . 168

Figura 33 – Testes estatísticos ao nível de 0.05 como confiança, para os resultados do sEGO adaptativo normalizado via MQ com P2. . . . . 174

Figura 34 – Escores acumulados por cada métrica ao longo da análise dos dezoito problemas propostos. . . . . 177



## Lista de tabelas

Tabela 2	– Principais características da função 1D-3. . . . .	144
Tabela 3	– Nomenclatura das funções analisadas. . . . .	151
Tabela 4	– Número dos piores resultados apresentados por cada variância inicial, entre todos os problemas. . . . .	167
Tabela 5	– Número de melhores resultados para cada métrica entre todos os problemas. . . . .	179
Tabela 6	– Resultados obtidos na comparação entre o sEGO com e sem normalização. Minimizador P2 e $\sigma_0^2 = 0.01$ . . . . .	225
Tabela 7	– Resultados obtidos em alguns problemas utilizando as cinco métricas e para três variâncias iniciais diferentes. . . . .	226
Tabela 8	– Escores acumulados de cada uma das cinco métricas com os dois minimizadores e para os dezoito proble- mas analisados. . . . .	228



## Lista de abreviaturas e siglas

AEI	<i>Augmented Expected Improvement</i>
CD	Extensão ( <i>range</i> ) do contradomínio de uma função
EGO	<i>Efficient Global Optimization</i>
EI	<i>Expected Improvement</i>
EIR	<i>Expected Improvement with Reinterpolation</i>
EQI	<i>Expected Quantile Improvement</i>
ER	Erro Relativo
IP	<i>Infill Points</i>
MCI	<i>Monte Carlo Integration</i>
MEF	Método dos Elementos Finitos
MQ	<i>Minimal Quantile</i>
MSE	<i>Mean Squared Error</i>
PSO	<i>Particle Swarm Optimization</i>
RBF	<i>Radial Basis Function</i>
RMSE	<i>Root Mean Squared Error</i>
SA	<i>Simulated Annealing</i>
sEGO	<i>Stochastic Efficient Global Optimization</i>
SGA	<i>Search Group Algorithm</i>
SK	<i>Stochastic Kriging</i>
TSSO	<i>Two-Stage Sequential Optimization</i>



## Lista de símbolos

$\mathbf{d}$	Vetor de variáveis de projeto ou variável independente das funções
$\mathbf{d}^+$	Ponto a ser predito pelo Kriging
$\mathbf{d}^{n+1}$	Novo IP a ser adicionado ao espaço amostral
$\mathbf{d}^{**}$	Melhor solução efetiva
$f$	Função determinística
$f_{\mathbf{x}}$	Função densidade de probabilidade para o vetor de variáveis aleatórias $\mathbf{X}$
$f_{\min}$	Valor mínimo dos valores amostrais de uma função determinística
$F$	Função normalizada
$h$	Função de correlação para o Kriging ou função de base para modelos polinomiais ou RBF
$\mathbf{h}$	Vetor de funções de correlação para o Kriging
$I$	Melhora esperada
$J$	Função integral
$J_{\min}$	Valor mínimo da função integral
$J_0$	Valor mínimo da função $J$ para a normalização
$\bar{J}$	Aproximação via MCI para a função integral
$k$	Dimensão do vetor de variáveis de projeto
$k_x$	Dimensão estocástica dos vetores aleatórios
$\mathbf{lb}$	Vetor com os limites inferiores para as variáveis de projeto

$L$	Função de verossimilhança
$L_{\ln}$	Logaritmo da função de verossimilhança
$M$	Tendência média do Kriging
$n$	Número de pontos amostrais
$n_c$	Número de pontos agrupados perto de um IP
$n_r$	Número de replicações do valor da função para o MCI
$p_r$	Parâmetro de suavidade do Kriging
$\mathbf{p}_r$	Vetor dos parâmetros $p_r$ do Kriging
$r_{hc}$	Distância do hipercubo centrado em $\mathbf{d}^{n+1}$
$s^2$	MSE em uma predição com o Kriging
$s$	RMSE em uma predição com o Kriging
$S$	Espaço amostral para construção do metamodelo
$\mathbf{ub}$	Vetor com os limites superiores para as variáveis de projeto
$\mathbf{x}$	Vetor de variáveis aleatórias
$\mathbf{X}$	Vetor de variáveis aleatórias
$w$	Ponderação da análise das métricas e variância
$\mathbf{y}$	Vetor de valores amostrais da função
$\hat{y}$	Predição via metamodelo
$\tilde{\mathbf{y}}$	Vetor aumentado de variáveis aleatórias para o Kriging
$\bar{y}$	Aproximação via MCI para um valor amostral da função integral

$\bar{y}$	Vetor de valores amostrais da função integral
$Y$	Variável aleatória na suposição do Kriging
$\mathbf{Y}$	Vetor dos valores aleatórios na suposição do Kriging
$\hat{Y}$	Predição via Kriging
$\hat{\mathbf{Y}}_r$	Vetor de predições do SK para os pontos amostrais
$Z$	Parcela extrínseca do Kriging
$\mathcal{D}$	Domínio ou região de busca dos problemas de otimização
$\mathbb{E}(I)$	Valor do EI
$\mathcal{L}_{\ln}$	Função logaritmo concentrada da verossimilhança
$\mathcal{N}$	Distribuição normal de probabilidade
$\gamma$	Constante de Tunelamento
$\delta_{ij}$	Delta de Kronecker
$\epsilon$	Erro cometido na aproximação via regressão polinomial
$\varepsilon$	Parcela intrínseca do SK
$\theta_r$	Parâmetro de importância do Kriging
$\boldsymbol{\theta}$	Vetor dos parâmetros $\theta_r$ do Kriging
$\mu$	Tendência central para o Kriging
$\hat{\mu}$	Estimador do parâmetro $\mu$ do Kriging
$\hat{\mu}_r$	Estimador da média do SK
$\sigma^2$	Variância da matriz de covariância do Kriging
$\hat{\sigma}^2$	Estimador para a variância $\sigma^2$ do Kriging

$\sigma_z^2$	Variância da matriz de covariância do Kriging Estocástico
$\hat{\sigma}_z^2$	Estimador da variância $\sigma_z^2$ do SK
$\bar{\sigma}^2$	Variância do erro cometido na aproximação via MCI
$\bar{\sigma}_{\text{alvo}}^2$	Variância alvo para o MCI
$\bar{\sigma}_0^2$	Variância alvo inicial
$\bar{\sigma}_{\text{adap}}^2$	Variância adaptativa
$\bar{\sigma}_{\text{min}}^2$	Variância mínima
$\sigma_x$	Desvio padrão para as variáveis aleatórias
$\Sigma_\epsilon$	Matriz de covariância do erro ou matriz de regularização
$\hat{\Sigma}_\epsilon$	Estimador para a matriz de covariância do erro
$\Sigma_Z$	Matriz de covariância da parcela extrínseca do SK
$\Sigma$	Matriz de covariância do SK
$\hat{\Sigma}$	Estimador da matriz de covariância do SK
$\psi$	Função estocástica
$\Psi$	Matriz de correlação para o Kriging ou a matriz de valores de base para os modelos polinomiais e em RBF
$\tilde{\Psi}$	Matriz de correlação aumentada para o Kriging
$\Omega$	Domínio estocástico

## OPERADORES

$(\cdot)^T$	Transposto de um vetor ou matriz
$(\cdot)^{(i)}$	I-ésima simulação ou i-ésimo valor amostral



$\  \cdot \ $	Norma Euclidiana
$\  \cdot \ _{\infty}$	Norma Infinita
$\text{Cov}(\cdot, \cdot)$	Covariância entre dois elementos
$\text{cor}(\cdot, \cdot)$	Correlação entre dois elementos
$\partial(\cdot)$	Derivada parcial
$P(\cdot)$	Probabilidade de uma variável aleatória
$\mathbb{E}(\cdot)$	Esperança matemática
$\Sigma_Z(\mathbf{d}, \cdot)$	Vetor das covariâncias extrínsecas do SK entre $\mathbf{d}$ e os demais pontos amostrais



## Sumário

1	INTRODUÇÃO . . . . .	29
1.1	CONTEXTO GERAL . . . . .	29
1.2	OBJETIVOS . . . . .	37
1.2.1	Objetivo Geral . . . . .	37
1.2.2	Objetivos Específicos . . . . .	37
1.3	ORGANIZAÇÃO DA DISSERTAÇÃO . . . . .	38
2	PROBLEMAS DE OTIMIZAÇÃO . . . . .	41
2.1	MODELO PADRÃO DE OTIMIZAÇÃO . . . . .	41
2.2	CONVEXIDADE . . . . .	43
2.3	ÓTIMO LOCAL E GLOBAL . . . . .	44
2.4	CARACTERIZAÇÃO DOS PROBLEMAS DE OTIMIZAÇÃO . . . . .	45
2.4.1	Tipos de problemas . . . . .	45
2.4.2	Algoritmos determinísticos de Otimização . . . . .	46
2.4.3	Algoritmos Heurísticos . . . . .	48
3	METAMODELOS . . . . .	51
3.1	METAMODELOS POLINOMIAIS . . . . .	54
3.2	FUNÇÕES RBF . . . . .	56
3.3	KRIGING . . . . .	58
3.3.1	Introdução ao Kriging Determinístico . . . . .	58
3.3.2	As bases de um modelo com o Kriging . . . . .	59
3.3.3	Definição dos parâmetros do Kriging: Verossi- milhança . . . . .	63
3.3.4	Predição com o metamodelo . . . . .	66
3.3.5	Erro na predição . . . . .	71
3.3.6	Alguns exemplos . . . . .	72
3.4	ALGORITMO EGO . . . . .	73
3.4.1	Definição da métrica EI . . . . .	76
4	SEGO ADAPTATIVO COM NORMALIZAÇÃO . . . . .	87

4.1	APROXIMAÇÃO DA FUNÇÃO DE ESTUDO	88
4.2	KRIGING ESTOCÁSTICO	91
4.2.1	Introdução ao Kriging regressor	91
4.2.2	O Kriging Estocástico - SK	96
4.2.3	Estimação dos parâmetros	98
4.3	sEGO PARA FUNÇÕES ESTOCÁSTICAS	101
4.3.1	MQ	104
4.3.2	AEI	105
4.3.3	EQI	107
4.3.4	TSSO	109
4.3.5	EIR	110
4.4	VARIÂNCIA ADAPTATIVA	114
4.4.1	Ajuste da variância adaptativa às métricas do sEGO	117
4.5	NORMALIZAÇÃO APLICADA AO SK	120
4.5.1	A normalização via tunelamento estocástico	122
4.5.2	Influência da normalização no modelo SK	127
4.6	CARACTERÍSTICAS DAS MÉTRICAS DE ADIÇÃO DE IPs	129
4.6.1	Análise do MQ	130
4.6.2	Análise do AEI	132
4.6.3	Análise do EQI	133
4.6.4	Análise do TSSO	134
4.6.5	Análise do EIR	136
5	RESULTADOS NUMÉRICOS	139
5.1	FUNÇÕES UTILIZADAS	140
5.1.1	Função 1D-1	140
5.1.2	Função 1D-2	140
5.1.3	Função 1D-3	141
5.1.4	Função Branin Modificada	144
5.1.5	Função Rosenbrock	145
5.1.6	Função Hartmann 3D	145

5.1.7	Função Colville . . . . .	147
5.1.8	Função Hartman 6D . . . . .	147
5.1.9	Função Levy 10D . . . . .	149
5.2	CONSIDERAÇÕES E CONFIGURAÇÕES GE- RAIS . . . . .	152
5.2.1	Configurações básicas para execução do sEGO	152
5.2.2	Considerações gerais acerca das soluções apre- sentadas . . . . .	153
5.3	sEGO ADAPTATIVO COM NORMALIZAÇÃO E SEM NORMALIZAÇÃO . . . . .	156
5.4	INFLUÊNCIA DA VARIÂNCIA INICIAL PARA O sEGO ADAPTATIVO . . . . .	161
5.4.1	Uma análise via testes estatísticos para a in- fluência da variância inicial . . . . .	163
5.4.2	Análise do pior desempenho do sEGO adap- tativo a partir das variâncias iniciais . . . . .	165
5.5	DETERMINAÇÃO DA MÉTRICA COM OS MELHORES RESULTADOS OBTIDOS PELO sEGO ADAPTATIVO . . . . .	175
5.5.1	Primeira análise: um olhar sobre a variabili- dade dos resultados na perspectiva do percen- til de 90% . . . . .	175
5.5.2	Segunda análise: um olhar sobre a significância de seis estatísticas dos resultados . . . . .	178
6	CONCLUSÃO E TRABALHOS FUTUROS .	181
6.1	CONCLUSÃO . . . . .	181
6.2	TRABALHOS FUTUROS . . . . .	186
	REFERÊNCIAS . . . . .	187

	<b>APÊNDICES</b>	<b>199</b>
	<b>APÊNDICE A – A DISTRIBUIÇÃO NOR-</b>	
	<b>MAL MULTIVARIADA . . .</b>	<b>201</b>
<b>A.1</b>	<b>INTEGRAL GAUSSIANA . . . . .</b>	<b>208</b>
<b>A.2</b>	<b>INTEGRAL AUXILIAR . . . . .</b>	<b>209</b>
	<b>APÊNDICE B – DERIVADAS DA FUNÇÃO</b>	
	<b>DE VEROSSIMILHANÇA</b>	
	<b>PARA O SK . . . . .</b>	<b>213</b>
	<b>APÊNDICE C – SIMPLIFICAÇÕES NUMÉRICAS</b>	
	<b>PARA O ALGORITMO . . .</b>	<b>217</b>
<b>C.1</b>	<b>TÉCNICAS MATRICIAIS PARA A ANÁLISE</b>	
	<b>NUMÉRICA . . . . .</b>	<b>217</b>
<b>C.2</b>	<b>CÁLCULO SIMPLIFICADO DA MÉTRICA</b>	
	<b>EI . . . . .</b>	<b>221</b>
	<b>APÊNDICE D – RESULTADOS ESTATÍSTICOS</b>	<b>225</b>

# 1 INTRODUÇÃO

## 1.1 CONTEXTO GERAL

A arte de projetar é um processo humano complexo que ao longo das últimas décadas vem passando por grandes modificações graças ao avanço computacional. Hoje em dia, grande parte das edificações, automóveis, alimentos, remédios e outros objetos que nos circundam são resultado de algum tipo de projeto. Criar tais projetos normalmente envolve processos de tomadas de decisão.

Dentro do contexto de tomada de decisão, examinamos vários projetos diferentes de modo a selecionar e executar o mais adequado. Quando realizamos uma descrição abstrata do problema utilizando expressões matemáticas, leis naturais e experiências passadas criamos o que chamamos de modelo matemático do problema. O modelo gerado pode conter muitos projetos alternativos e, portanto, devemos adicionar certos critérios que nos permitam a escolha do projeto adequado. Normalmente, são escolhidos aqueles projetos que otimizam determinada métrica de análise, ou a chamada função objetivo.

O recente avanço computacional permitiu que os modelos matemáticos dos problemas de engenharia se tornassem extremamente complexos, incluindo uma quantidade maior de dados, detalhes e refinamentos. Como exemplo, citamos os recentes avanços nas análises via Método dos Elementos Finitos (MEF) ou Métodos Computacionais para Mecânica dos Fluídos.

O nível de refino nas análises dos modelos, nos conduziu a problemas de otimização que possuem um alto custo computacional associado a funções objetivo de difícil tratamento analítico. Uma classe de problemas que agrava muito a questão do custo computacional são aqueles em que desejamos minimizar integrais multidimensionais, cuja função objetivo dependa da chamada destes códigos de MEF.

Nesta dissertação, estamos interessados na minimização de

funções que são definidas por

$$J(\mathbf{d}) = \int_{\Omega} \psi(\mathbf{d}, \mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}, \quad (1.1)$$

onde  $\mathbf{d} \in \mathbb{R}^k$  é o vetor de projeto com  $k$  dimensões,  $\mathbf{x} \in \mathbb{R}^{k_x}$  é um vetor de parâmetros estocásticos de  $k_x$  dimensões que segue algum tipo de distribuição de probabilidade.  $\psi(\mathbf{d}, \mathbf{x})$  é a medida de performance do sistema a ser otimizado e que possui parâmetros incertos (*e.g.* custo, probabilidade de falha, peso, etc.).  $f_{\mathbf{x}}(\mathbf{x})$  é uma função peso conhecida, normalmente tomada como a função densidade de probabilidade conjunta das variáveis aleatórias do vetor  $\mathbf{x}$  e  $\Omega \subseteq \mathbb{R}^{k_x}$  é o domínio de integração (*e.g.* os limites para a distribuição de probabilidade).

O problema de minimização que resolveremos é considerado sem restrições e dado por

$$\min_{\mathbf{d} \in \mathcal{D}} J(\mathbf{d}), \quad (1.2)$$

onde  $\mathcal{D} \subset \mathbb{R}^k$  representa o domínio de busca das variáveis de projeto. Dessa forma nosso problema fica limitado a restrições somente do tipo caixa.

Esse tipo de problema de otimização é encontrado na maximização da performance esperada de um sistema mecânico, amplamente aplicado em otimização robusta e apresentada por [Capiez-Lernout e Soize \(2008\)](#), [Soize, Capiez-Lernout e Ohayon \(2008\)](#), [Ritto et al. \(2011\)](#), [Lopez et al. \(2014\)](#), [Miguel, Miguel e Lopez \(2016\)](#), [Miguel et al. \(2016\)](#). Nessa classe de problemas, é realizada a adição das incertezas sobre os parâmetros de projeto, e conforme à quantificação das incertezas vai sendo considerada, o modelo matemático que representa a performance do sistema se torna mais realista. Como consequência, a performance do sistema será medida pela esperança matemática do valor da função objetivo, levando em consideração as variabilidades dos parâmetros. Normalmente as variabilidades são medidas via distribuições de probabilidade e a esperança matemática é calculada via integração no domínio de variação dos parâmetros. Como exemplo, podemos tomar a [Figura 1](#)



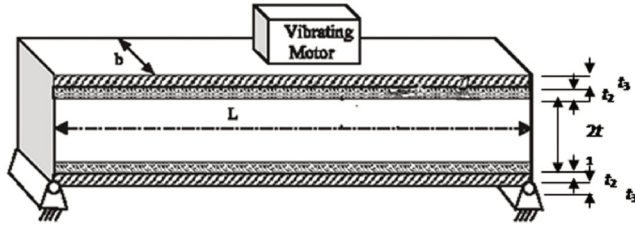


Figura 1 Plataforma simplesmente apoiada sujeita a vibrações.  
(Fonte: [Bhattacharjya e Chakraborty \(2018\)](#))

que apresenta uma viga sanduíche, simplesmente apoiada, e que suporta um motor vibratório. O projeto deve ser escolhido de modo que as dimensões da viga sanduíche resultem no menor custo e satisfaça uma frequência natural mínima. As incertezas associadas à viga estão em todas as propriedades materiais de todas as camadas do sanduíche.

As integrais multidimensionais aqui tratadas também surgem nos problemas de otimização utilizando a abordagem de Projeto Baseado em Performance (*Performance Based Design* - PBD) ([BECK; KOUGIOUMTZOGLOU; SANTOS, 2014; SPENCE; KAREEM, 2014; BOBBY; SPENCE; KAREEM, 2016](#)). O PBD pode ser definido como o projeto, avaliação e construção de obras de engenharia que atendam, de forma mais econômica possível, demandas futuras incertas de seus proprietários e usuários assim como as ações acidentais naturais (e.g. vento, ondas, terremoto) e causadas pelo homem (e.g. ataque terrorista, incêndio). Esta é uma filosofia de projeto estrutural que considera a formulação gradativa de dano para estruturas sujeitas a diferentes tipos e níveis de fenômenos (os quais podem ter diferentes características ao longo da vida útil da estrutura) para os quais diversos níveis de desempenho (e.g. conforto, dano leve, dano profundo e colapso) são aceitos com diferentes probabilidades. Na Figura 2 temos um edifício que pode ser analisado considerando a natureza estocástica dos ventos e as incertezas intrínsecas dos materiais que formam a estrutura. A performance desse exemplo ([BOBBY; SPENCE; KAREEM, 2016](#)) está

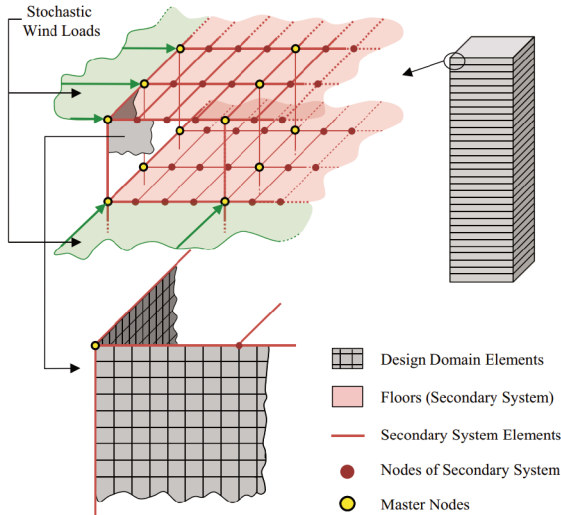


Figura 2 Estrutura sujeita a variabilidade estocástica dos ventos e incerteza nos materiais.

(Fonte: [Bobby, Spence e Kareem \(2016\)](#))

em termos da taxa média anual em que a medida de dano excede um certo limite, indicando que a estrutura entra em um estado de dano (como por exemplo o colapso de uma parede divisória devido ao deslize entre dois andares consecutivos).

Como último exemplo, temos a integral dupla que surge nas aplicações de Projeto Ótimo de Experimento (*Optimal Experimental Design* - OED), estudadas por [Huan e Marzouk \(2013\)](#), [Beck et al. \(2018\)](#) e [Carlon et al. \(2019\)](#). O OED tem por objetivo encontrar os parâmetros de um experimento que maximizem o ganho de informação a partir dos dados medidos. Tem como base a minimização de uma integral dupla que exige a solução de centenas (ou milhares) de problemas probabilísticos inversos. Há interesse de sua aplicação em diversas áreas da engenharia como poços de petróleo, propagação de contaminantes e engenharia aeronáutica. Neste último, pode-se citar a utilização da Tomografia por Impedância Elétrica (*Electrical Impedance Tomography*

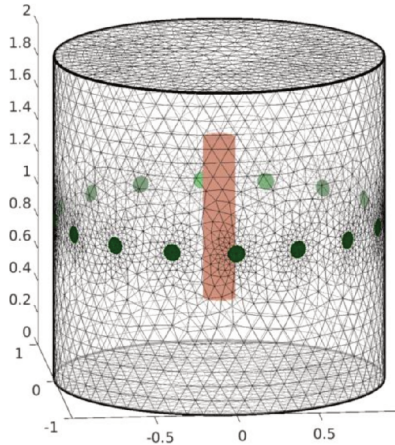


Figura 3 Elemento MEF em 3D com um alvo circular.  
(Fonte: EIDORS (2019))

- EIT) para encontrar trincas/falhas em asas de aviões. Na Figura 3 temos a representação de um experimento de EIT onde monta-se o experimento por meio do OED.

Nos problemas que exemplificamos anteriormente, as funções objetivo a serem otimizadas, possuem na sua grande maioria um comportamento não convexo e multimodal (vários mínimos locais) e ainda podem ser funções do tipo *black-box*, ou seja, funções da qual não se conheça nenhuma informação adicional ao seu valor funcional. Assim, para a solução dos problemas de otimização, deve-se evitar os algoritmos que possuam dependência pelo gradiente da função. Esse tipo de solução além de necessitar da informação adicional das derivadas da função objetivo, assumem uma condição de serem algoritmos de pesquisa local pela solução. Indo além, devido ao seu alto custo computacional, não recomenda-se o uso de algoritmos que utilizem de muitas avaliações da função objetivo, como por exemplo os algoritmos heurísticos.

Então, os problemas de otimização que visamos resolver envolvem: integrais multidimensionais que requerem alto custo computacional

para sua avaliação, e levam a funções objetivos não convexas e multimodais. Uma classe de algoritmos aptos a tratar de tais problemas é a Otimização Global Eficiente (*Efficient Global Optimization* - EGO), desenvolvido por Jones, Schonlau e Welch (1998). O algoritmo base do EGO começa construindo um metamodelo a partir de um plano inicial de pontos. Esse metamodelo substitui a função objetivo realizando predições sobre seus valores, bem como associando um erro a essas predições. Essas duas informações são utilizadas para se adicionar novos pontos no plano, realizando uma busca pelo mínimo global. A forma na qual os novos pontos são adicionados é o que difere os vários métodos de EGO, e tem suma importância na sua eficiência e robustez.

O trabalho de Jones, Schonlau e Welch (1998) foi idealizado para funções que não possuem incertezas. Porém, para que possamos utilizar a mesma abordagem nos nossos problemas, devemos utilizar um método que leve em consideração o erro associado à aproximação das funções integrais via simulação, uma vez que elas não podem ser resolvidas analiticamente. Logo, utilizamos o algoritmo Otimização Global Estocástica Eficiente (*Stochastic Efficient Global Optimization* - sEGO).

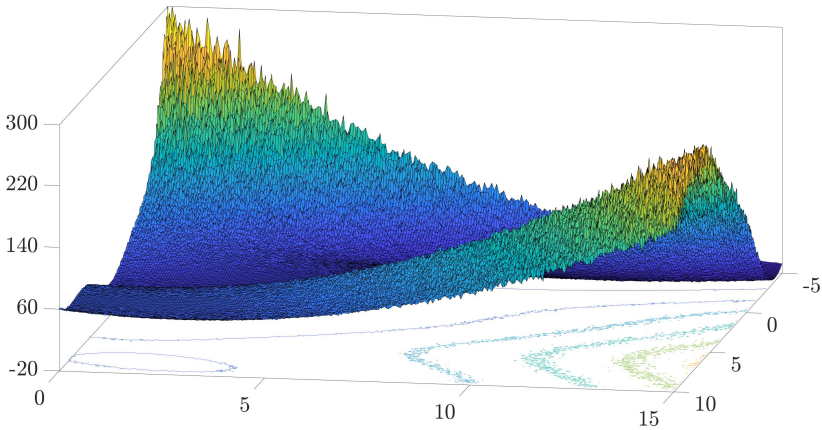
No contexto do sEGO aplicado a minimização de integrais, Carraro et al. (2019) desenvolveram uma forma adaptativa para a aproximação da função objetivo, realizando uma aproximação das integrais por meio da Integração de Monte Carlo (*Monte Carlo Integration* - MCI). Nessa aproximação, a abordagem adaptativa é responsável por ditar qual a qualidade da aproximação em cada ponto do metamodelo, ajustando uma variância alvo para a variância do erro cometido durante as simulações do MCI. Em suas simulações, Carraro et al. (2019) concluíram que uma variância alvo grande poderá fazer com que o processo de otimização fique preso em uma região com altas incertezas da função. Em contrapartida, aproximar a função por meio de uma variância alvo pequena fará com que o custo de MCI seja proibitivo. Portanto, a adaptatividade possui como característica principal realizar um consumo

inteligente do orçamento computacional, dando preferência para que as avaliações sejam gastas em uma boa aproximação dos pontos que possuam uma alta tendência a serem os minimizadores.

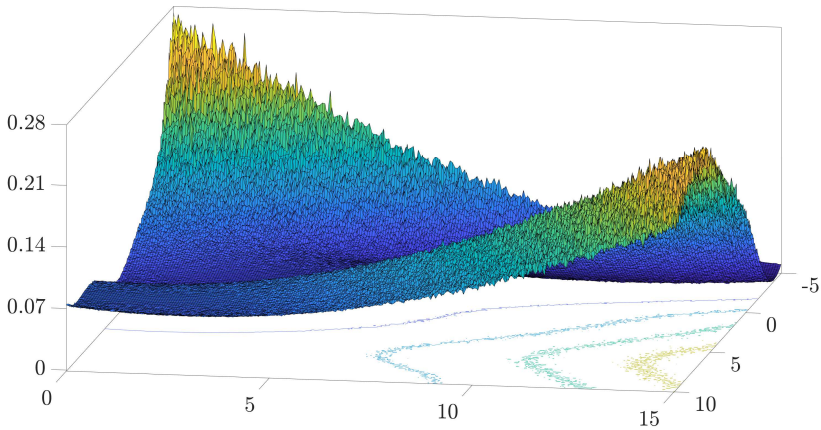
Outra fonte de consumo do orçamento computacional, e que não foi abordado por Carraro et al. (2019), são os diferentes valores que a função objetivo pode assumir. Funções que possuem um contradomínio muito amplo e geram valores discrepantes nas simulações do MCI, podem necessitar de inúmeras avaliações até se atingir a variância alvo. Um exemplo típico desse caso pode ser conferido na Figura 4(a) que apresenta uma função estocástica em duas dimensões.

Portanto, este trabalho é uma extensão do que foi realizado por Carraro et al. (2019), onde o ineditismo proposto será a realização de uma transformação não linear sobre a função objetivo. Chamamos essa transformação não linear de normalização. A ideia principal dessa normalização é não alterar o comportamento estocástico da função objetivo, mas converter o contradomínio original para um contradomínio com valores menores, fazendo com que a variância do erro na aproximação do MCI convirja rapidamente, resguardando o orçamento computacional. A aplicação da normalização pode ser vista na Figura 4(b) que apresenta o mesmo caso bidimensional citado anteriormente. Podemos ver que a superfície da função estocástica continua a mesma, mantendo as mesmas incertezas, porém o valor funcional com normalização pertence a um contradomínio muito menor do que o original.

Espera-se que o sEGO utilizando a abordagem normalizada seja capaz de adicionar mais pontos durante sua execução, do que o mesmo sEGO utilizado por Carraro et al. (2019). Outro ponto que cabe destaque, é que a eficiência e robustez do sEGO estão diretamente associados às métricas que são utilizadas para a adição de pontos. Como propomos uma nova abordagem normalizada e adaptativa para o sEGO, é realizada neste trabalho a análise de cinco diferentes técnicas de adição de pontos propostas na literatura. Cada técnica possui características únicas em como o erro associado ao valor da função é utilizado para



(a) Sem normalização



(b) Com normalização

Figura 4 – Função estocástica bidimensional.

a pesquisa de pontos; dessa forma, somos capazes de testar técnicas que prezam por uma busca local e técnicas que possuem uma maior exploração do domínio.

Os testes realizados são capazes de dizer como cada técnica se comporta no cenário estocástico, além de mostrar que com um maior número de pontos adicionados, temos uma maior exploração

(local ou global) do domínio de busca. Dessa forma, somos capazes de concluir que aplicando a normalização, aumentamos a robustez ao algoritmo, fornecendo ferramentas que o façam resolver os problemas aqui propostos, com mais chances de se obter o mínimo global. Também somos capazes de realizar uma análise estatística para ditar qual técnica de adição de pontos possui melhor comportamento robusto, prezando por valor mínimo obtido e menor variabilidade nos resultados.

## 1.2 OBJETIVOS

### 1.2.1 Objetivo Geral

O objetivo geral deste trabalho é aumentar a eficiência e robustez do sEGO adaptativo proposto por [Carraro et al. \(2019\)](#).

### 1.2.2 Objetivos Específicos

Para alcançarmos o objetivo geral deste trabalho precisaremos atingir com sucesso alguns outros objetivos. São eles:

- Estudar o uso de metamodelos estocásticos combinados com a variância do erro na aproximação via MCI;
- Implementar os algoritmos sEGO;
- Realizar um estudo gráfico e mais aprofundado, focando no comportamento e descrição das métricas de adição de pontos;
- Investigar o efeito de diferentes métricas de adição de pontos na performance do sEGO;
- Analisar as consequências da abordagem normalizada para os metamodelos estocásticos;
- Investigar a utilização do tunelamento estocástico para aumento da eficiência do sEGO;
- Verificar a eficácia do método proposto para funções de variadas dimensões.

### 1.3 ORGANIZAÇÃO DA DISSERTAÇÃO

Para que possamos fornecer a melhor estrutura possível deste manuscrito, a parte que contempla a revisão bibliográfica é composta por três capítulos. O capítulo 2 apresenta uma rápida introdução aos problemas de otimização e como eles são definidos (Seção 2.1). Mostramos também algumas classificações importantes para o conceito de otimização global (Seções 2.2 e 2.3). Finalizamos o capítulo apresentado a nomenclatura dos algoritmos mais utilizados para resolução de tais problemas (Seção 2.4).

No Capítulo 3 temos a introdução à criação dos Metamodelos. Começamos por apresentar a classe de metamodelos mais simples e tradicionais que são os polinomiais (Seção 3.1). Em seguida são apresentados os metamodelos baseados nos Processos Gaussianos por meio das Funções de Bases Radiais (*Radial Basis Functions*) (Seção 3.2). Dessa forma, conduzimos de forma clara e didática o leitor para a construção do Kriging determinístico (Seção 3.3) elucidando todas as etapas fundamentais para sua determinação, entre elas a definição do modelo, o ajuste dos parâmetros via maximização da verossimilhança, a predição com o Kriging e por fim o erro cometido na predição. Fechamos esse importante capítulo apresentando o EGO (Seção 3.4).

O Capítulo 4, apresenta todas as etapas da construção do algoritmo que resolve os problemas de otimização propostos neste texto. Começamos o capítulo definindo, de forma matemática, como serão os problemas e as funções integrais (Seção 4.1). Apresentamos também a aproximação via MCI e o cálculo da variância do erro cometido na aproximação. Na sequência, apresentamos os fundamentos do Kriging Estocástico (Seção 4.2), onde apresentamos a extensão do Kriging interpolador para o regressor, construímos a predição do metamodelo e calculamos os parâmetros de ajuste via maximização da verossimilhança. Na seção seguinte são apresentadas as métricas de adição de pontos que farão parte do algoritmo sEGO (Seção 4.3), as quais serão ajustadas para fazer parte de todo o organismo do método. Após, apresentamos a



abordagem via variância adaptativa que será utilizada (Seção 4.4). Por fim introduzimos a normalização (Seção 4.5) e fazemos uma análise das principais características de cada métrica de adição de pontos, dentro de todo o contexto abordado (Seção 4.6).

O Capítulo 5 apresenta a metodologia proposta e os resultados obtidos por esta dissertação. Começamos por definir as nove funções objetivo que farão parte dos problemas de minimização propostos (Seção 5.1). Na sequência definimos os parâmetros mais importantes que fazem com que todo o algoritmo implementado funcione (Seção 5.2). Começamos os resultados por meio de uma comparação entre as abordagens do sEGO com e sem normalização (Seção 5.3). Na próxima seção mostramos como o valor da variância inicial da abordagem adaptativa influencia o processo de otimização e como a normalização atua para diminuir essa influência (Seção 5.4). Para finalizar os resultados, apresentamos duas técnicas que classificam, da melhor para a pior, as métricas estudadas de adição de pontos (Seção 5.5).

Para encerrar o manuscrito, o Capítulo 6 apresenta duas seções. Na primeira temos todas as conclusões que tiramos sobre os resultados analisados e a abordagem proposta. Na segunda apresentamos pontos de estudo que julgamos serem relevantes para trabalhos futuros na área.

Na parte pós textual do trabalho apresentamos quatro apêndices. O Apêndice A é responsável por apresentar a demonstração da função de verossimilhança utilizada para a obtenção dos parâmetros de ajuste do Kriging determinístico e estocástico. No Apêndice B apresentamos as expressões das derivadas de primeira ordem para a função de verossimilhança do Kriging Estocástico. O Apêndice C tem por função apresentar algumas técnicas numéricas que foram utilizadas na implementação do código computacional desta dissertação. Por fim, apresentamos no Apêndice D algumas tabelas com os valores estatísticos obtidos na seção de resultados.



## 2 PROBLEMAS DE OTIMIZAÇÃO

Problemas de otimização surgem naturalmente em diferentes áreas do conhecimento. Como exemplos temos: um engenheiro estrutural que projeta um edifício com vários andares deve escolher materiais e proporções para diferentes componentes estruturais no edifício, a fim de obter uma estrutura segura que seja a mais econômica possível. O controlador de produção de uma fábrica deve programar as operações da linha de produção de modo que a fábrica produza produtos que maximizem as receitas da empresa, atendendo às demandas dos clientes por diferentes produtos e permanecendo dentro das limitações de recursos disponíveis.

Segundo [Bhatti \(2012\)](#) os problemas de otimização possuem três características em comum:

1. Existe uma meta ou objetivo geral para a atividade;
2. Além do objetivo geral, geralmente há outras exigências, ou restrições, que devem ser satisfeitas;
3. Implícito em todos os problemas está a noção de que existem escolhas disponíveis que, quando feitas adequadamente, atenderão às metas e exigências.

Assim, a otimização pode ser definida como a ciência de determinar as “melhores” soluções para certos problemas matematicamente definidos, que são frequentemente modelos da realidade física ([FLETCHER, 1987](#)). De forma geral, deve-se determinar as variáveis de projeto, de maneira que a função objetivo possua seu melhor valor.

### 2.1 MODELO PADRÃO DE OTIMIZAÇÃO

De acordo com [Arora \(2017\)](#) um modelo matemático geral pode ser proposto para problemas de otimização. Esse modelo é chamado de *Modelo Padrão de Otimização* e é proposto para a minimização de uma

função objetivo, onde as variáveis de projeto devem satisfazer condições de restrição de igualdade (=) e desigualdade (do tipo  $\leq$ ). Portanto um problema geral em otimização deve ser posto como: Determine o vetor de variáveis de projeto  $\mathbf{d} \in \mathbb{R}^k$  dado por

$$\mathbf{d} = \{d_1, d_2, \dots, d_k\}^T, \quad (2.1)$$

que minimiza a função objetivo  $f : \mathbb{R}^k \times \mathbb{R}^l \rightarrow \mathbb{R}$  definida por

$$f_{\text{obj}} = f(\mathbf{d}, \mathbf{x}), \quad (2.2)$$

onde o  $\mathbf{x}$  é um vetor de parâmetros da forma

$$\mathbf{x} = \{x_1, x_2, \dots, x_l\}^T, \quad (2.3)$$

e sujeita as seguintes restrições

$$g_i(\mathbf{d}, \mathbf{x}) \leq 0; \quad i = 1, 2, \dots, m \quad (2.4)$$

$$h_j(\mathbf{d}, \mathbf{x}) = 0; \quad j = 1, 2, \dots, p, \quad (2.5)$$

onde  $g_i : \mathbb{R}^k \times \mathbb{R}^l \rightarrow \mathbb{R}$ ,  $h_j : \mathbb{R}^k \times \mathbb{R}^l \rightarrow \mathbb{R}$ ,  $m$  e  $p$  representam, respectivamente, as funções que definem as restrições de desigualdade e igualdade, e o número de restrições de desigualdade e igualdade. Problemas de maximização são convertidos facilmente tomando o negativo do valor de uma função objetivo a minimizar. [Arora \(2017\)](#) basicamente apresenta o mesmo problema; porém, não há em sua formulação a consideração do vetor de parâmetros  $\mathbf{x}$ . Levamos essa informação para nossos problemas, pois usualmente, esses parâmetros são a quantificação das incertezas nos problemas propostos neste trabalho.

Quando um problema de otimização puder ser colocado nessa forma padrão, diremos que se trata de um *Problema Com Restrições*. Porém, em algumas aplicações práticas poderão surgir problemas de otimização onde deve-se minimizar uma função objetivo  $f(\mathbf{d})$  sem quaisquer restrições sobre o vetor de variáveis  $\mathbf{d}$ . Neste caso, o problema é chamado de *Problema Sem Restrições* e sua forma padrão é dada por

$$\min_{\mathbf{d} \in \mathbb{R}^k} f(\mathbf{d}). \quad (2.6)$$

Se estivermos trabalhando em um problema com restrições, dizemos que, um vetor de variáveis de projeto  $\mathbf{d}$ , é uma solução viável, se ele satisfizer todas as restrições do problema. Caso uma ou mais restrições sejam violadas, então ele será chamado de solução inviável. A região viável do problema de otimização é o conjunto de todos os vetores de projeto viáveis.

## 2.2 CONVEXIDADE

Os conceitos de conjuntos convexos e funções convexas são de grande importância para problemas de otimização. Na análise desses problemas, utiliza-se da convexidade para caracterização das soluções ótimas e para desenvolvimento de procedimentos computacionais (BAZARAA; SHERALI; SHETTY, 2006).

Por definição, um conjunto  $\mathcal{D} \in \mathbb{R}^k$  é convexo se, e somente se, dados  $\mathbf{d}_1, \mathbf{d}_2 \in \mathcal{D}$ , para qualquer  $\Lambda \in [0, 1]$  temos

$$(1 - \Lambda)\mathbf{d}_1 + \Lambda\mathbf{d}_2 \in \mathcal{D}.$$

Portanto, se  $\mathcal{D}$  é conjunto convexo, então um segmento de linha que une dois pontos quaisquer do conjunto está contido em  $\mathcal{D}$ .

Uma função  $f(\mathbf{d})$ , definida sobre um conjunto convexo  $\mathcal{D}$ , será convexa se, e somente se,

$$f((1 - \Lambda)\mathbf{d}_1 + \Lambda\mathbf{d}_2) \leq (1 - \Lambda)f(\mathbf{d}_1) + \Lambda f(\mathbf{d}_2).$$

Um exemplo de função convexa por regiões é mostrada na Figura 5. Pode-se notar que em  $I_1 = [0.5, 6]$  a função é convexa, pois o segmento de linha que une os pontos  $d_1$  a  $d_2$  se mantém acima da curva da função para todo ponto em  $I_1$ . Porém em  $I_2 = [0.5, 17]$  ela não é convexa, pois o segmento que liga  $d_1$  a  $d_3$  não permanece sempre acima da curva da função para todo ponto em  $I_2$ .

Um problema de otimização convexo é aquele em que a região viável, a função objetivo e as restrições de desigualdade são convexas, e as restrições de igualdade são lineares.

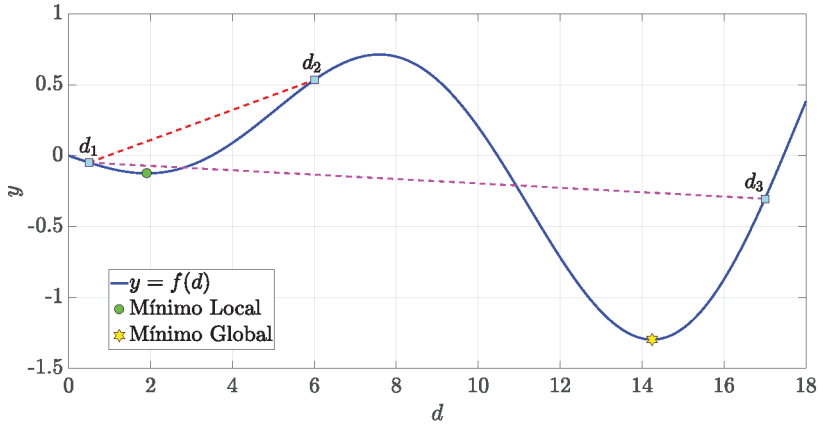


Figura 5 Convexidade de uma função 1D.

### 2.3 ÓTIMO LOCAL E GLOBAL

A distinção entre valores ótimos locais e globais é de extrema importância quando tratamos de funções que não são convexas em seus domínios. Uma função  $f(\mathbf{d})$  de  $k$  variáveis possui um mínimo local em  $\mathbf{d}^*$  se  $f(\mathbf{d}^*) \leq f(\mathbf{d})$  para todo  $\mathbf{d}$  em uma pequena vizinhança de  $\mathbf{d}^*$  dentro do conjunto viável  $\mathcal{D}$ . Caso  $f(\mathbf{d}^*) \leq f(\mathbf{d})$  para todo  $\mathbf{d}$  dentro da região viável  $\mathcal{D}$ , então temos um mínimo global em  $\mathbf{d}^*$  (ARORA, 2017).

Pode-se observar que a função apresentada na Figura 5 é não convexa em  $I = [0, 18]$  e, pela definição acima, possui um mínimo local em  $d_1^* = 1,9097$  e um mínimo global em  $d_2^* = 14,2470$ .

Se vamos resolver um problema de otimização, então estamos interessados em levar a função objetivo ao seu melhor valor possível, ou seja, até seu ótimo global. Se considerarmos problemas de otimização convexas, uma propriedade fundamental é que qualquer ótimo local é também um ótimo global. Essa propriedade faz com que esses problemas sejam de alguma maneira fáceis de se analisar, bem como de garantir que a solução seja em um ótimo global (LUENBERGER, 1969). Porém, uma grande dificuldade surge em problemas que são não lineares ou não

convexos, pois a existência de vários mínimos locais dificulta consideravelmente a determinação do mínimo global (KAGAN et al., 2009). Tais problemas são chamados de multimodais.

Garantir que uma determinada solução do problema seja ótima se torna uma tarefa quase impossível se estamos trabalhando com problemas não convexos, como o caso da Figura 5. Nesses casos uma alternativa seria a de realizar uma busca exaustiva sobre o espaço viável de solução. Porém, computacionalmente, isso se torna uma tarefa intratável. Portanto, como muitas funções objetivo na área de engenharia são não convexas, necessitamos cuidado ao utilizar algoritmos de otimização para que os mesmos não nos deixem confinados em regiões de um ótimo local.

## 2.4 CARACTERIZAÇÃO DOS PROBLEMAS DE OTIMIZAÇÃO

Os algoritmos de otimização funcionam de forma iterativa. Eles começam com um valor inicial de projeto  $\mathbf{d}_0$  e geram uma sequência de melhores valores mínimos a cada iteração, até que se atinja um critério de parada, e com certa esperança, uma solução viável (NOCEDAL; WRIGHT, 2006). O que torna cada algoritmo único é a estratégia utilizada para passar de uma iteração para outra melhorando o valor mínimo atual.

A grande maioria dos otimizadores existentes utilizam os valores da função objetivo, os valores das restrições (caso existam) e possivelmente informações advindas das derivadas de primeira e segunda ordem das funções. Caso a função objetivo ou as restrições sejam formadas por funções das quais se conheça apenas seu valor em alguns pontos de  $\mathcal{D}$ , e não seja possível obter nenhuma outra informação, tais como as derivadas, então essas funções serão chamadas de funções “*black-box*”.

### 2.4.1 Tipos de problemas

De acordo com Nocedal e Wright (2006) em alguns problemas de otimização o modelo padrão de problema não será completamente

especificado, pois poderá existir uma dependência de quantidades que são desconhecidas no momento de sua formulação. A presença de parâmetros incertos (ou randômicos) faz com que os algoritmos de otimização se comportem de maneiras diferentes. Assim, duas classificações quanto aos algoritmos são realizadas:

- ▶ **Determinísticos:** São os tradicionais problemas de programação matemática, onde não há incerteza sobre nenhum parâmetro do problema. Nesse contexto os algoritmos trabalham melhorando a solução iterativamente. Quando executado diversas vezes a partir do mesmo ponto, os algoritmos sempre deverão encontrar o mesmo resultado;
- ▶ **Estocásticos:** Classe de problemas que possuem parâmetros incertos, e que em sua grande maioria serão determinados por distribuições de probabilidades definidas por uma média e variância do valor. Nesse caso a incerteza gera ruídos nos valores da função objetivo ou das restrições, portanto um mesmo processo de otimização poderá levar a resultados diferentes (mesmo se executado a partir do mesmo ponto).

### 2.4.2 Algoritmos determinísticos de Otimização

A maioria dos algoritmos para problemas de otimização possuem características que os classificam como otimizadores locais, ou seja, são capazes de encontrar um mínimo local, e nem sempre encontram a solução global. Para problemas que são convexos todas as soluções locais são também soluções globais, porém problemas multimodais, com e sem restrições, podem conter vários mínimos locais e somente um mínimo global (NOCEDAL; WRIGHT, 2006).

Os algoritmos determinísticos foram os pioneiros na solução de tais problemas, onde os mais importantes são aqueles que fazem uso do valor da função objetivo e restrições, bem como das informações sobre suas derivadas. O sucesso da otimização está diretamente associado



ao comportamento das funções e ao ponto inicial de iteração, onde diferentes mínimos poderão ser encontrados a partir de pontos distintos.

Alguns algoritmos serão apresentados aqui somente por seus nomes, para um maior aprofundamento acerca dos mesmos sugere-se ao leitor [Luenberger \(1969\)](#), [Fletcher \(1987\)](#), [Bazararaa, Sherali e Shetty \(2006\)](#), [Bhatti \(2012\)](#), [Izmailov e Solodov \(2012\)](#) e [Arora \(2017\)](#).

Para problemas de otimização sem restrições os algoritmos com maior destaque são:

- Pesquisa em Linha sem utilizar derivadas. Alguns algoritmos dessa classe são: Método da Seção Áurea e Pesquisa de Fibonacci;
- Pesquisa em Linha utilizando derivadas. Para essa técnica temos os algoritmos: Método da Bisseção e o Método de Newton;
- Método Simplex para problemas lineares;
- Descida mais Íngreme;
- Método do Gradiente Conjugado;
- Direção de Pesquisa: Método de Newton e o Método de Newton Modificado;
- Métodos Quase-Newton: Método DFP e Método BFGS.

Os algoritmos destacados anteriormente, com exceção do Método Simplex, também servem para problemas com restrições; porém, o problema padrão de otimização passa por uma transformação e deixa de ser um problema com restrições para um problema sem restrições. O Método Simplex já é capaz de tratar as restrições lineares em seu algoritmo. Os métodos de transformações mais conhecidos são:

- Método da Penalidade;
- Lagrangeano Aumentado.

Outros dois métodos para se tratar problemas com restrições são

- Programação Quadrática Sequencial;
- Métodos de Pontos Interiores.

Alguns pontos negativos são de importante destaque acerca esses algoritmos de otimização:

- ▷ São extremamente sensíveis ao valor inicial  $\mathbf{d}_0$ , podendo o valor final ser mínimo local, mínimo global, ou nenhum dos dois;
- ▷ Podem realizar um alto número de avaliações das funções até que se atinja um critério de parada estabelecido;
- ▷ Possuem, na sua grande maioria, a dependência do gradiente e da hessiana das funções;
- ▷ São de difícil aplicação a problemas com ruídos ou erros.

Podemos citar como pontos positivos:

- ▷ Possuem uma rápida convergência de acordo com os parâmetros iniciais;
- ▷ Não dependem de parâmetros randômicos, portanto a cada execução obtemos o mesmo valor;
- ▷ Possuem uma rápida implementação se comparado aos métodos heurísticos.

### 2.4.3 Algoritmos Heurísticos

Apesar de bem consolidados, os algoritmos determinísticos podem falhar na pesquisa global além de, na grande maioria, precisar dos cálculos do gradiente e da hessiana de suas funções. Logo, esses

algoritmos devem ser evitados para problemas formados por funções do tipo *black-box*, ou mesmo funções cujas derivadas são de difícil obtenção.

Para contornar esses problemas, e principalmente para solucionar a obtenção de mínimos globais, nas últimas décadas muitos algoritmos foram desenvolvidos inspirados em fenômenos naturais (ARORA, 2017). Normalmente esses métodos utilizam apenas o valor da função objetivo, não sendo influenciados por sua suavidade ou continuidade.

Os algoritmos heurísticos são métodos iterativos e estocásticos, onde durante a etapa de melhora, as decisões e transformações da solução são tomadas por números randômicos, o que torna cada execução do método única, de tal forma que não temos garantia de obter o mesmo valor mínimo em cada simulação.

Abaixo são listados alguns algoritmos dessa classe:

- Algoritmos Genéticos (GOLDBERG, 1989);
- Recozimento Simulado (KIRKPATRICK; GELATT; VECCHI, 1983);
- Colônia de Formigas (DORIGO, 1992);
- Enxame de Partículas (*Particle Swarm Optimization* - PSO) (KENNEDY; EBERHART, 1995);
- Pesquisa em Grupo (*Search Group Algorithm* - SGA) (GONÇALVES; LOPEZ; MIGUEL, 2015).

Entre os pontos negativos dos algoritmos heurísticos temos o grande número de avaliações das funções que devem ser realizadas, até para problemas de pequena dimensão. Nesses casos, a depender da complexidade das funções, o tempo computacional pode ficar intratável. Outro ponto negativo é que não possuímos nenhuma garantia de que o mínimo global será obtido.

Como pontos positivos temos a não dependência de derivadas e de hessianas, tratamento de funções descontínuas e a otimização de problemas com funções objetivo do tipo *black-box*.

Todos os métodos apresentados nesta seção, são capazes de resolver um problema de otimização, com ou sem restrições, pelo menos no âmbito de solução local. Para a solução global, nenhum deles possui convergência garantida caso a função seja não convexa. Outro empecilho para o uso destes algoritmos, nos problemas que desejamos resolver, é que em nenhum deles há um tratamento especial quanto às variáveis estocásticas. Quando tais algoritmos são aplicados em casos estocásticos, na maioria dos casos é utilizado uma média dos valores funcionais ao invés de se levar em consideração a variabilidade da função para sua avaliação. Essa abordagem eleva consideravelmente o o número de avaliações das funções objetivo necessário para a convergência, o que consequentemente aumenta o tempo computacional de solução.

Portanto, como os problemas propostos por este texto possuem, na sua maioria, funções estocásticas não convexas, caras computacionalmente de serem avaliadas e que nem sempre possuem um tratamento analítico simples, é necessário que outras formas de solução sejam propostas. Serão levados em consideração métodos que prezem pela busca global da solução e que requeiram um baixo número de avaliações da função objetivo.

### 3 METAMODELOS

Neste capítulo abordaremos as bases fundamentais dos meta-modelos, que são aplicados no cenário proposto por esta dissertação. Sua escrita preza por um detalhamento extenso nas bases do Kriging e do EGO, por questões didáticas, e para que se crie uma referência do referido tema em língua portuguesa.

Em muitas aplicações práticas, é desejável que seus modelos matemáticos sejam os mais fiéis possíveis à realidade. Após a construção desse modelo, podemos recair em processos que sejam extremamente complexos e que exijam uma enorme quantidade de cálculo computacional. Por exemplo, imagine uma simulação com o MEF dos efeitos sísmicos em uma estrutura, onde o objetivo é minimizar a probabilidade de falha da estrutura. Nesse caso as funções que envolvem a modelagem matemática, além de possuírem formulações complexas, podem levar dias até serem completamente avaliadas.

A otimização de tais sistemas pode ser difícil, ou até mesmo impossível, se não dispusermos do tempo necessário para sua conclusão, e ainda podemos sofrer os riscos de uma otimização local quando utilizamos os métodos baseados em gradiente comentados anteriormente e nos casos onde são presentes funções não convexas.

Nesse cenário, durante os últimos anos, a modelagem via Superfícies de Resposta, ou também conhecida como Metamodelagem, vem se destacando no papel de realizar uma otimização global. Os Metamodelos (ou superfícies de resposta) são criados pelo ajuste de uma curva ou superfície a um conjunto inicial de Pontos de Suporte, pertencentes ao domínio de busca das variáveis de projeto (HOYLE, 2006). A ideia básica é que o metamodelo aja como uma “curva interpoladora” dos dados disponíveis, de modo que os resultados possam ser previstos sem recorrer ao uso da fonte primária (função objetivo). Essa abordagem baseia-se no pressuposto de que a superfície de resposta, uma vez construída, será muitas ordens de grandeza mais rápida que a

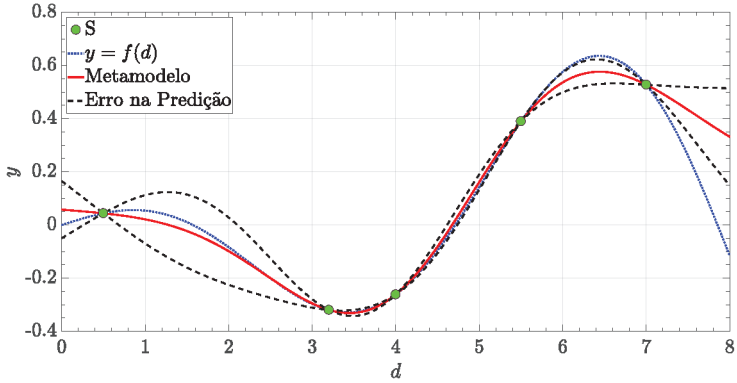
fonte primária, e ainda será útil para prever outros pontos da função (FORRESTER; SOBESTER; KEANE, 2008).

Inicialmente o uso de metamodelos era baseado na abordagem de se construir uma superfície de resposta utilizando uma regressão em base polinomial de segunda ordem, utilizando o método dos mínimos quadrados. Esta abordagem foi extensivamente utilizada em muitos campos da engenharia, como pode ser visto nos trabalhos de Jr (1997), Liu, Haftka e Akgün (2000), Eom et al. (2011), Torii e Lopez (2011) e Torii, Lopez e Biondini (2012). Porém se trata de uma abordagem pouco flexível e, segundo Hussain, Barton e Joshi (2002), possui resultados apreciáveis quando aplicado em otimização local ou em funções convexas. Dessa forma, essa abordagem é imprópria para problemas não lineares e para funções não convexas, que são o objetivo desse trabalho.

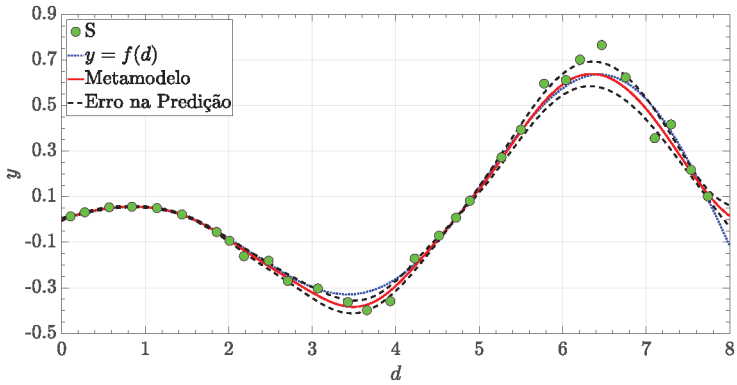
Outros tipos de curvas podem ser utilizadas para a criação/ajuste dos metamodelos, como as Funções de Base Radial (*Radial Basis Function* - RBF) (BROOMHEAD; LOWE, 1988; POWELL, 1987), Kriging (MATHERON, 1963) e Máquinas de Vetores de Suporte (*Support Vector Machines* - SRV) (CRISTIANINI; SHAW-TAYLOR, 2000).

Os metamodelos podem ser construídos com um entre dois propósitos: ser um modelo interpolador da superfície ou ser um modelo em regressão da superfície. O metamodelo é construído a partir de um conjunto de pontos de suporte  $S$ , chamado de espaço amostral. Se considerado interpolador, o metamodelo deverá possuir o mesmo valor da função original em todos os pontos de suporte de  $S$ . No entanto, como regressor, basta que o metamodelo fique o mais próximo possível dos valores da função em  $S$ .

A Figura 6 apresenta os dois tipos de metamodelos que podemos criar, onde nas duas figuras temos o espaço de pontos de suporte  $S$ , a curva pontilhada azul representa a função original, a curva contínua vermelha representa o metamodelo ajustado aos pontos de suporte, e por fim, as curvas pontilhadas pretas representam o valor do metamodelo acrescido (curva superior) ou diminuído (curva inferior) o erro



(a) Interpolador



(b) Regressor

Figura 6 Tipos de metamodelos.

na aproximação. Na Figura 6(a) temos o caso interpolador, onde cada ponto de suporte foi aproximado pelo metamodelo com exatamente o mesmo valor da função original. Podemos ver nesse caso que, o erro na aproximação do metamodelo é nulo nos pontos de suporte. Na Figura 6(b) temos o caso regressor, onde podemos ver que, nos pontos de suporte, o metamodelo não possui exatamente o valor da função original. Podemos ver nesse caso que, o erro na aproximação, não é nulo nos pontos de suporte.

A seguir, apresentamos uma breve introdução à criação de metamodelos interpoladores por polinômios e por RBF. Essa abordagem propicia os requisitos básicos para o tratamento do Kriging, objeto principal de estudo desse trabalho.

### 3.1 METAMODELOS POLINOMIAIS

Seja o espaço amostral  $S = \{\mathbf{d}^{(1)}, \mathbf{d}^{(2)}, \dots, \mathbf{d}^{(n)}\}$  formado por  $n$  pontos de suporte onde cada vetor  $\mathbf{d}^{(i)}$  ( $i = 1, 2, \dots, n$ ) contém  $k$  variáveis,  $\mathbf{d}^{(i)} = \{d_1^{(i)}, d_2^{(i)}, \dots, d_k^{(i)}\}^T$ . A cada um dos  $n$  pontos de suporte, estará associado um valor da função a ser substituída, logo podemos formar o vetor  $\mathbf{y} = \{y^{(1)}, y^{(2)}, \dots, y^{(n)}\}^T$ , onde  $y^{(i)} = f(\mathbf{d}^{(i)})$ .

Os modelos polinomiais podem ser generalizados por

$$\hat{y}(\mathbf{d}) = \sum_{i=1}^m \beta_i h_i(\mathbf{d}) + \epsilon, \quad (3.1)$$

onde  $\mathbf{d}$  é o ponto a ser aproximado pelo metamodelo,  $\beta_i$  são os parâmetros de ajuste do modelo e  $h_i(\mathbf{d})$  são as funções base do espaço vetorial  $\mathcal{P} = \{h_i(\mathbf{d}) \mid i = 1, 2, \dots, m\}$  de todos os polinômios em  $\mathbf{d}$  de grau  $g$ . O termo  $\epsilon$  representa o erro cometido pela aproximação polinomial e é suposto como sendo independente e identicamente distribuído por uma distribuição normal de média zero.

Para a determinação dos parâmetros de ajuste, podemos fazer uso do artifício de que o modelo polinomial possui os mesmos valores da função original nos pontos de suporte. Assim, temos que  $\hat{y}(\mathbf{d}^{(i)}) = y^{(i)}$  para todo  $i = 1, 2, \dots, n$ , e da equação (3.1) temos que

$$\begin{bmatrix} h_1(\mathbf{d}^{(1)}) & h_2(\mathbf{d}^{(1)}) & \dots & h_m(\mathbf{d}^{(1)}) \\ h_1(\mathbf{d}^{(2)}) & h_2(\mathbf{d}^{(2)}) & \dots & h_m(\mathbf{d}^{(2)}) \\ \vdots & \vdots & \ddots & \vdots \\ h_1(\mathbf{d}^{(n)}) & h_2(\mathbf{d}^{(n)}) & \dots & h_m(\mathbf{d}^{(n)}) \end{bmatrix} \begin{Bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{Bmatrix} = \begin{Bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{Bmatrix}, \quad (3.2)$$

ou em forma matricial temos que

$$\Psi \boldsymbol{\beta} = \mathbf{y}, \quad (3.3)$$



onde  $\Psi$  é uma matriz retangular de ordem  $n \times m$ . De (3.3) podemos determinar o vetor de parâmetros de ajuste utilizando a pseudo inversa de Moore-Penrose (PENROSE, 1955), logo

$$\beta = (\Psi^T \Psi)^{-1} \Psi^T \mathbf{y}. \quad (3.4)$$

Caso tenhamos  $m = n$ , então  $(\Psi^T \Psi)^{-1} \Psi^T = \Psi^{-1}$ .

A solução obtida por 3.4 é a estimativa dos parâmetros de ajuste do modelo pelo método dos mínimos quadrados (MMQ) ou também vista como a minimização da média da soma dos erros quadrados,  $\epsilon^2$ , dado por

$$\text{MMQ} = \frac{1}{n} \sum_{i=1}^n \left( y^{(i)} - \hat{y}(\mathbf{d}^{(i)}) \right)^2. \quad (3.5)$$

A Figura 7 apresenta uma aproximação via metamodelos polinomiais. A função da qual se deseja a aproximação é  $f : [0, 8] \rightarrow \mathbb{R}$  dada pela lei

$$f(d) = 0.1d \cos(ad), \quad (3.6)$$

onde  $a$  é um parâmetro igual a 1 para a figura da esquerda e igual a 2 para a figura da direita. Na figura da esquerda temos a aproximação via polinômios de graus 3, 5 e 8 e na figura da direita temos apenas a aproximação para polinômios de grau 8.

Pode-se ver que para uma função mais comportada (esquerda) os polinômios de grau 5 e 8 já conseguem uma boa aproximação para a função. Entretanto, conforme a função fica mais multimodal (direita) tem-se a necessidade de aumentar o grau dos polinômios de aproximação ou aumentar o número de pontos de suporte. Porém essa solução não é uma alternativa satisfatória. Polinômios de ordens maiores tendem a oscilar com grandes amplitudes nos extremos dos pontos de suporte, logo por essa razão eles não são considerados boas soluções para modelos substitutos (SASENA, 2002).

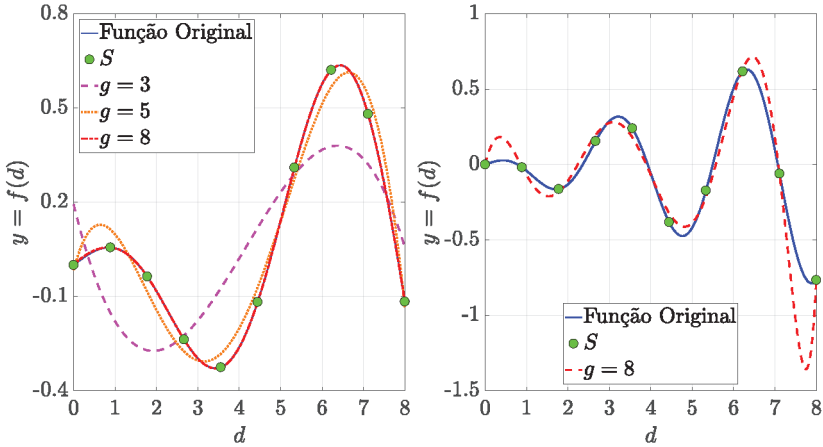


Figura 7 Aproximação em metamodelos polinomiais.

### 3.2 FUNÇÕES RBF

Os metamodelos construídos a partir das RBF diferem dos modelos polinomiais apenas na escolha das funções de base para a aproximação. Como o nome sugere as bases são funções dependentes da distância Euclidiana entre o ponto a ser predito (estimado) e os pontos de suporte. Assim, dado um espaço amostral  $S$  formado por  $n$  pontos  $\mathbf{d}$ , um metamodelo via aproximação por RBF será dado por

$$\hat{y}(\mathbf{d}) = \sum_{i=1}^n \beta_i h(\|\mathbf{d} - \mathbf{d}^{(i)}\|), \quad (3.7)$$

onde  $\beta_i$  são os parâmetros de ajuste do modelo,  $h$  serão as funções de base radial e  $\|\cdot\|$  representa a distância ou norma Euclidiana.

Existem diferentes escolhas para o tipo de função de base  $h$ , entre as mais tradicionais se encontram:

$$\begin{aligned} h(\|\mathbf{d} - \mathbf{d}^{(i)}\|) &= \|\mathbf{d} - \mathbf{d}^{(i)}\| && \text{Linear} \\ &= \|\mathbf{d} - \mathbf{d}^{(i)}\|^3 && \text{Cúbica} \\ &= \|\mathbf{d} - \mathbf{d}^{(i)}\|^2 \log(\|\mathbf{d} - \mathbf{d}^{(i)}\|) && \text{Thin-Plate Spline} \end{aligned}$$

$$\begin{aligned}
&= \exp\left(\frac{\|\mathbf{d} - \mathbf{d}^{(i)}\|^2}{2\sigma^2}\right) && \text{Gaussiana} \\
&= (\|\mathbf{d} - \mathbf{d}^{(i)}\|^2 + \sigma^2)^{1/2} && \text{Multiquadrática} \\
&= (\|\mathbf{d} - \mathbf{d}^{(i)}\|^2 + \sigma^2)^{-1/2} && \text{Multiquadrática Inversa.}
\end{aligned}$$

Caso seja escolhida uma entre as três primeiras funções de aproximação, a estimação dos parâmetros de ajuste se dará da mesma forma que na polinomial, pelas equações (3.3) e (3.4), com a facilidade de  $\Psi$  ser uma matriz quadrada de ordem  $n$ .

Quando uma das três últimas funções for a escolhida, então ocorrerá a adição de mais um parâmetro de ajuste,  $\sigma$ , e com isso a utilização da abordagem que leva a (3.3) não será possível. Nesses casos, os parâmetros deverão ser estimados de outra maneira, e a que será utilizada nesse trabalho será a de se encontrar o máximo da verossimilhança, discutido na Seção 3.3.3. Estimados os parâmetros de ajuste, a escolha correta de  $\beta$  garante que a aproximação possa reproduzir os mesmos valores do espaço amostral  $S$ , enquanto a estimativa correta dos parâmetros adicionais,  $\sigma$ , nos permite minimizar o erro de estimação do modelo (FORRESTER; SOBESTER; KEANE, 2008).

Outro fator chave da utilização das RBF é a possibilidade de uma estimativa dos erros cometidos na previsão do modelo para o ponto  $\mathbf{d}$ . Além disso, quando o objetivo dos metamodelos for o da otimização de uma função objetivo, esse erro nos permite desenvolver uma métrica para a expectativa da melhora no valor mínimo (ou máximo) da função em relação ao seu mínimo (ou máximo) obtido até o momento atual do processo iterativo de otimização (FORRESTER; SOBESTER; KEANE, 2008). Essa métrica será desenvolvida para os conceitos do Kriging como metamodelo, cujas funções de aproximação são formas alternativas da aproximação Gaussiana das RBF.

### 3.3 KRIGING

#### 3.3.1 Introdução ao Kriging Determinístico

**Matheron (1963)** concebeu o termo “*Krigagem*”, em homenagem ao Engenheiro de Minas Sul Africano, Danie Krige, que foi o primeiro a desenvolver o método hoje chamado de “*Kriging*” (**KRIGE, 1951**). A *Krigagem* entrou para o rol de aplicações em projetos de engenharia seguindo o trabalho de **Sacks et al. (1989)**, que aplicou o método à aproximação de experimentos computacionais (**FORRESTER; SOBESTER; KEANE, 2008**).

Enquanto na regressão clássica os coeficientes são calculados para descrever uma função, no Kriging o foco é estimar parâmetros que descrevam como a função tipicamente se comporta (**JONES; SCHONLAU; WELCH, 1998**). A ideia principal por trás do Kriging é a de que os erros cometidos nas predições não são independentes, diferentemente da abordagem na regressão clássica (como a visto nos modelos polinomiais), onde os erros são supostos como independentes, idênticos e seguem uma variável aleatória de distribuição normal (**SASENA, 2002**).

Para ilustração, tomemos a Figura 8, onde a função (3.6) com  $a = 1$  é apresentada junto com seu modelo polinomial de grau 5. Nela destacamos o ponto de suporte  $d^{(i)}$  do qual possuímos o valor da função original  $y^{(i)}$  e o erro no valor predito pela regressão é igual a  $\epsilon(d^{(i)}) = y(d^{(i)}) - \hat{y}(d^{(i)})$ . Seja agora um valor pequeno  $\delta > 0$ , então para o ponto  $d^{(i)} + \delta$ , que não é de suporte, conheceremos apenas o valor da predição  $\hat{y}(d^{(i)} + \delta)$ .

Para os modelos regressivos o erro nesse novo ponto será devido aos pontos de suporte e ao processo de obtenção dos parâmetros de ajuste da aproximação, e não dependerá do valor  $d^{(i)} + \delta$ . Na abordagem pelo Kriging temos que se o erro em  $d^{(i)}$  é grande então o mais correto a se afirmar sobre o erro em  $d^{(i)} + \delta$  é que ele também seja grande, ou conforme  $\delta \rightarrow 0$  temos que  $\hat{y}(d^{(i)} + \delta)$  seja diferente de  $y = f(d^{(i)})$  pelo mesmo erro  $\epsilon(d^{(i)})$ . Isso é consequência sistemática dos erros associados

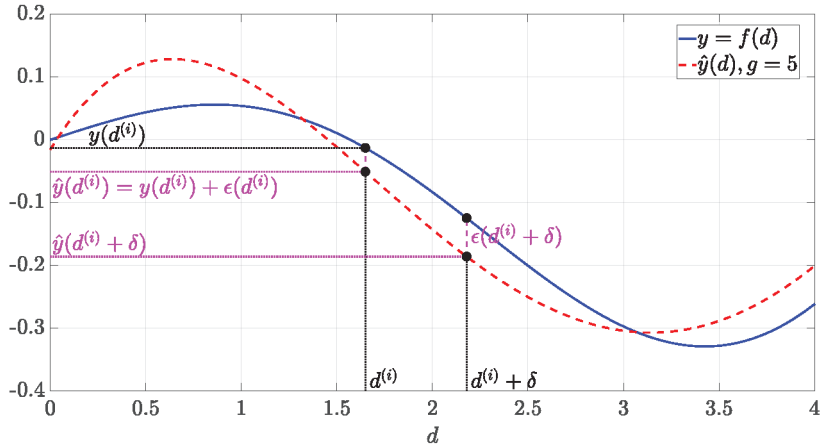


Figura 8 Análise dos erros na regressão.

à forma da função de  $y = f(d)$  (SASENA, 2002), dessa forma o erro poderá ser colocado como função de  $\mathbf{d}$ .

### 3.3.2 As bases de um modelo com o Kriging

A resposta,  $\hat{y}(\mathbf{d})$ , utilizando o Kriging como metamodelo pode ser definida como a combinação de duas funções que analisam as respostas do modelo, a saber

$$\hat{Y}(\mathbf{d}) = M(\mathbf{d}) + Z(\mathbf{d}). \quad (3.8)$$

A função  $M(\mathbf{d})$  representa a tendência de resposta do modelo global, normalmente uma combinação de funções polinomiais (similar ao modelo polinomial). A função  $Z(\mathbf{d})$  fornece uma métrica para a tendência localizada da resposta apresentando as incertezas sobre o valor da predição em  $\mathbf{d}$ . Usualmente a função  $Z(\mathbf{d})$  é definida como a realização de um processo Gaussiano com média zero, variância  $\sigma^2$  e uma covariância não nula entre pontos (SIMPSON et al., 2001) dada por

$$\text{Cov}[Z(\mathbf{d}^{(i)}), Z(\mathbf{d}^{(j)})] = \sigma^2 h(\mathbf{d}^{(i)}, \mathbf{d}^{(j)}), \quad (3.9)$$

onde  $h(\mathbf{d}^{(i)}, \mathbf{d}^{(j)})$  é a função de correlação (ou função de base) entre os pontos  $\mathbf{d}^{(i)}$  e  $\mathbf{d}^{(j)}$ . O valor de  $\sigma^2$  pode ser interpretado como a variância

de  $Z(\mathbf{d})$  para todos  $\mathbf{d}$ , ou seja, representa a variabilidade de incerteza medida por  $Z(\mathbf{d})$ . Normalmente seu valor é considerado constante para todos os pontos  $\mathbf{d}$  e será ajustado de acordo com os pontos de suporte, juntamente com os parâmetros da função de correlação  $h$ . Já a função de correlação  $h$  é a responsável por determinar como o metamodelo irá interpolar os dados e aproximar a função, fornecendo um valor para o grau de dependência entre as  $k$  variáveis de  $\mathbf{d}$ . Da premissa de  $Z(\mathbf{d})$  ser um processo Gaussiano, temos que o modelo adotado trata uma resposta determinística  $y = f(\mathbf{d})$  como sendo uma variável aleatória normalmente distribuída,  $Y(\mathbf{d})$ .

Sendo  $Y(\mathbf{d})$  uma variável aleatória normalmente distribuída que possui uma alta correlação para  $\mathbf{d}$  próximos e uma baixa correlação para  $\mathbf{d}$  afastados, e o termo  $Z(\mathbf{d})$  é construído utilizando um processo Gaussiano, podemos definir a função de correlação ou de base como

$$\text{cor} \left[ Z(\mathbf{d}^{(i)}), Z(\mathbf{d}^{(j)}) \right] = h(\mathbf{d}^{(i)}, \mathbf{d}^{(j)}) = \exp \left( - \sum_{r=1}^k \theta_r \left| d_r^{(i)} - d_r^{(j)} \right|^{p_r} \right). \quad (3.10)$$

Pode-se notar na expressão acima que, se  $\|\mathbf{d}^{(i)} - \mathbf{d}^{(j)}\| \rightarrow 0$  então a correlação tende para 1, enquanto que se  $\|\mathbf{d}^{(i)} - \mathbf{d}^{(j)}\| \rightarrow \infty$  então a correlação tende a zero (SIMPSON et al., 2001; JONES, 2001; FORRESTER; SOBESTER; KEANE, 2008). O fato de  $h(\mathbf{d}^{(i)}, \mathbf{d}^{(i)}) = 1$  é uma das qualidades da função 3.10, pois permite que o preditor do Kriging seja exatamente o valor da função em qualquer ponto de suporte.

A escolha da correlação dada por (3.10) é robusta o bastante para permitir que dois parâmetros,  $\theta_r$  e  $p_r$ , sejam ajustados para cada dimensão  $r = 1, 2, \dots, k$  do ponto de suporte. Isso permite uma visão maior do comportamento da função a ser modelada, onde um dos parâmetros será responsável pela dominância de determinada dimensão e a outra pela suavidade do modelo (HOYLE, 2006).

Suponha o caso  $k = 1$ , colocando  $\theta = 1$  em (3.10) temos que

$$h(d^{(i)}, d) = \exp \left( - \left| d^{(i)} - d \right|^p \right),$$

pela Figura 9(a) podemos ver que o parâmetro  $p$  influencia na suavidade da correlação, onde para  $p = 2$  temos uma correlação bem suave e com gradiente contínuo quando  $|d^{(i)} - d| = 0$ . Ao reduzirmos os valores de  $p$ , aumentamos a taxa na qual a correlação inicialmente começa a diminuir, conforme a distância  $|d^{(i)} - d|$  cresce, e pode-se ver que para pontos que sejam muito próximos podemos obter uma queda imediata da correlação, mas ainda a mantendo igual a 1 para o mesmo ponto.

Ainda para  $k = 1$ , fixando  $p = 2$  temos

$$h(d^{(i)}, d) = \exp\left(-\theta |d^{(i)} - d|^2\right),$$

e pela Figura 9(b) podemos ver que o parâmetro  $\theta$  é uma medida de “importância” da variável  $d$ . Pode-se notar que um valor de  $\theta$  baixo implica que todos os pontos possuem uma alta correlação, com  $Y(\mathbf{d})$  sendo similar em toda a amostra e pontos distantes possuem uma alta influência sobre o ponto a ser predito. Já valores altos de  $\theta$  significarão que os pontos possuirão uma alta correlação somente se estiverem próximos, possuindo uma alta taxa de decréscimo na correlação conforme a distância aumenta. Dessa forma, pontos distantes do ponto a ser predito possuem uma baixa influência sobre o valor predito (JONES; SCHONLAU; WELCH, 1998; JONES, 2001; FORRESTER; SOBES-TER; KEANE, 2008; FORRESTER; KEANE, 2009).

Como possuímos  $n$  pontos de suporte para criar o metamodelo, de (3.10) podemos montar a matriz de correlação da amostra para o metamodelo

$$\Psi = \begin{bmatrix} h(\mathbf{d}^{(1)}, \mathbf{d}^{(1)}) & h(\mathbf{d}^{(1)}, \mathbf{d}^{(2)}) & \dots & h(\mathbf{d}^{(1)}, \mathbf{d}^{(n)}) \\ h(\mathbf{d}^{(2)}, \mathbf{d}^{(1)}) & h(\mathbf{d}^{(2)}, \mathbf{d}^{(2)}) & \dots & h(\mathbf{d}^{(2)}, \mathbf{d}^{(n)}) \\ \vdots & \vdots & \ddots & \vdots \\ h(\mathbf{d}^{(n)}, \mathbf{d}^{(1)}) & h(\mathbf{d}^{(n)}, \mathbf{d}^{(2)}) & \dots & h(\mathbf{d}^{(n)}, \mathbf{d}^{(n)}) \end{bmatrix}, \quad (3.11)$$

e também a matriz de covariância

$$\text{Cov}(\mathbf{Y}, \mathbf{Y}) = \sigma^2 \Psi, \quad (3.12)$$

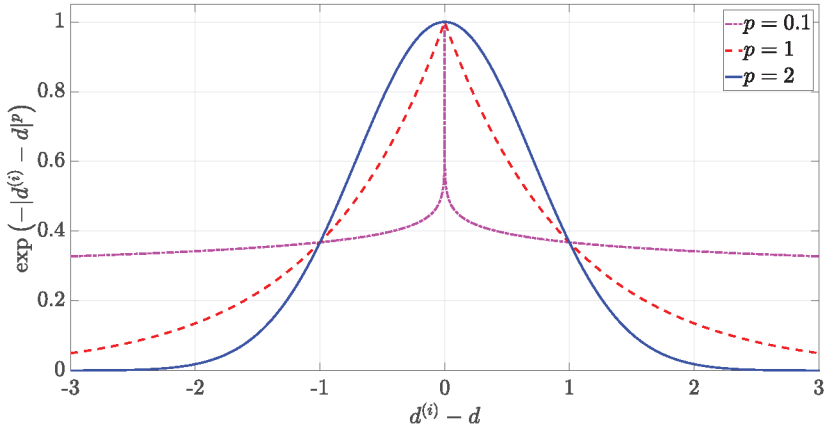
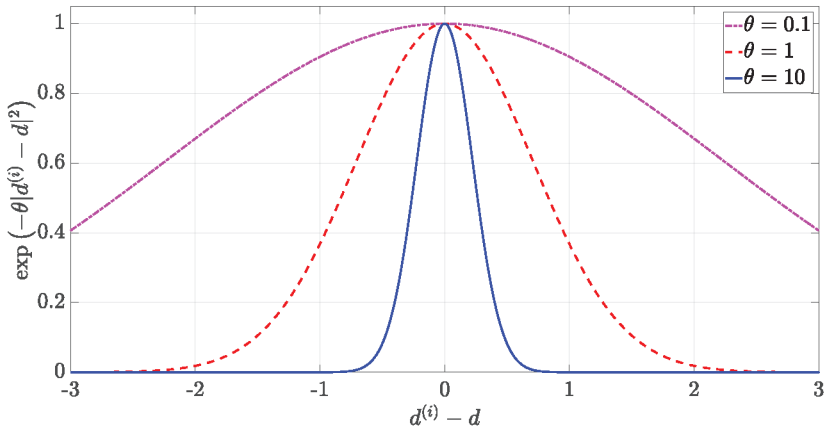
(a)  $p$ (b)  $\theta$ 

Figura 9 Influência dos parâmetros na correlação.

onde

$$\mathbf{Y} = \begin{Bmatrix} Y(\mathbf{d}^{(1)}) \\ \vdots \\ Y(\mathbf{d}^{(n)}) \end{Bmatrix},$$

é o vetor de respostas, onde cada elemento será suposto como normal-



mente distribuído. Faremos a suposição de que esse vetor possui média igual a  $\boldsymbol{\mu} = \mathbf{1}\mu$ , onde  $\mathbf{1}$  é um vetor de dimensão  $n \times 1$  formado por uns e  $\mu$  representa o valor médio da suposta variável aleatória  $Y$  avaliada no ponto  $\mathbf{d}^{(i)}$ .

Definidos então uma média e uma matriz de covariância, temos o conjunto essencial de parâmetros para definir uma distribuição multivariada de probabilidade para o vetor de variáveis aleatórias  $\mathbf{Y}$ . Como supomos um processo Gaussiano, então esta distribuição multivariada é suposta como normal e denotada por  $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Psi})$

Para efetivamente podermos aplicar um modelo substituto em Kriging para uma função, precisamos de uma forma efetiva de se determinar todos os parâmetros desconhecidos que fazem parte do modelo, a saber,  $\mu, \sigma^2, \boldsymbol{\theta} = \{\theta_1, \theta_2, \dots, \theta_k\}^T$  e  $\mathbf{p} = \{p_1, p_2, \dots, p_k\}^T$ . Definimos que nosso metamodelo irá interpolar exatamente os dados amostrais, logo não consideraremos erros em  $\mathbf{y} = \{y^{(1)}, y^{(2)}, \dots, y^{(n)}\}^T$  e nem que surjam erros na criação da superfície de resposta. Portanto uma das alternativas para determinação dos parâmetros é escolhê-los de maneira que maximizemos a probabilidade de se obter uma resposta a partir da distribuição de  $\mathbf{Y}$ . Assim, utilizamos a abordagem de se maximizar a verossimilhança das respostas  $\mathbf{y}$  (FORRESTER; SOBESTER; KEANE, 2008).

### 3.3.3 Definição dos parâmetros do Kriging: Verossimilhança

Seja  $Y^{(i)} = Y(\mathbf{d}^{(i)})$  uma das variáveis aleatórias do vetor  $\mathbf{Y}$ . Suponha que a média e a covariância do vetor  $\mathbf{Y}$  são dados por  $\boldsymbol{\mu}$  e  $\sigma^2 \boldsymbol{\Psi}$ , respectivamente. Como demonstrado no Apêndice A, a função de verossimilhança será dada por

$$\begin{aligned} L(\mu, \sigma^2, \boldsymbol{\theta}, \mathbf{p}) &= (2\pi)^{-\frac{n}{2}} |\sigma^2 \boldsymbol{\Psi}|^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} (\mathbf{y} - \mathbf{1}\mu)^T [\sigma^2 \boldsymbol{\Psi}]^{-1} (\mathbf{y} - \mathbf{1}\mu) \right] \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} |\boldsymbol{\Psi}|^{-\frac{1}{2}} \exp \left[ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{1}\mu)^T [\boldsymbol{\Psi}]^{-1} (\mathbf{y} - \mathbf{1}\mu) \right]. \end{aligned} \tag{3.13}$$

Temos que os parâmetros  $\hat{\mu}$ ,  $\hat{\sigma}^2$ ,  $\hat{\boldsymbol{\theta}}$  e  $\hat{\mathbf{p}}$  que são capazes de maximizar a função (3.13), também são minimizadores para a função  $\ln(L(\mu, \sigma^2, \boldsymbol{\theta}, \mathbf{p}))$ . Portanto, podemos escrever a função logaritmo de verossimilhança como

$$L_{\ln}(\mu, \sigma^2, \boldsymbol{\theta}, \mathbf{p}) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \ln |\boldsymbol{\Psi}| - \frac{(\mathbf{y} - \mathbf{1}\mu)^T \boldsymbol{\Psi}^{-1} (\mathbf{y} - \mathbf{1}\mu)}{2\sigma^2}. \quad (3.14)$$

Os parâmetros  $\boldsymbol{\theta}$  e  $\mathbf{p}$  somente influenciam a matriz de correlação  $\boldsymbol{\Psi}$ . Portanto, para simplificar a obtenção dos parâmetros, podemos primeiro otimizar (3.14) somente para os parâmetros  $\mu$  e  $\sigma^2$ . Logo os estimadores dos parâmetros que maximizam a verossimilhança devem resolver simultaneamente as seguintes derivadas parciais

$$\frac{\partial L_{\ln}(\mu, \sigma^2, \boldsymbol{\theta}, \mathbf{p})}{\partial \mu} = 0 \quad (3.15)$$

$$\frac{\partial L_{\ln}(\mu, \sigma^2, \boldsymbol{\theta}, \mathbf{p})}{\partial \sigma^2} = 0. \quad (3.16)$$

De (3.15) temos que

$$\begin{aligned} -\frac{1}{2\sigma^2} \left[ -\mathbf{1}^T \boldsymbol{\Psi}^{-1} (\mathbf{y} - \mathbf{1}\mu) - (\mathbf{y} - \mathbf{1}\mu)^T \boldsymbol{\Psi}^{-1} \mathbf{1} \right] &= 0 \\ \frac{1}{\sigma^2} \mathbf{1}^T \boldsymbol{\Psi}^{-1} (\mathbf{y} - \mathbf{1}\mu) &= 0 \\ \mathbf{1}^T \boldsymbol{\Psi}^{-1} \mathbf{1}\mu &= \mathbf{1}^T \boldsymbol{\Psi}^{-1} \mathbf{y} \\ \mu &= \frac{\mathbf{1}^T \boldsymbol{\Psi}^{-1} \mathbf{y}}{\mathbf{1}^T \boldsymbol{\Psi}^{-1} \mathbf{1}}. \end{aligned}$$

De (3.16) temos que

$$-\frac{n}{\sigma^2} + \frac{(\mathbf{y} - \mathbf{1}\mu)^T \boldsymbol{\Psi}^{-1} (\mathbf{y} - \mathbf{1}\mu)}{(\sigma^2)^2} = 0,$$

e por fim encontramos

$$\sigma^2 = \frac{(\mathbf{y} - \mathbf{1}\mu)^T \boldsymbol{\Psi}^{-1} (\mathbf{y} - \mathbf{1}\mu)}{n}.$$

Portanto como as duas derivadas devem ser satisfeitas ao mesmo tempo, podemos tomar primeiro o estimador para a média como sendo

$$\hat{\mu} = \frac{\mathbf{1}^T \Psi^{-1} \mathbf{y}}{\mathbf{1}^T \Psi^{-1} \mathbf{1}} \quad (3.17)$$

e assim obter o estimador para a variância como

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{1}\hat{\mu})^T \Psi^{-1} (\mathbf{y} - \mathbf{1}\hat{\mu})}{n} \quad (3.18)$$

Substituindo os estimadores (3.17) e (3.18) em (3.14) obtemos

$$L_{\ln}(\boldsymbol{\theta}, \mathbf{p}) = -\frac{n}{2} \ln \hat{\sigma}^2 - \frac{1}{2} \ln |\Psi| - \frac{n}{2} [\ln(2\pi) + 1],$$

do qual podemos remover o termo constante e definir a função logaritmo concentrada da verossimilhança como

$$\mathfrak{L}_{\ln}(\boldsymbol{\theta}, \mathbf{p}) \approx -\frac{n}{2} \ln(\hat{\sigma}^2) - \frac{1}{2} \ln |\Psi|. \quad (3.19)$$

Podemos ver que (3.19) é uma função dependente apenas de  $\Psi$  e por consequência somente de  $\boldsymbol{\theta}$  e  $\mathbf{p}$ , porém de forma implícita. Como desejamos obter o maior valor que (3.19) pode assumir, então devemos maximizá-la novamente. Entretanto, a função (3.19) é não convexa e normalmente multimodal com longos platôs constantes. Isso faz com que a otimização seja uma tarefa mais árdua de se realizar com os algoritmos baseados em gradiente. Portanto, uma abordagem mais confiável é utilizar algum tipo de metaheurístico, como aqueles apresentados na Seção 2.4.3 para poder realizar tal otimização (FORRESTER; SOBESTER; KEANE, 2008).

A etapa mais intensa de cálculo do modelo se encontra na determinação dos parâmetros da correlação. Isto é devido ao fato de, por (3.18) e (3.19), termos que encontrar a inversa e o determinante da matriz de correlação  $\Psi$  (CARRARO, 2017). Porém, a matriz de correlação do Kriging será uma matriz positiva definida e essa característica permite que calculemos sua decomposição Cholesky, aumentando assim, a eficiência do cálculo da inversa com o artifício da substituição

progressiva e regressiva, e no cálculo do determinante (FORRESTER; SOBESTER; KEANE, 2008).

Outra abordagem comumente utilizada na literatura é a de se fixar o valor do parâmetro  $p_r = 2$ , prezando pela suavidade entre todas as dimensões (SACKS et al., 1989; SCHONLAU; WELCH; JONES, 1998; JONES; SCHONLAU; WELCH, 1998; JONES, 2001; SASENA, 2002; BEERS; KLEIJNEN, 2003; FORRESTER; KEANE; BRESSLOFF, 2006; FORRESTER; SOBESTER; KEANE, 2008; FORRESTER; KEANE, 2009; ANKENMAN; NELSON; STAUM, 2010; PICHENY; WAGNER; GINSBOURGER, 2013; CHAUDHURI; HAFTKA, 2014; JALALI; NIEUWENHUYSE; PICHENY, 2017). Dessa forma, somos capazes de reduzir a dimensão do subproblema de otimizar a função logaritmo concentrada da verossimilhança e conseqüentemente resguardamos tempo computacional para outras etapas intensas do processo de otimização.

### 3.3.4 Predição com o metamodelo

Após a maximização de (3.19), esperamos que uma nova predição  $\hat{Y}$  em um novo ponto  $\mathbf{d}^+$  seja consistente com os dados amostrados e conseqüentemente com os parâmetros otimizadores. Para se alcançar essa proposição utilizamos a abordagem baseada nos trabalhos de Jones (2001) e Forrester, Sobester e Keane (2008) para se encontrar um preditor com o Kriging, o qual consiste na maximização da verossimilhança (3.14) de forma análoga a utilizada na seção anterior para determinação dos parâmetros do modelo aproximado.

Primeiro suponhamos que  $\hat{Y}$  seja uma nova previsão por (3.8) dada a função de correlação (3.10) com os parâmetros obtidos na maximização de (3.19). Seja o vetor aumentado

$$\tilde{\mathbf{y}} = \left\{ \mathbf{y}, \hat{Y} \right\}^T, \quad (3.20)$$

e o vetor de correlações entre os dados observados e nossa nova previsão

$$\mathbf{h} = \begin{pmatrix} h(\mathbf{d}^1, \mathbf{d}^+) \\ h(\mathbf{d}^2, \mathbf{d}^+) \\ \vdots \\ h(\mathbf{d}^n, \mathbf{d}^+) \end{pmatrix}. \quad (3.21)$$

Seja a matriz de correlação aumentada  $\tilde{\Psi}$  de ordem  $(n+1) \times (n+1)$  definida por

$$\tilde{\Psi} = \begin{pmatrix} \Psi & \mathbf{h} \\ \mathbf{h}^T & 1 \end{pmatrix}, \quad (3.22)$$

onde  $[\tilde{\Psi}]_{(n+1)(n+1)} = 1$  representa a correlação entre  $\mathbf{d}^+$  e ele mesmo.

Substituindo (3.20) e (3.22) em (3.14) temos

$$L_{\ln}(\hat{Y}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\hat{\sigma}^2) - \frac{1}{2} \ln |\tilde{\Psi}| - \frac{(\tilde{\mathbf{y}} - \mathbf{1}\hat{\mu})^T \tilde{\Psi}^{-1} (\tilde{\mathbf{y}} - \mathbf{1}\hat{\mu})}{2\hat{\sigma}^2}, \quad (3.23)$$

onde somente o último termo dessa nova função é dependente de  $\hat{Y}$ , logo, podemos considerar apenas esse termo para a maximização e realizar uma nova aproximação considerando a função logaritmo da verossimilhança como

$$L_{\ln}(\hat{Y}) \approx -\frac{\begin{Bmatrix} \mathbf{y} - \mathbf{1}\hat{\mu} \\ \hat{Y} - \hat{\mu} \end{Bmatrix}^T \begin{bmatrix} \Psi & \mathbf{h} \\ \mathbf{h}^T & 1 \end{bmatrix}^{-1} \begin{Bmatrix} \mathbf{y} - \mathbf{1}\hat{\mu} \\ \hat{Y} - \hat{\mu} \end{Bmatrix}}{2\hat{\sigma}^2}. \quad (3.24)$$

A ideia para se ter a melhor predição é de que a função (3.24) seja maximizada para o valor de  $\hat{Y}$ , porém se faz necessário encontrar a inversa de  $\tilde{\Psi}$  para a otimização. Supondo que  $\tilde{\Psi}$  é não singular, então ela possui uma única inversa  $\tilde{\Psi}^{-1}$  que satisfaz  $\tilde{\Psi}^{-1}\tilde{\Psi} = \mathbf{I}$ . Como  $\tilde{\Psi}$  é dada por (3.22), então fazemos a suposição de que  $\Psi$  é não singular e

que desejamos determinar uma inversa definida por

$$\tilde{\Psi}^{-1} = \begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}^T & C \end{pmatrix},$$

onde  $\mathbf{A}$  é uma matriz não singular,  $\mathbf{b}$  é um vetor qualquer com  $\mathbf{b}^T$  sendo seu transposto e por fim  $C$  é um escalar real. Logo temos que

$$\tilde{\Psi}^{-1}\tilde{\Psi} = \mathbf{I} \rightarrow \begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}^T & C \end{pmatrix} \begin{pmatrix} \Psi & \mathbf{h} \\ \mathbf{h}^T & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}.$$

Essa equação resulta em quatro equações

$$\begin{aligned} \mathbf{A}\Psi + \mathbf{b}\mathbf{h}^T &= \mathbf{I} \\ \mathbf{A}\mathbf{h} + \mathbf{b} &= \mathbf{0} \\ \mathbf{b}^T\Psi + C\mathbf{h}^T &= \mathbf{0} \\ \mathbf{b}^T\mathbf{h} + C &= 1. \end{aligned} \tag{3.25}$$

Da equação (3.25)<sub>3</sub> temos que

$$\mathbf{b}^T = -C\mathbf{h}^T\Psi^{-1}$$

que se substituído em (3.25)<sub>4</sub> resulta em

$$C(1 - \mathbf{h}^T\Psi^{-1}\mathbf{h}) = 1$$

logo temos que

$$C = (1 - \mathbf{h}^T\Psi^{-1}\mathbf{h})^{-1} \tag{3.26}$$

e que

$$\mathbf{b}^T = - (1 - \mathbf{h}^T\Psi^{-1}\mathbf{h})^{-1} \mathbf{h}^T\Psi^{-1}. \tag{3.27}$$

Pode-se observar que o termo  $(1 - \mathbf{h}^T\Psi^{-1}\mathbf{h})$  trata-se de um escalar e que  $\Psi^T = \Psi$  pois  $\Psi$  é uma matriz simétrica, então aplicando as

propriedades da transposição  $(\mathbf{A}_1\mathbf{A}_2)^T = \mathbf{A}_2^T\mathbf{A}_1^T$  e  $[\mathbf{A}_1^{-1}]^T = [\mathbf{A}_1^T]^{-1}$ , temos que

$$\begin{aligned}
 \mathbf{b} &= - \left[ \left( \mathbf{1} - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h} \right)^{-1} \mathbf{h}^T \boldsymbol{\Psi}^{-1} \right]^T \\
 &= - \left[ \mathbf{h}^T \boldsymbol{\Psi}^{-1} \right]^T \left[ \left( \mathbf{1} - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h} \right)^{-1} \right]^T \\
 &= - \left[ \boldsymbol{\Psi}^{-1} \right]^T \mathbf{h} \left[ \left( \mathbf{1} - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h} \right)^T \right]^{-1} \\
 &= - \left[ \boldsymbol{\Psi}^T \right]^{-1} \mathbf{h} \left( \mathbf{1} - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h} \right)^{-1} \\
 &= - \boldsymbol{\Psi}^{-1} \mathbf{h} \left( \mathbf{1} - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h} \right)^{-1}.
 \end{aligned} \tag{3.28}$$

Substituindo esse valor de  $\mathbf{b}$  na equação (3.25)<sub>1</sub> temos que

$$\begin{aligned}
 \mathbf{A} \boldsymbol{\Psi} - \boldsymbol{\Psi}^{-1} \mathbf{h} \left( \mathbf{1} - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h} \right)^{-1} \mathbf{h}^T &= \mathbf{I} \\
 \mathbf{A} \boldsymbol{\Psi} &= \mathbf{I} + \boldsymbol{\Psi}^{-1} \mathbf{h} \left( \mathbf{1} - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h} \right)^{-1} \mathbf{h}^T \\
 \mathbf{A} &= \left[ \mathbf{I} + \boldsymbol{\Psi}^{-1} \mathbf{h} \left( \mathbf{1} - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h} \right)^{-1} \mathbf{h}^T \right] \boldsymbol{\Psi}^{-1} \\
 \mathbf{A} &= \boldsymbol{\Psi}^{-1} + \boldsymbol{\Psi}^{-1} \mathbf{h} \left( \mathbf{1} - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h} \right)^{-1} \mathbf{h}^T \boldsymbol{\Psi}^{-1}.
 \end{aligned} \tag{3.29}$$

De volta a equação (3.24) temos que

$$\begin{aligned}
 L_{\ln}(\hat{Y}) &\approx -\frac{1}{2\hat{\sigma}^2} \left[ \begin{array}{c} \mathbf{y} - \mathbf{1}\hat{\mu} \\ \hat{Y} - \hat{\mu} \end{array} \right]^T \left[ \begin{array}{cc} \boldsymbol{\Psi} & \mathbf{h} \\ \mathbf{h}^T & 1 \end{array} \right]^{-1} \left[ \begin{array}{c} \mathbf{y} - \mathbf{1}\hat{\mu} \\ \hat{Y} - \hat{\mu} \end{array} \right] \\
 &= -\frac{1}{2\hat{\sigma}^2} \left[ \begin{array}{c} \mathbf{y} - \mathbf{1}\hat{\mu} \\ \hat{Y} - \hat{\mu} \end{array} \right]^T \left[ \begin{array}{cc} \mathbf{A} & \mathbf{b} \\ \mathbf{b}^T & C \end{array} \right] \left[ \begin{array}{c} \mathbf{y} - \mathbf{1}\hat{\mu} \\ \hat{Y} - \hat{\mu} \end{array} \right] \\
 &= -\frac{1}{2\hat{\sigma}^2} \left[ \mathbf{A} (\mathbf{y} - \mathbf{1}\hat{\mu})^2 + (\mathbf{b} + \mathbf{b}^T) (\mathbf{y} - \mathbf{1}\hat{\mu}) (\hat{Y} - \hat{\mu}) + C (\hat{Y} - \hat{\mu})^2 \right].
 \end{aligned}$$

Dessa última equação, considerando que a matriz  $\boldsymbol{\Psi}^{-1}$  é simétrica, podemos ver que o segundo termo será dado por

$$\mathbf{b} + \mathbf{b}^T = -\boldsymbol{\Psi}^{-1} \mathbf{h} \left( \mathbf{1} - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h} \right)^{-1} - \left( \mathbf{1} - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h} \right)^{-1} \mathbf{h}^T \boldsymbol{\Psi}^{-1}$$

$$\begin{aligned}
&= -\frac{\boldsymbol{\Psi}^{-1}\mathbf{h}}{\left(1 - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h}\right)} - \frac{\mathbf{h}^T \boldsymbol{\Psi}^{-1}}{\left(1 - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h}\right)} \\
&= -2 \frac{\mathbf{h}^T \boldsymbol{\Psi}^{-1}}{\left(1 - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h}\right)}.
\end{aligned}$$

Eliminando o termo  $\mathbf{A}(\mathbf{y} - \mathbf{1}\hat{\mu})^2$  por não conter  $\hat{Y}$  temos que

$$L_{\ln}(\hat{Y}) = \left[ -\frac{1}{2\hat{\sigma}^2 (1 - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h})} \right] (\hat{Y} - \hat{\mu})^2 + \left[ \frac{\mathbf{h}^T \boldsymbol{\Psi}^{-1} (\mathbf{y} - \mathbf{1}\hat{\mu})}{\hat{\sigma}^2 (1 - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h})} \right] (\hat{Y} - \hat{\mu}). \quad (3.30)$$

Como temos  $L_{\ln}$  em função de  $\hat{Y}$  e queremos o valor máximo para a função de verossimilhança, então derivando (3.30) em função de  $\hat{Y}$  e igualando a zero temos

$$\begin{aligned}
&\frac{dL_{\ln}(\hat{Y})}{d\hat{Y}} = 0 \\
&\left[ -\frac{1}{\hat{\sigma}^2 (1 - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h})} \right] (\hat{Y} - \hat{\mu}) + \left[ \frac{\mathbf{h}^T \boldsymbol{\Psi}^{-1} (\mathbf{y} - \mathbf{1}\hat{\mu})}{\hat{\sigma}^2 (1 - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h})} \right] = 0 \\
&\hat{Y} = \hat{\mu} + \mathbf{h}^T \boldsymbol{\Psi}^{-1} (\mathbf{y} - \mathbf{1}\hat{\mu}).
\end{aligned}$$

Portanto, pelo máximo da verossimilhança temos que a predição via metamodelo com o Kriging é dado por

$$\hat{Y}(\mathbf{d}^+) = \hat{\mu} + \mathbf{h}^T \boldsymbol{\Psi}^{-1} (\mathbf{y} - \mathbf{1}\hat{\mu}), \quad (3.31)$$

onde  $\mathbf{h}$  é dado por (3.21) e  $\boldsymbol{\Psi}$  é definida em (3.11).

Podemos ver facilmente que a predição dada por (3.31) interpola todos os pontos de suporte. Tomando  $\mathbf{d}^+ = \mathbf{d}^{(i)}$  temos que (3.21) se torna a  $i$ -ésima coluna da matriz de correlação  $\boldsymbol{\Psi}$  e com isso temos que  $\boldsymbol{\Psi}^{-1}\mathbf{h} = \mathbf{e}_i$ , onde  $\mathbf{e}_i$  é o  $i$ -ésimo vetor canônico, portanto

$$\begin{aligned}
\hat{Y}(\mathbf{d}^{(i)}) &= \hat{\mu} + (\mathbf{y} - \mathbf{1}\hat{\mu})^T \mathbf{e}_i \\
&= \hat{\mu} + y^{(i)} - \hat{\mu} \\
&= y^{(i)}.
\end{aligned} \quad (3.32)$$



Cabe ressaltar que fizemos todo o processo para predição de um valor via o Kriging através da suposição de uma variável aleatória  $Y(\mathbf{d})$ , porém em (3.31) nós não temos uma variável aleatória. O que determina-se com (3.31) é exatamente o valor da predição no ponto. Porém, associado aos pontos que não são de suporte, teremos um erro cometido nessas predições. O processo de determinação desse erro torna-se possível do fato de o Kriging ser um processo Gaussiano.

### 3.3.5 Erro na predição

Um dos grandes potenciais de se utilizar o Kriging como um metamodelo é que o seu processo permite o cálculo de uma estimativa para o erro das respostas do modelo. Essa estimativa do potencial erro no preditor está associada de forma inversamente proporcional à curvatura da função logaritmo concentrada de verossimilhança. Uma curvatura baixa (curva mais plana) sugere um alto potencial de erro, já uma curvatura alta (curvas com picos) sugere um baixo potencial de erro (JONES, 2001). Sacks et al. (1989), usando um processo padrão estocástico, propuseram que o Erro Quadrado Médio (*Mean Squared Error - MSE*) fosse dado por

$$s^2(\mathbf{d}) = \hat{\sigma}^2 \left[ 1 - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h} + \frac{(1 - \mathbf{1}^T \boldsymbol{\Psi}^{-1} \mathbf{h})^2}{\mathbf{1}^T \boldsymbol{\Psi}^{-1} \mathbf{1}} \right]. \quad (3.33)$$

O terceiro termo dentro dos colchetes representa a incerteza decorrente do fato de não conhecermos exatamente  $\mu$  e termos estimado seu valor a partir da amostra inicial de  $n$  pontos. Para uma dedução completa de (3.33) sugere-se os trabalhos de Sacks et al. (1989), Sasena (2002), Hoyle (2006).

Se tomarmos um ponto de suporte qualquer e colocarmos  $\mathbf{d} = \mathbf{d}^{(i)}$  podemos fazer uma análise análoga à da predição com

$$\mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h} = \mathbf{h}^T \mathbf{e}_i = [\mathbf{h}]_i = 1$$

e

$$\mathbf{1}^T \boldsymbol{\Psi}^{-1} \mathbf{h} = \mathbf{1}^T \mathbf{e}_i = [\mathbf{1}]_i = 1,$$

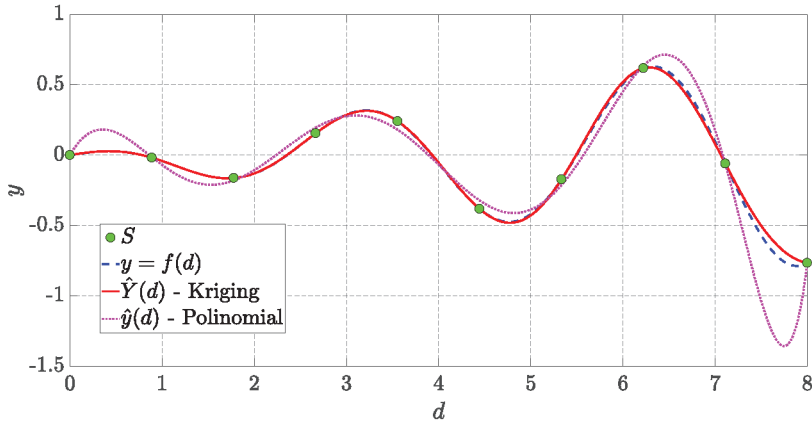


Figura 10 Modelos substitutos da função (3.6).

logo vemos, de (3.33), que  $s^2(\mathbf{d}^{(i)}) = 0$ , e assim verificamos que o nosso modelo interpola exatamente os dados amostrados. Essa lógica somente se aplica quando não temos nenhuma dúvida acerca da fonte dos valores amostrados  $y^{(i)} = y(\mathbf{d}^{(i)})$  (FORRESTER; SOBESTER; KEANE, 2008).

### 3.3.6 Alguns exemplos

A primeira aplicação analisada é a aproximação da função (3.6) com  $a = 2$ . Foram utilizados  $n = 10$  pontos de suporte escolhidos de forma que o intervalo  $[0, 8]$  foi particionado em 9 subintervalos iguais. A Figura 10 apresenta os modelos substitutos do Kriging e de uma Regressão Polinomial de grau 8 (a mesma utilizada para a Figura 7). Podemos ver como o metamodelo obtido pelo Kriging é superior ao polinomial, tendo o Kriging conseguido captar o formato da função em quase todo o domínio.

Um segundo exemplo é dado pela aproximação da função  $f : [-0.5, 4.5] \rightarrow \mathbb{R}$  dada por

$$f(d) = (2d - 4) \exp[-(d^2 - 4d + 3)] \text{sen}(0.7d^2 + 4.9d). \quad (3.34)$$

A Figura 11 apresenta a função, o modelo substituto em Kriging e também as curvas referentes à variação do valor da predição quando

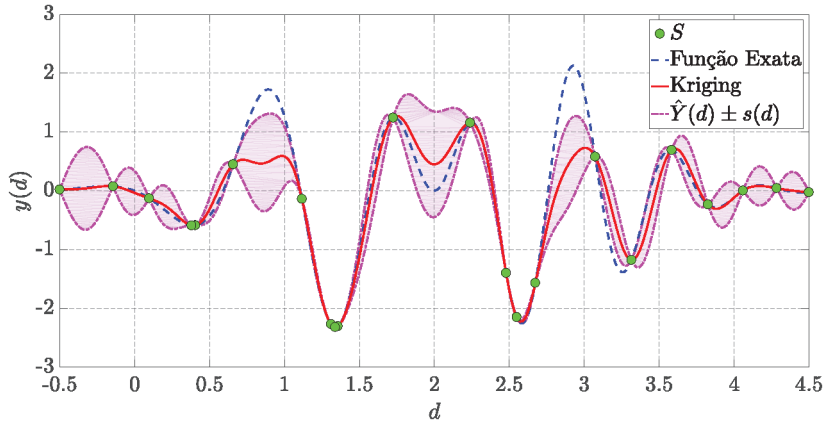


Figura 11 Interpolação com o Kriging e seu erro.

considerados os erros obtidos por (3.33). O preenchimento entre as curvas  $\hat{Y}(d) \pm s(d)$  mostram a amplitude de variação do erro na predição. Podemos notar que em regiões onde o Kriging não conseguiu captar a forma da função o erro está maior, indicando que pontos de suporte nessas regiões podem ser tomados para que o metamodelo seja mais exato. Vale ressaltar que essa característica será explorada para a otimização global na Seção 3.4.

### 3.4 ALGORITMO EGO

Nas seções anteriores vimos como podemos ajustar um modelo substituto em Kriging para um determinado conjunto amostral de pontos. Nesta seção, desenvolvemos a primeira abordagem de otimização de uma função multimodal, não convexa e cara computacionalmente, e com foco no problema padrão sem restrições definido por (2.6) e modificado para

$$\min_{\mathbf{d} \in \mathcal{D}} f(\mathbf{d}),$$

onde  $\mathcal{D} = \{\mathbf{d} \in \mathbb{R}^k \mid \mathbf{lb} \leq \mathbf{d} \leq \mathbf{ub}\}$ .

O interesse na aplicação dos metamodelos é contornar o caro processo de avaliação da função  $f$ . Para tanto, vamos supor que o modelo substituto  $\hat{Y}$  seja uma representação fiel de  $f$ , logo podemos buscar o minimizador através da avaliação exaustiva (porém barata) do modelo substituto até que encontremos o ótimo global. Obtendo esse minimizador, sabemos que ele minimiza o modelo  $\hat{Y}$  porém o processo de otimização ainda não estará encerrado até que façamos a avaliação da função de alta fidelidade nesse possível minimizador (FORRESTER; SOBESTER; KEANE, 2008).

Porém, para que esse processo imaginário funcione e obtenhamos um excelente candidato a minimizador, devemos garantir a melhor substituição possível  $\hat{Y}$ , pelo menos na região onde se encontra o mínimo global. Para tanto, devemos explorar a informação barata do metamodelo de modo a identificar quais são as bacias de mínimos promissoras, e conseqüentemente, reduzir o erro do metamodelo nessas regiões.

Uma forma de garantir regiões de mínimo com baixo erro na modelagem é sucessivamente ir adicionando novos pontos de suporte ao conjunto  $S$ , que pertençam à vizinhança dessas bacias. Dessa forma, melhoramos a correlação entre os pontos dessas vizinhanças e temos uma maior confiabilidade do substituto nessa região. Esses novos pontos de suporte são chamados de Pontos de Preenchimento (*Infill Points* - IPs) e são determinados utilizando somente as informações do metamodelo, em específico o preditor e o MSE dados pelas equações (3.31) e (3.33), respectivamente.

O EGO é o nome dado por Jones, Schonlau e Welch (1998) ao seu conjunto de soluções para a otimização global. Em seu trabalho os autores exploram as informações sobre o metamodelo em Kriging para iterativamente ir adicionando IPs ao mesmo tempo que buscam pelo ótimo global. Em Jones (2001) temos várias formas diferentes de se determinar os IPs, aqui iremos focar no estudo da Melhora Esperada (*Expected Improvement* - EI) e que será utilizada em todos os processos de otimização desenvolvidos nesse trabalho.

De acordo com Jones (2001) basicamente o processo de otimização via metamodelos pode ser dividido nas seguintes etapas:

1. Definir um espaço amostral inicial de pontos de suporte e realizar o ajuste de um metamodelo a estes pontos;
2. Calcular o próximo IP utilizando uma métrica de exploração;
3. Adicionar este novo ponto de suporte ao espaço amostral, reajustar um modelo aos novos dados e ir para o passo 2. Fazer isso até se atingir um critério de parada.

No passo 1 o espaço amostral normalmente é criado utilizando algum processo que garanta o maior preenchimento do espaço de busca possível. Isso irá garantir a maior eficiência para podermos obter informações globais da função (PRONZATO; MÜLLER, 2012).

Um dos recursos mais utilizados na literatura para tal feito é o uso dos Hipercubos Latinos (Latin Hypercube - LH). Essa técnica objetiva realizar amostras, garantindo uma distribuição de pontos em todos os eixos, de forma que, nenhuma dimensão seja avaliada mais de uma vez no mesmo valor (CARRARO, 2017). Mais detalhes sobre a criação de LHs poderá ser encontrada nos trabalhos de Johnson, Moore e Ylvisaker (1990), Mitchell e Morris (1992). Após a construção do espaço amostral passamos para o ajuste do Kriging conforme especificado nas seções anteriores.

O passo 2 constitui-se de um subproblema de otimização, no qual a função objetivo é a métrica EI e o espaço de busca é o mesmo do problema de otimização,  $\mathcal{D}$ . O ponto maximizador do EI é considerado um IP e então adicionado ao espaço amostral. Por fim, no passo três ajustamos um novo modelo aos dados atualizados pelo IP. Essas duas etapas são realizadas até que se atinja um número máximo de avaliações da função objetivo ou algum outro critério de parada específico.

### 3.4.1 Definição da métrica EI

Uma forma natural de se conhecer novos IPs seria o de otimizar um metamodelo já ajustado e utilizar esse minimizador como novo IP. Porém, Jones, Schonlau e Welch (1998) mostraram que esse processo, mesmo se repetido várias vezes poderia levar a uma otimização local e não global. Isso ocorre pois não levamos em consideração a incerteza sobre todo o metamodelo e começamos a obter um metamodelo extremamente aproximado na região do mínimo já obtido, representando a exploração local do domínio de busca.

Para podermos ter um cenário mais global da nossa aproximação via metamodelos, precisamos colocar certa ênfase na busca das regiões do domínio com algum grau de incerteza sobre o modelo substituído. Uma forma de concretizar essa ação é a pesquisa da raiz quadrada do erro quadrado médio (*Root Mean Squared Error - RMSE*), calculada extraíndo a raiz quadrada do valor dado por (3.33).

Na Figura 12 temos representado a função (3.34) e a aproximação com o Kriging em uma amostra de 12 pontos, definidos pelo autor. A curva laranja abaixo das demais representa o RMSE obtido por (3.33) dessa aproximação via Kriging. Se procurássemos pelo minimizador de  $\hat{Y}$  teríamos sucesso em encontrar um minimizador local

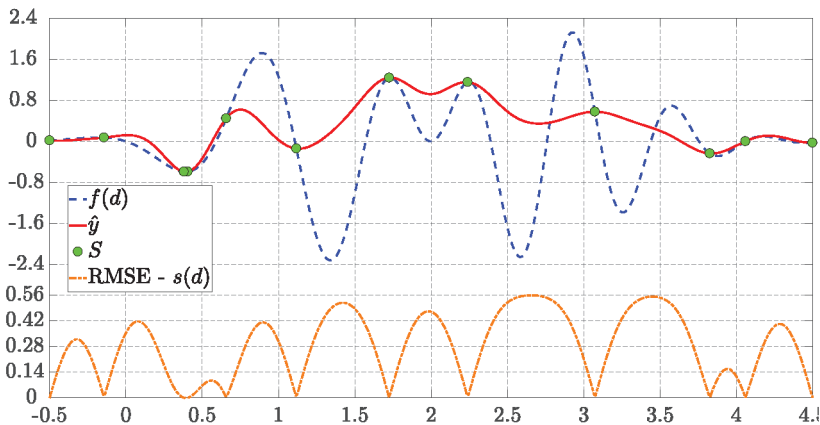


Figura 12 Metamodelo em Kriging e seu RMSE.

de  $f$  em  $d^* = 0.3918$  graças a nulidade do RMSE nesse ponto. Como consequência dessa primeira aproximação, não seríamos capazes de encontrar a região do mínimo global via minimização do metamodelo.

Pode-se ver que o RMSE assume seu valor máximo, aproximadamente 0.56, no ponto  $d = 2.6563$ , e este poderia ser um excelente candidato a IP do ponto de vista da pesquisa global, já que conseguiríamos sair da bacia do mínimo encontrado até o momento. Porém, fazer a amostragem desse IP seria equivalente a depositar toda nossa fé somente na pesquisa global e isso pode ser até pior do que uma busca local (JONES; SCHONLAU; WELCH, 1998). Para piorar nossa situação nesse cenário, iríamos ajustar um novo modelo considerando um IP em outra bacia de mínimo local. Isso mostra que a métrica a ser escolhida para a adição dos IPs deve conter um bom balanço entre pesquisa local e pesquisa global, além de envolver o valor mínimo da função já obtido.

Uma métrica mais robusta do que o RMSE é o EI, e seu conceito pode ser encontrado na literatura desde 1978 nos trabalhos de Mockus, Tiesis e Zilinskas (1978), Locatelli (1997), Schonlau (1997), Schonlau, Welch e Jones (1998), Forrester, Sobester e Keane (2007), Forrester, Sobester e Keane (2008), Nascentes et al. (2018). Como o nome indica, o EI basicamente calcula o quanto de melhora esperamos alcançar se fizermos uma amostragem em determinado ponto.

Suponha que  $f_{\min} = \min\{y^{(1)}, y^{(2)}, \dots, y^{(n)}\}$  seja o menor valor obtido atualmente para a função. Ao calcularmos uma aproximação  $\hat{Y}(\mathbf{d})$  para  $f(\mathbf{d})$  temos que alguma incerteza sobre as predições em pontos que não são de suporte irá ocorrer. Vamos tratar as predições pelo Kriging como se fossem a realização de uma variável aleatória normalmente distribuída dada por  $Y(\mathbf{d}) \sim \mathcal{N}(\hat{Y}(\mathbf{d}), s(\mathbf{d}))$ , onde  $\hat{Y}(\mathbf{d})$  e  $s(\mathbf{d})$  são dadas, respectivamente, por (3.31) e pela raiz quadrada de (3.33). A ideia dessa suposição é apresentada na Figura 13, onde para o ponto  $d^a = 1.3104$  desenhamos a curva da distribuição normal  $Y(d^a) \sim \mathcal{N}(0.0968, 0.4519)$ . Podemos observar que existe alguma probabilidade de que o valor predito  $Y(d^a)$  seja menor que o nosso melhor

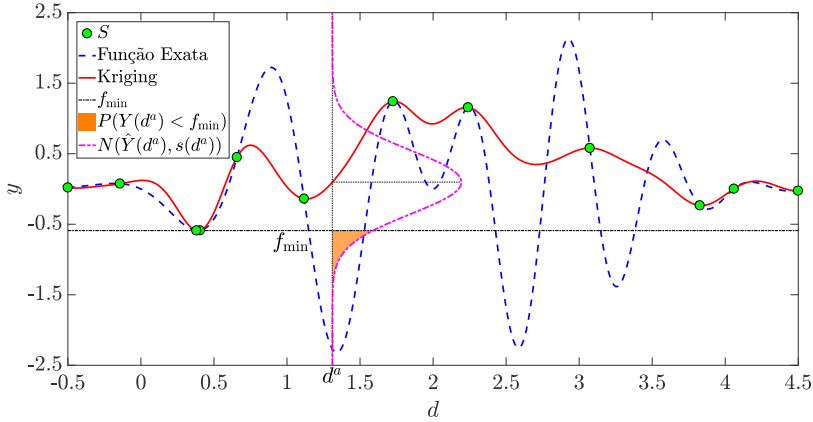


Figura 13 – Incerteza sobre o valor predito em  $d^a$ .

valor até o momento  $f_{\min} = -0.5899$ , graças a parte da cauda da distribuição localizada abaixo de  $f_{\min}$ . A probabilidade  $P(Y(d^a) < f_{\min})$  está representada pela área hachurada em laranja na figura.

Diferentes quantidades de melhora, ou diferentes distâncias abaixo de  $f_{\min}$ , estão associadas a variabilidade do RMSE ao longo do domínio no metamodelo, e conseqüentemente à densidade de probabilidade  $P(Y(d^a) < f_{\min})$ . Se ponderarmos todas as possíveis melhoras pela sua densidade de probabilidade associada, obteremos a métrica que chamamos de EI (JONES; SCHONLAU; WELCH, 1998).

Para podermos formalizar matematicamente essa teoria, iremos fazer uma simplificação omitindo a dependência de  $\mathbf{d}$  da notações de  $\hat{Y}(\mathbf{d})$ ,  $s(\mathbf{d})$  e  $Y(\mathbf{d})$ . Seja a melhora esperada em um ponto  $\mathbf{d}$  dada por

$$I = \max\{f_{\min} - Y, 0\}. \quad (3.35)$$

Essa definição de  $I$  é uma variável aleatória pela dependência de  $Y$ . Para obtermos o valor do EI precisamos tomar a esperança matemática (média) dessa melhora. Tomando o caso onde  $I \geq 0$  e conseqüentemente  $Y = f_{\min} - I$ , então a função de densidade de probabilidade para



obtermos algum valor melhor é dada por

$$F(I) = \frac{1}{s\sqrt{2\pi}} \exp \left[ -\frac{(f_{\min} - I - \hat{Y})^2}{2s^2} \right]. \quad (3.36)$$

Dessa forma, o EI é dado por

$$\mathbb{E}(I) = \int_0^\infty I \left[ \frac{1}{s\sqrt{2\pi}} \exp \left[ -\frac{(f_{\min} - I - \hat{Y})^2}{2s^2} \right] \right] dI. \quad (3.37)$$

Seja a seguinte mudança de variáveis

$$T = \frac{Y - \hat{Y}}{s} \quad \text{e} \quad u = \frac{f_{\min} - \hat{Y}}{s}, \quad (3.38)$$

da qual temos que  $s dT = dY$ . Como  $Y = f_{\min} - I$  então

$$I = f_{\min} - Y = su + \hat{Y} - Y = su - sT = s(u - T) \quad (3.39)$$

e

$$dI = -dY.$$

Para garantirmos que o valor de  $I$  seja não negativo fazemos

$$I = \begin{cases} s(u - T) & \text{se } T < u \\ 0 & \text{se } T \geq u. \end{cases} \quad (3.40)$$

Substituindo (3.38) e (3.39) em (3.37) e levando em consideração (3.40) temos que

$$\begin{aligned} \mathbb{E}(I) &= \frac{1}{s\sqrt{2\pi}} \int_{-\infty}^u s(u - T) \exp \left( -\frac{T^2}{2} \right) s dT \\ &= \frac{s}{\sqrt{2\pi}} \left[ \int_{-\infty}^u u \exp \left( -\frac{T^2}{2} \right) dT - \int_{-\infty}^u T \exp \left( -\frac{T^2}{2} \right) dT \right] \\ &= \frac{s}{\sqrt{2\pi}} \left[ u \int_{-\infty}^u \exp \left( -\frac{T^2}{2} \right) dT + \left( \exp \left( -\frac{T^2}{2} \right) \Big|_{-\infty}^u \right) \right] \\ &= s \left[ u \int_{-\infty}^u \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{T^2}{2} \right) dT + \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{u^2}{2} \right) \right]. \end{aligned}$$

A integral dentro dos colchetes é a representação da distribuição cumulativa de probabilidade da variável aleatória normal padrão  $T \sim \mathcal{N}(0,1)$  de  $-\infty$  até  $u$ , que será representada por  $\Phi(u)$ . O segundo termo da soma dentro dos colchetes é igual a função densidade de probabilidade de uma variável aleatória normal padrão avaliada em  $u$ , que será denotada por  $\phi(u)$ , logo

$$\mathbb{E}(I) = s [u\Phi(u) + \phi(u)], \quad (3.41)$$

e da segunda igualdade em (3.38) temos que a métrica EI em um ponto qualquer  $\mathbf{d}$  é calculada por

$$\mathbb{E}(I(\mathbf{d})) = \left( f_{\min} - \hat{Y}(\mathbf{d}) \right) \Phi \left( \frac{f_{\min} - \hat{Y}(\mathbf{d})}{s(\mathbf{d})} \right) + s(\mathbf{d}) \phi \left( \frac{f_{\min} - \hat{Y}(\mathbf{d})}{s(\mathbf{d})} \right). \quad (3.42)$$

O primeiro termo em (3.42) é a diferença entre o mínimo atual e o valor predito em  $\mathbf{d}$ , penalizado pela probabilidade de haver melhora no valor predito. O segundo termo será grande quando  $\hat{Y}(\mathbf{d})$  estiver próximo do valor de  $f_{\min}$  e  $s(\mathbf{d})$  for um valor alto, ou seja, quando possuímos muita incerteza se  $\hat{Y}(\mathbf{d})$  irá conseguir alcançar  $f_{\min}$  (SCHONLAU, 1997). Isso nos mostra o balanço entre exploração local e a exploração global do EI, que tenderá a possuir um valor alto quando o valor predito for menor que  $f_{\min}$  ou quando muita incerteza estiver associada à predição.

A Figura 14 traz uma apresentação de diferentes etapas do EGO via EI. No gráfico 14(a) temos o mesmo metamodelo ajustado para a função (3.34) apresentado na Figura 12 com 12 pontos de suporte. No gráfico 14(b) temos a curva que representa a métrica do EI em todo o domínio de busca da função. Podemos ver que o ponto que traz o máximo do EI é  $\mathbf{d} = 1.3136$  representado pela estrela azul na figura, e este será o próximo IP a ser adicionado. Podemos ver que esse valor é diferente do valor que teríamos obtido se utilizássemos somente o RMSE como condição para a adição do IP. Os gráficos 14(c), 14(e) e 14(g) mostram as atualizações do metamodelo quando reajustamos eles considerando no espaço amostral os IP indicados nos gráficos 14(b),

14(d) e 14(f), respectivamente. Vemos que três IPs foram adicionados na bacia do mínimo global, isso mostra o poder de exploração local do EI. Podemos ver que o gráfico 14(f) apresenta um ponto de exploração global obtido pelo EI.

O gráfico 14(i) não apresenta o modelo após a adição do IP apresentado pelo gráfico 14(h), ele apresenta o modelo final com um total de 30 pontos de suporte na amostra, incluindo o ponto obtido em 14(h). Podemos ver que não há distinção entre o metamodelo e a função original, bem como o EI é totalmente nulo em todo o domínio, como mostra o gráfico 14(j). Quando comparamos as Figuras 14(a) e 14(i) podemos ver efetivamente o balanço entre exploração local e global, onde nas bacias que apresentam mínimos foram colocados entre 1 e 4 IPs, e mais 8 IPs foram colocados em regiões fora dessas bacias.

A análise gráfica anterior serve também para ilustrar uma dificuldade no subproblema de se otimizar (3.42) sobre uma região contínua. Como nos pontos de suporte temos EI igual a zero, conforme nos afastamos destes, o RMSE tende a aumentar e como consequência temos o aumento do EI, isso causa os vários “picos” que podemos ver nos Figuras 14(b), 14(d), 14(f) e 14(h). Portanto, maximizar o EI, já é por si só outro problema de otimização global; logo, os métodos de otimização tradicionais ou os heurísticos podem falhar durante esse subprocesso.

Porém, essa dificuldade na otimização do EI não prejudica o EGO, como pode ser confirmado nos trabalhos de Mockus (1994) e por Schonlau (1997). Mais especificamente, Mockus (1994) afirma que não há necessidade de uma exata otimização da métrica do EI, isso por que essa otimização só determina o próximo ponto de observação para o novo metamodelo (IP).

A garantia da otimização global via EI pode ser conferida no trabalho de Schonlau (1997), onde ele prova o seguinte teorema: Suponha que usando o modelo Gaussiano dado por (3.8) e a função de correlação (3.10) temos que o MSE dado por (3.33) é sempre positivo para qualquer

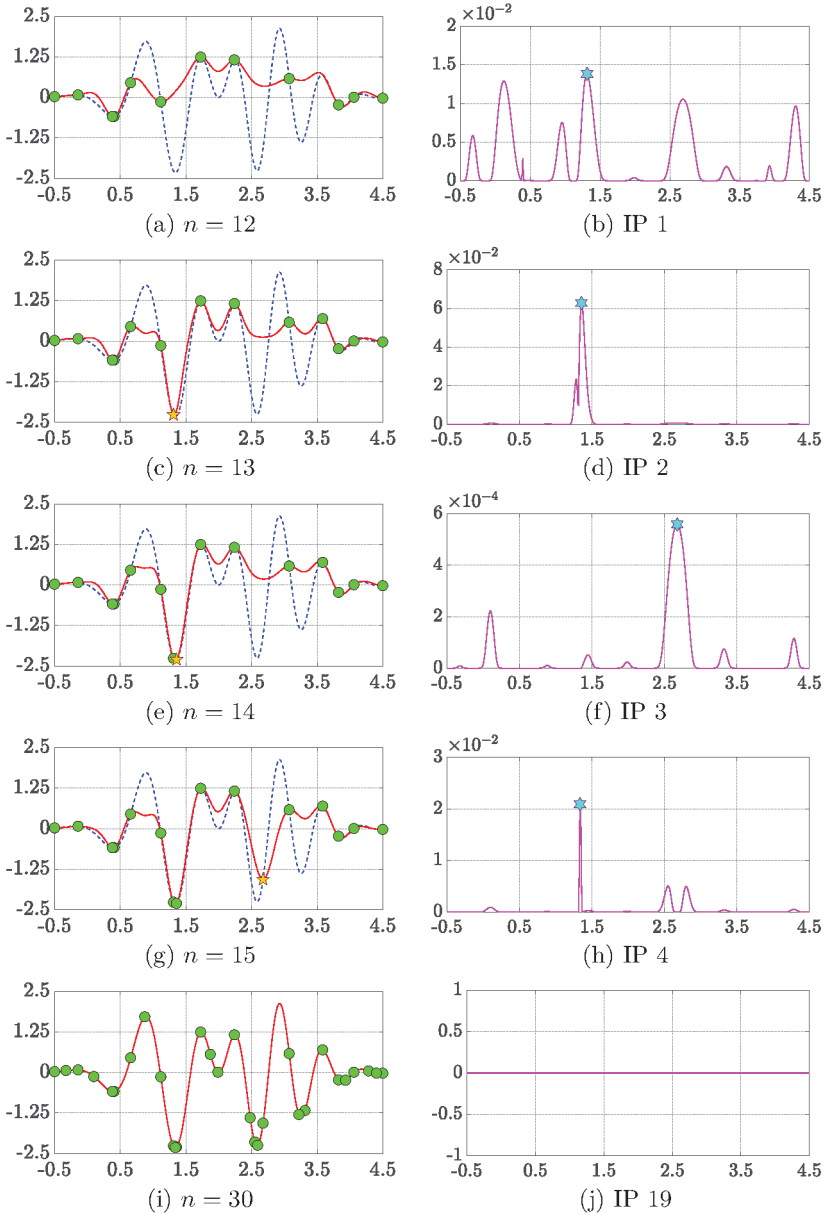


Figura 14 Diferentes avaliações do EI durante o EGO.

ponto que não seja de suporte  $\mathbf{d}$ . Além disso, suponha que o número possível de pontos de suporte é finito. Então o algoritmo EI irá visitar todos estes pontos e encontrar o mínimo global.

A Figura 15 mostra um fluxograma para a execução completa da otimização via EGO.

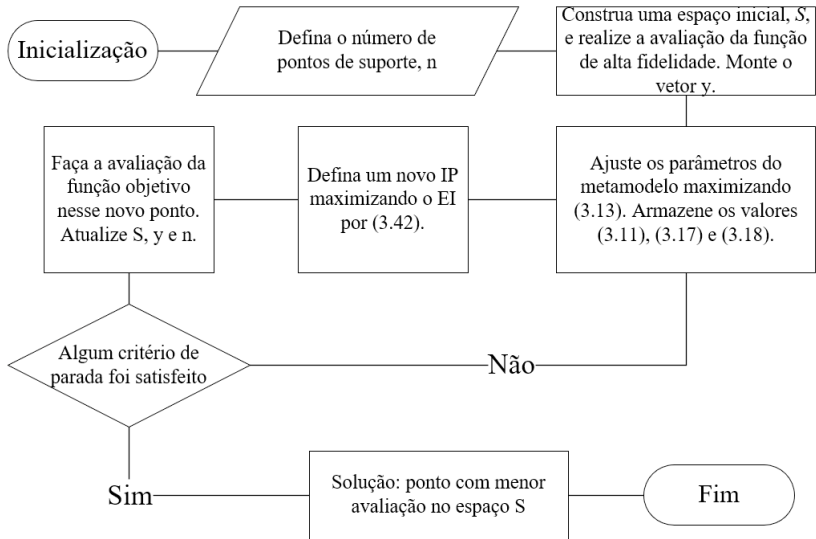


Figura 15 – Fluxograma do EGO.

Podemos ver que o EGO é um exímio otimizador global, que consegue varrer todo o espaço de busca procurando pelo ponto que traz o menor valor possível da função objetivo, e que não pode ser melhorado. Ao fim da pesquisa do domínio, temos que o metamodelo criado substitui perfeitamente a função original, pelo menos na bacia onde se encontra o mínimo global. Outra vantagem que podemos citar para o EGO é a diminuição considerável de avaliações da função objetivo que deveremos fazer. Essa diminuição leva a uma consequente diminuição no tempo de processamento.

Porém, o EGO também possui algumas desvantagens que cabe

destaque:

- Devemos tomar um cuidado com o número de dimensões do problema a ser analisado. Em dimensões elevadas, necessitamos de um número maior de pontos de suporte no espaço amostral inicial, e o EGO pode demorar algumas iterações até encontrar a região de mínimo global. Como a dimensão da matriz de covariância do Kriging é a mesma do número de pontos de suporte, esta matriz pode se tornar intratável computacionalmente. Usualmente é indicado que  $k < 20$  (JONES; SCHONLAU; WELCH, 1998; JONES, 2001; SASENA, 2002; FORRESTER; SOBESTER; KEANE, 2008; PICHENY et al., 2013; JALALI; NIEUWENHUYSE; PICHENY, 2017);
- A parte mais custosa do algoritmo, conseqüentemente mais demorada, é o ajuste dos parâmetros via maximização da verossimilhança, onde necessitamos do cálculo da inversa da matriz de covariância do Kriging. Logo, aqui deve-se utilizar algoritmos que possam, de forma rápida, alcançar uma solução;
- Se comparado com outros algoritmos de otimização, o EGO possui alguma dificuldade de implementação.

No formato apresentado até o momento, não se é possível resolver as funções integrais que analisaremos, ou nem utilizar o EGO no contexto estocástico. Isso ocorre pois durante o ajuste dos parâmetros do Kriging fazemos a suposição de que o valor da função é determinístico. Logo, como mostraremos a seguir, isso faz que em situações estocásticas, o Kriging tome os valores da função e as interpole sem levar em consideração o erro associado àquele valor, fazendo com que o Kriging superajuste os dados.

Temos também que em situações estocásticas o EI não se aplica, pois não temos certezas sobre o valor mínimo que a função objetivo já possui. Dessa forma devemos ser capazes de trazer as incertezas acerca

das avaliações das funções para dentro da métrica de IP. Isso torna todo o processo mais confiável e robusto.





## 4 sEGO ADAPTATIVO COM NORMALIZAÇÃO

Os métodos de substituição e de otimização desenvolvidos na seção anterior foram formulados sob a premissa de que a função a ser modelada era suave, contínua e não continha incertezas sobre o valor experimentado. Isso nos garantia a hipótese de que para os pontos amostrados  $\mathbf{d}^{(i)}$ , toda simulação numérica resultaria sempre em  $\hat{y}^{(i)} = f(\mathbf{d}^{(i)})$ .

Em algumas aplicações de engenharia, muitas funções possuem uma tendência suave e contínua. Entretanto, as avaliações dessas funções são normalmente uma dispersão em torno dessa tendência devido a erros no experimento físico ou na simulação computacional usada para avaliar a função (FORRESTER; SOBESTER; KEANE, 2008).

Nos últimos anos houve um grande aumento no estudo de processos ou simulações estocásticos, envolvendo funções que só podem ser observadas na presença de ruído. Entre eles podemos destacar as aplicações a: eventos de simulações discretas (ANKENMAN; NELSON; STAUM, 2010), medidas experimentais na área de engenharia mecânica (BIERMANN; WEINERT; WAGNER, 2008), avaliações de segurança nuclear (MISS; JACQUET; HEULERS, 2005), propagação de ondas acústicas em fluídos turbulentos (IOOSS; LHUILLIER; JEANNEAU, 2002), otimização robusta de aerofólios (LI; HUYSE; PADULA, 2002), projeto de materiais compostos (SAKATA; ASHIDA; ZAKO, 2008; SAKATA; ASHIDA., 2009) e propagação de trincas em metais (STEPHENS; FUCHS, 2001).

Os tipos de funções que aparecem nos problemas citados anteriormente também podem ser substituídos por metamodelos, porém pela presença do erro numérico ou incerteza de alguns parâmetros, é mais natural pensarmos em um metamodelo como regressor do que como interpolador. Dessa forma, um substituto regressor pode extrair a tendência suave dos dados e filtrar os ruídos não passando por todos os valores da função.

Nesse sentido, vários foram os esforços de se estender o Kriging Determinístico para o que é chamado de Kriging Estocástico (*Stochastic Kriging* - SK). Entre os diversos autores que trabalharam em cima dessa nova abordagem citamos [Beers e Kleijnen \(2003\)](#), [Forrester, Keane e Bressloff \(2006\)](#), [Huang et al. \(2006\)](#), [Staum \(2009\)](#), [Ankenman, Nelson e Staum \(2010\)](#), [Picheny, Wagner e Ginsbourger \(2012\)](#), [Chen e Kim \(2014\)](#), [Plumlee e Tuo \(2014\)](#), [Kleijnen e Mehdad \(2015\)](#).

#### 4.1 APROXIMAÇÃO DA FUNÇÃO DE ESTUDO

Relembrando, a função integral de estudo (1.1) é dada por

$$J(\mathbf{d}) = \int_{\Omega} \psi(\mathbf{d}, \mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x},$$

e o problema de minimização (1.2) é escrito como

$$\min_{\mathbf{d} \in \mathcal{D}} J(\mathbf{d}).$$

Na definição da função acima, vemos que existe uma diferença entre o número da dimensão  $k$  do vetor de variáveis de projeto e a dimensão estocástica  $k_x$  da integral. Nos casos onde  $\psi(\mathbf{d}, \mathbf{x})$  apresentar uma avaliação custosa computacionalmente, a avaliação de um único valor para a função (1.1) pode se tornar intratável. Algumas técnicas de integração numérica tais como os processos de quadratura podem ser aplicados, porém possuem baixa eficiência para problemas de alta dimensão.

Em tais casos, uma maneira mais eficiente de se calcular (1.1) é por meio de técnicas amostrais, tais como a já citada MCI ([LEPAGE, 1978](#); [CAFLISCH, 1998](#)), *Importance Sampling* ([RUBINSTEIN; KROESE, 2017](#)), *Multi Level Monte Carlo* ([GILES, 2008](#)) e *Multi Index Monte Carlo* ([HAJI-ALI; NOBILE; TEMPONE, 2016](#)).

Utilizando a técnica de MCI podemos aproximar (1.1) por

$$J(\mathbf{d}) \approx \bar{J}(\mathbf{d}) = \frac{1}{n_r} \sum_{i=1}^{n_r} \psi(\mathbf{d}, \mathbf{x}^{(i)}), \quad (4.1)$$

onde  $n_r$  é o número de pontos amostrais utilizado pelo MCI e  $\mathbf{x}^{(i)}$  é um ponto amostral escolhido aleatoriamente a partir da distribuição  $f_{\mathbf{x}}(\mathbf{x})$ . Não é difícil extrair a conclusão de que cada replicação do valor de  $J(\mathbf{d})$  pode ser diferente, visto a aleatoriedade de  $\mathbf{x}^{(i)}$ .

A utilização de uma técnica amostral como a MCI possui a vantagem de nos possibilitar a estimação de um valor para a variância do erro cometido durante a aproximação. Podemos utilizar um estimador não viciado para a variância em um ponto fixo  $\mathbf{d}$  como sendo

$$\bar{\sigma}^2(\mathbf{d}) = \frac{1}{n_r(n_r - 1)} \sum_{i=1}^{n_r} \left( \psi(\mathbf{d}, \mathbf{x}^{(i)}) - \bar{J}(\mathbf{d}) \right)^2. \quad (4.2)$$

Dessa forma, conforme o número de pontos amostrais do MCI aumenta, a variância do erro no ponto  $\mathbf{d}$  diminui tornando a aproximação (4.1) mais próxima do valor original (1.1). Os trabalhos de [Hammersley e Handscomb \(1964\)](#) e [Kalos e Whitlock \(1986\)](#) demonstraram que a taxa de convergência do MCI é de  $n_r^{-1/2}$ .

Para mostrarmos como a aplicação de (4.1) pode ser utilizada para aproximar (1.1), vamos utilizar a função (3.34) onde aqui a definimos por

$$\psi(d, X) = \left[ (2d - 4) \exp[-(d^2 - 4d + 3)] \operatorname{sen}(0.7d^2 + 4.9d) \right] X, \quad (4.3)$$

com  $X$  sendo uma variável aleatória normal com média 1 e desvio padrão igual a 0.5, ou seja,  $X \sim \mathcal{N}(1, 0.5)$ . Logo, temos que a esperança matemática da função (4.3) será dada por

$$J(d) = \mathbb{E}[\psi(d, X)] = \int_{-\infty}^{\infty} \psi(d, X) f_X(X) dX, \quad (4.4)$$

onde  $f_X(X)$  é a função densidade de probabilidade normal para a variável. Facilmente podemos observar que  $J(d) = f(d)$ , com  $f(d)$  definida em (3.34). Porém, quando avaliamos (4.4) por MCI essa igualdade já não será verdadeira mais, isso obviamente devido à integral não ser calculada em todo o domínio e ser aproximada por uma soma discreta no domínio de integração.

Para verificarmos a influência de  $n_r$  na aproximação pelo MCI, escolhemos, aleatoriamente,  $m$  pontos  $d$  em  $[-0.5, 4.5]$  e definimos o erro relativo da aproximação por

$$\text{ER} = \sqrt{\frac{\sum_{j=1}^m [f(d^{(j)}) - J(d^{(j)})]^2}{\sum_{j=1}^m f(d^{(j)})^2}}. \quad (4.5)$$

Na Figura (16) temos representado o gráfico de (4.4) para quatro diferentes valores de  $n_r$ . É apresentado também o valor do erro relativo para cada uma das aproximações utilizando  $m = 400$  pontos linearmente distribuídos no domínio de  $d$ .

Podemos observar que para  $n_r = 10$  nossa aproximação consegue captar a suavidade da função, porém o cálculo fica com um ruído muito alto nas partes onde o valor da função é maior. Conforme o valor de  $n_r$  aumenta podemos observar que o ruído começa a diminuir e o erro relativo confirma uma melhor aproximação reduzindo o seu valor. No entanto, pode-se ver a necessidade de um alto esforço computacional para melhorar a aproximação, sendo que com  $n_r = 1000$  ainda não conseguimos um erro menor do que 1%.

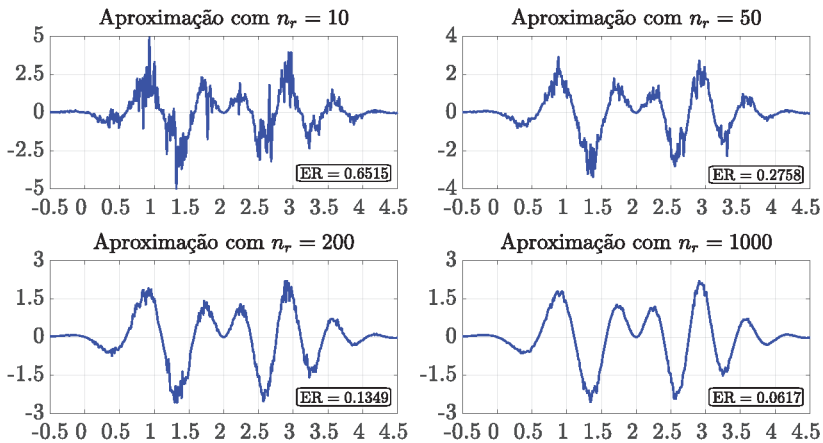


Figura 16 Influência de  $n_r$  para o MCI.

Tomando o ponto onde ocorre o mínimo global,  $d = 1.3403$ , temos que a variância dada por (4.2) para  $n_r = 10, 50, 200, 1000$  será, respectivamente,  $\bar{\sigma}^2 = 0.2644, 0.0265, 0.0070, 0.0014$ . Esse comportamento da variância do erro confirma a afirmação de  $\bar{\sigma}^2$  diminuir conforme  $n_r$  aumenta.

O exemplo anterior serve para mostrar um dos ruídos mais comuns que acontecem em processos computacionais, o ruído multiplicativo. Porém para funções mais complexas, ou os problemas de engenharia citados anteriormente, o ruído acontece durante o processo de construção do campo estocástico do vetor de variáveis aleatórias  $\mathbf{x} = \{x_1, x_2, \dots, x_{k_x}\}$ , onde a quantidade de pontos na amostra para o MCI afeta a qualidade de uma boa aproximação para o valor da função.

## 4.2 KRIGING ESTOCÁSTICO

### 4.2.1 Introdução ao Kriging regressor

Como citado anteriormente, metamodelos também podem ser substitutos para funções sobre a presença de ruídos ou incertezas, porém ajustes devem ser realizados para que os modelos interpoladores passem a ser regressores. Uma das formas mais simples de se filtrar os erros foi assumido na Seção 3.1, que é a regressão polinomial. Ao se usar qualquer substituto para filtrar ruídos, nós assumimos que a função é suave, porém os modelos polinomiais vão além e assumem que a função possui uma forma específica (nesse caso polinomial). Essa abordagem evidentemente leva a um sub ou superajuste do modelo, com as tendências suaves dos dados sendo filtrados junto com o erro (FORRESTER; KEANE; BRESSLOFF, 2006; FORRESTER; SOBESTER; KEANE, 2008).

Para esclarecer melhor, suponha que a função (4.3) dada por (4.4) seja considerada apenas no domínio reduzido  $\mathcal{D} = [-0.5, 1]$ . Tome-mos  $n = 30$  pontos de suporte linearmente distribuídos em  $\mathcal{D}$ , formando o espaço de pontos de suporte  $S$ . Avaliando com (4.1) a função (4.4) utilizando  $n_r = 10$  amostras, temos o vetor de avaliações  $\bar{\mathbf{y}}$  com  $\bar{y}^{(i)} = \bar{J}(d^{(i)})$ .

A Figura 17 apresenta uma aproximação via Kriging Determinístico para a função estocástica. Podemos observar que o Kriging agindo como interpolador passa por todos os pontos de suporte considerando-os como o valor exato da função, não levando em consideração os erros associados. Dessa forma perdemos a tendência da suavidade dos dados quando o valor do erro torna-se maior.

Pode-se ver ainda que o erro RMSE do Kriging é nulo nos pontos de suporte e muito baixo nos demais pontos, mostrando que o Kriging realiza um superajuste do modelo. A nulidade do RMSE nos pontos de suporte é teoricamente errônea pois sabemos que não há certeza no valor da função nestes pontos.

Essa abordagem foi utilizada inicialmente por [Beers e Kleijnen \(2003\)](#) que também utilizam como valor amostral da função a média de um certo número de repetições das avaliações da função no mesmo ponto. Em seu trabalho, os autores mostraram que essa abordagem era ainda melhor do que uma regressão polinomial. Porém, para que o Kriging determinístico possa efetivamente ser melhor, deve-se diminuir a incerteza sobre o valor amostrado da função, e como consequência

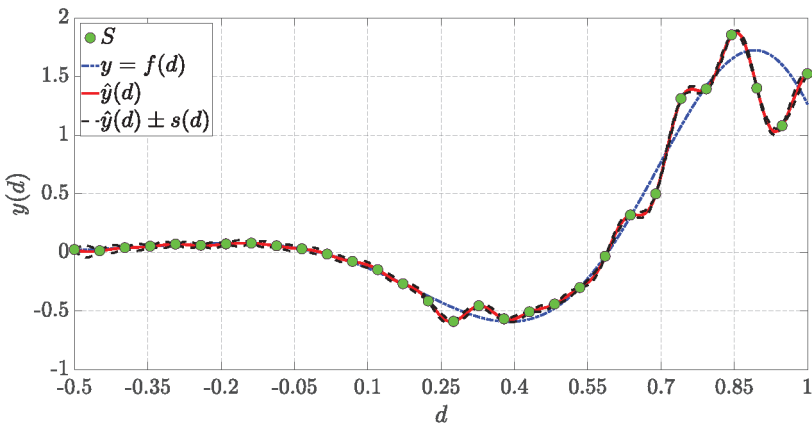


Figura 17 Kriging determinístico para a função estocástica.

elevar o número de repetições no cálculo da função. Sendo assim, essa abordagem se torna inviável para nossos problemas, já que o custo computacional da função é extremamente elevado.

Vemos assim a necessidade de permitir que o Kriging possa ser capaz de realizar uma regressão ao invés de uma interpolação. Durante a derivação do preditor para o Kriging (3.31) na Seção 3.3.4, mostramos como a previsão interpola os dados. Isso ocorre porque em um determinado ponto de suporte, o vetor de correlações do ponto previsto com os dados da amostra,  $\mathbf{h}$  (3.21), é uma coluna da matriz de correlação  $\Psi$  (3.11).

Para sermos capazes de filtrar o ruído da função, uma abordagem comumente utilizada é a proposta por Hoerl e Kennard (1970) e Tikhonov e Arsenin (1977) onde uma constante de regressão (as vezes chamada de constante de regularização (FORRESTER; KEANE; BRESSLOFF, 2006)) pode ser adicionada à diagonal principal de  $\Psi$ , dessa forma passamos a ter  $\Psi + \lambda \mathbf{I}$  (onde  $\mathbf{I}$  é a matriz identidade quadrada de ordem  $n$ ). Com isso, observamos que, conforme  $\|\mathbf{d} - \mathbf{d}^{(i)}\| \rightarrow 0$  temos  $h(\mathbf{d}, \mathbf{d}^{(i)}) = 1 + \lambda$ , assim o vetor  $\mathbf{h}$  deixa de ser uma coluna de  $\Psi$ , e consequentemente os dados não serão mais interpolados.

Segundo Keane e Nair (2005) a constante de regressão deve coincidir com a variância do erro cometido na avaliação da função. Porém, usualmente essa variância é desconhecida. Dessa forma, Forrester, Keane e Bressloff (2006), Huang et al. (2006), Forrester, Sobester e Keane (2008), Ankenman, Nelson e Staum (2010), entre outros autores, propõem que esse parâmetro adicional seja estimado pelo máximo da verossimilhança, junto com os demais parâmetros do Kriging.

Entretanto, a aproximação da função (1.1) por meio do MCI nos proporciona o cálculo da variância do erro na avaliação da função, por meio de (4.2), dessa forma podemos colocar  $\lambda$  como sendo essa variância estimada. Na Seção 4.2.2 mostraremos como essa abordagem altera o modelo tradicional para o Kriging e como podemos ajustar seus parâmetros.

Nos trabalhos de Hoerl e Kennard (1970) e Tikhonov e Arsenin (1977) a constante de regressão é tomada como sendo constante e idêntica para todos os pontos de suporte, ou nos termos de Huang et al. (2006) o erro é suposto como independente, idêntico e normalmente distribuído. Assim a variância do erro não depende do ponto  $\mathbf{d}$ . Quando a variância do erro é considerada dessa forma, dizemos que o problema possui um ruído homogêneo. A técnica de modelagem utilizando essa teoria pode ser encontrada nos trabalhos de Huang et al. (2006) e Picheny et al. (2013).

Entretanto, para cada ponto avaliado por (4.1) podemos obter uma variância diferente, logo a utilização de  $\lambda$  constante não carregaria todas as informações que possuímos acerca dos pontos de suporte. Portanto, para trazer todas as informações disponíveis sobre o erro das avaliações nos pontos de suporte  $\mathbf{d}^{(i)}$ , com  $i = 1, 2, \dots, n$ , podemos definir a matriz de regularização  $\Sigma_\varepsilon$ , quadrada de ordem  $n$ , tal que

$$[\Sigma_\varepsilon]_{ij} = \bar{\sigma}^2(\mathbf{d}^{(i)}) \delta_{ij}, \quad (4.6)$$

onde  $\delta_{ij}$  representa o Delta de Kronecker. De posse da nova matriz de regularização, podemos somá-la à matriz de correlação  $\Psi$  no lugar de  $\lambda \mathbf{I}$ , assim nossa nova matriz de correlação é  $\Psi + \Sigma_\varepsilon$ . Nesse novo cenário, quando  $\|\mathbf{d} - \mathbf{d}^{(i)}\| \rightarrow 0$  temos  $h(\mathbf{d}, \mathbf{d}^{(i)}) = 1 + \bar{\sigma}^2(\mathbf{d}^{(i)})$ , e com isso garantimos a regressão do modelo. Utilizando essa abordagem dizemos que o erro na avaliação da função é heterogêneo.

Outra consequência pode ser analisada quando  $n_r \rightarrow \infty$  tornando a aproximação via MCI o cálculo do valor original da função. Dessa forma, nos pontos de suporte temos que  $\bar{\sigma}^2(\mathbf{d}^{(i)}) \rightarrow 0$  e  $\Sigma_\varepsilon$  se torna a matriz nula de ordem  $n$ , consequentemente transformando a matriz de correlação para  $\Psi$  e retornando o Kriging a interpolador.

Voltando ao caso avaliado anteriormente da função (4.3), supondo as mesmas condições, um novo modelo substituto com o Kriging foi criado, porém agora utilizando a nova matriz de correlação  $\Psi + \Sigma_\varepsilon$ . A Figura 18 apresenta os pontos de suporte  $S$  dos quais agora possuímos



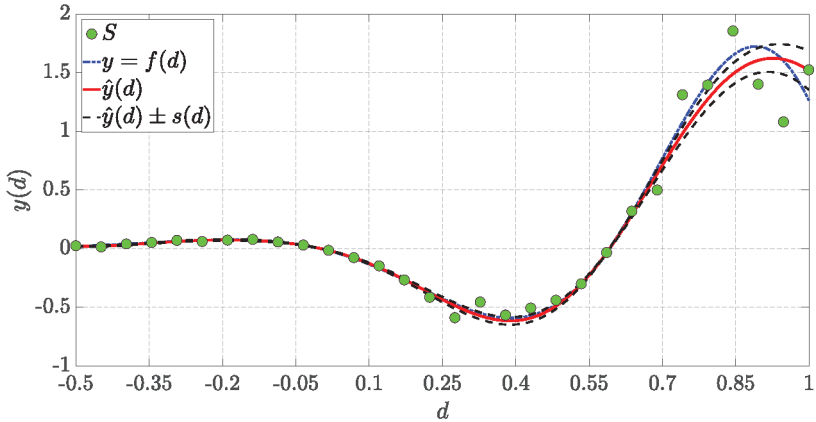


Figura 18 Metamodelo considerando a nova matriz de correlação.

uma variância do erro na aproximação por MCI, a curva da função exata dada por (4.4), a regressão utilizando o Kriging com a nova matriz de correlação e as curvas representando o RMSE para a aproximação.

Podemos ver que agora o regressor conseguiu captar todas as nuances da curva original mantendo a suavidade dos dados. Dessa vez, o RMSE não se anula nos pontos de suporte e as curvas  $\hat{y}(d) \pm s(d)$  se comportam mantendo a mesma tendência que os pontos de suporte ditam. A não nulidade do RMSE é mais uma comprovação da regressão do Kriging, onde agora temos computado para todo ponto de suporte um erro intrínseco ao valor da função no ponto, e este erro ocorre devido à realização do processo estocástico na aproximação da função.

Uma quantificação da qualidade das aproximações apresentada nas Figuras 17 e 18 pode ser realizada por meio do erro relativo dado por (4.5). Tomando  $m = 200$  pontos uniformemente distribuídos sobre o espaço  $\mathcal{D} = [-0.5, 1]$ , temos que o erro relativo para a aproximação interpoladora é de  $ER = 0.1896$  e para o caso regressor é de  $ER = 0.0820$ . Logo, mesmo a interpolação tendo dado um erro relativo baixo, temos que a regressão possibilitou reduzir o erro relativo da aproximação em

cerca de 56.75%.

### 4.2.2 O Kriging Estocástico - SK

Para podermos ser mais formais quanto ao Kriging regressor, que agora passará a ser chamado de Kriging Estocástico (*Stochastic Kriging* - SK) iremos tomar como base o trabalho de [Ankenman, Nelson e Staum \(2010\)](#), que efetivamente propuseram a extensão do Kriging determinístico para lidar com processos estocásticos.

No trabalho citado, são definidas duas diferentes fontes de incerteza envolvidas na regressão: a primeira é a Incerteza Intrínseca que está relacionada à natureza estocástica da avaliação da função objetivo e a segunda é a Incerteza Extrínseca vinculada a natureza estocástica imposta para o processo Gaussiano da superfície de resposta. Seguindo essas premissas, o seguinte modelo é apresentado por Ankenman para representar o processo estocástico de uma replicação  $j$  da função no ponto  $\mathbf{d}$

$$y_j(\mathbf{d}) = M(\mathbf{d}) + Z(\mathbf{d}) + \varepsilon_j(\mathbf{d}). \quad (4.7)$$

O termo  $M(\mathbf{d})$  como visto para o Kriging determinístico é responsável por mostrar a tendência geral da resposta. A parcela  $Z(\mathbf{d})$  será a representante da incerteza extrínseca do modelo e continuará sendo um processo estocástico com média zero e variância  $\sigma_z^2$ , mantendo a correlação espacial entre os pontos de suporte.

A última parcela,  $\varepsilon_j(\mathbf{d})$ , é a representação da incerteza intrínseca e a responsável por fazer diferir o modelo do SK para o modelo determinístico. Em seu trabalho, [Ankenman, Nelson e Staum \(2010\)](#) consideram que a sequência de erros intrínsecos  $\varepsilon_1(\mathbf{d}), \varepsilon_2(\mathbf{d}), \dots, \varepsilon_{n_r}(\mathbf{d})$ , para cada uma das  $n_r$  replicações do valor da função em  $\mathbf{d}$ , é naturalmente independente, idêntica e normalmente distribuída. Aqui, é permitido que a variância do erro não seja constante. Também pode ser considerada uma certa correlação espacial entre erros de pontos distintos, ou seja, que  $\text{cor}[\varepsilon(\mathbf{d}^{(i)}), \varepsilon(\mathbf{d}^{(j)})] > 0$ . Todavia, para poder realizar as estimações

dos parâmetros, os autores fazem a suposição de que a correlação supracitada seja zero, considerando que erros em pontos distintos sejam independentes. Como em nossa proposta introduzimos o erro por meio da aproximação via MCI, para garantir que não haja correlação espacial entre erros para pontos distintos, consideramos que as amostras  $\mathbf{x}^{(i)}$  utilizadas para a aproximação por MCI sejam independentes.

Iremos considerar que um ponto de suporte é constituído pelo par  $(\mathbf{d}^{(i)}, n_r)$  onde  $i = 1, 2, \dots, n$  e  $n_r$  é o número de replicações no cálculo aproximado de MCI. Denotaremos por

$$\bar{y}^{(i)} = \bar{y}(\mathbf{d}^{(i)}) = \bar{J}(\mathbf{d}^{(i)}) \quad (4.8)$$

a aproximação do valor da função por (4.1) e deixamos

$$\bar{\mathbf{y}} = \left\{ \bar{y}^{(1)}, \bar{y}^{(2)}, \dots, \bar{y}^{(n)} \right\}^T \quad (4.9)$$

representar o vetor de avaliações da função nos  $n$  pontos de suporte.

**Ankenman, Nelson e Staum (2010)** propuseram que as predições realizadas pelo metamodelo fossem dadas por  $Y(\mathbf{d}^+) = M(\mathbf{d}^+) + Z(\mathbf{d}^+)$  para qualquer  $\mathbf{d}^+$  amostral ou não. Por simplificação, os autores consideraram  $M(\mathbf{d}) = \mu$ ; dessa forma, somente uma constante irá representar a tendência geral da resposta do modelo substituto.

Seja  $\Sigma_Z$  a matriz de covariância obtida pela correlação espacial extrínseca entre todos os pontos de suporte  $\mathbf{d}^{(1)}, \mathbf{d}^{(2)}, \dots, \mathbf{d}^{(n)}$ , ou seja,

$$\begin{aligned} \Sigma_Z \left( \mathbf{d}^{(i)}, \mathbf{d}^{(j)} \right) &= \text{Cov} \left[ Z(\mathbf{d}^{(i)}), Z(\mathbf{d}^{(j)}) \right] \\ &= \sigma_z^2 h(\mathbf{d}^{(i)}, \mathbf{d}^{(j)}), \end{aligned}$$

onde  $h(\mathbf{d}^{(i)}, \mathbf{d}^{(j)})$  é a mesma base radial definida por (3.10). Seja  $\Sigma_Z(\mathbf{d}^+, \cdot)$  o vetor  $n \times 1$  formado pelas covariâncias entre o ponto predito e os pontos de suporte. Por fim, coloquemos  $\Sigma_\varepsilon$  como sendo a matriz  $n \times n$  de covariância obtida pelo ruído intrínseco entre todos os pontos de suporte. Utilizando o mínimo do MSE, **Ankenman, Nelson e Staum (2010)** mostraram que o melhor preditor não viciado é dado por

$$Y(\mathbf{d}^+) = \mu + \Sigma_Z(\mathbf{d}^+, \cdot)^T [\Sigma_Z + \Sigma_\varepsilon]^{-1} (\bar{\mathbf{y}} - \mu \mathbf{1}), \quad (4.10)$$

onde  $\mathbf{1}$  é um vetor  $n \times 1$  de uns. Esse é o preditor chamado por ele de SK. Podemos observar que na ausência de incerteza intrínseca, ou seja, na ausência de ruído, o modelo (4.10) é reduzido ao modelo (3.31) do Kriging determinístico. Ankenman, Nelson e Staum (2010) também mostram que o melhor MSE para esse preditor é dado por

$$s^2(\mathbf{d}^+) = \Sigma_Z(\mathbf{d}^+, \mathbf{d}^+) - \Sigma_Z(\mathbf{d}^+, \cdot)^T [\Sigma_Z + \Sigma_\varepsilon]^{-1} \Sigma_Z(\mathbf{d}^+, \cdot). \quad (4.11)$$

### 4.2.3 Estimação dos parâmetros

Para efetivamente construirmos um metamodelo utilizando o SK, devemos realizar as estimações dos parâmetros desconhecidos, a saber,  $\mu$ ,  $\sigma_z^2$ ,  $\boldsymbol{\theta}$  (continuaremos supondo que  $\mathbf{p} = 2$ ) e  $\Sigma_\varepsilon$ . Novamente utilizamos a maximização da verossimilhança para garantirmos a maior probabilidade de se obter a resposta  $Y(\mathbf{d}^+)$ , dado o vetor  $\bar{\mathbf{y}}$ .

Um problema que antes deve ser resolvido é a estimação da matriz de covariância para o erro intrínseco. Ankenman, Nelson e Staum (2010) citam que naturalmente a variância do erro intrínseco, denotada por  $V(\mathbf{d}) = \text{Var}[\varepsilon(\mathbf{d})]$ , é desconhecida. Dessa forma os autores propõem uma solução baseada na criação de um Kriging determinístico que interpole os valores das variâncias estimadas para a replicação dos valores da função no mesmo ponto. Ou seja, suponha que sejam realizadas  $n_r^{(i)}$  replicações do valor da função  $y^{(i)}$  no ponto  $\mathbf{d}^{(i)}$ , então ele supõe que a variância nos pontos de suporte pode ser aproximada por

$$\mathcal{V}^2(\mathbf{d}^{(i)}) = \frac{1}{n_r^{(i)} - 1} \sum_{j=1}^{n_r^{(i)}} \left( y_j^{(i)} - \bar{y}^{(i)} \right)^2,$$

e a partir destes valores os autores criam um metamodelo em Kriging que interpola essas variâncias. Portanto, obtemos uma estimação onde  $\hat{V}(\mathbf{d}^{(i)}) = \mathcal{V}^2(\mathbf{d}^{(i)})$  para a variância do erro nos pontos de suporte.

Após a obtenção de uma estimação para as variâncias, é definida a seguinte estimação para os termos da matriz de covariância do erro

intrínseco:

$$\left[ \hat{\Sigma}_\varepsilon \right]_{ij} = \frac{\hat{V}(\mathbf{d}^{(i)})}{n_r^{(i)}} \delta_{ij}. \quad (4.12)$$

Podemos ver que a matriz de covariância estimada por [Ankenman, Nelson e Staum \(2010\)](#) é a mesma que supomos em (4.6), com  $\hat{V}(\mathbf{d}^{(i)})/n_r^{(i)} = \hat{\sigma}^2(\mathbf{d}^{(i)})$ .

Assumindo então que a matriz de covariância para o erro é conhecida, calculamos os demais parâmetros maximizando a função logaritmo da verossimilhança (como apresentado em (3.14) )

$$L_{\ln}(\mu, \sigma_z^2, \boldsymbol{\theta}) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} (\bar{\mathbf{y}} - \mu \mathbf{1})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{y}} - \mu \mathbf{1}), \quad (4.13)$$

onde  $\boldsymbol{\Sigma} = \sigma_z^2 \boldsymbol{\Psi} + \boldsymbol{\Sigma}_\varepsilon$  sendo  $\boldsymbol{\Psi}$  a matriz de correlação espacial do termo extrínseco  $Z(\mathbf{d})$  entre todos os pontos de suporte.

Para obtenção dos estimadores que maximizam (4.13) devemos resolver simultaneamente as seguintes equações não lineares

$$\frac{\partial L_{\ln}(\mu, \sigma_z^2, \boldsymbol{\theta})}{\partial \mu} = 0 \quad \frac{\partial L_{\ln}(\mu, \sigma_z^2, \boldsymbol{\theta})}{\partial \sigma_z^2} = 0 \quad \frac{\partial L_{\ln}(\mu, \sigma_z^2, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0} \quad (4.14)$$

e encontrar os valores  $\hat{\mu}$ ,  $\hat{\sigma}_z^2$  e  $\hat{\boldsymbol{\theta}}$ . Como apresentado no Apêndice B, vemos que as equações acima formam um sistema não linear de  $k + 2$  incógnitas e  $k + 2$  equações ( $k$  é o número de dimensões do vetor  $\boldsymbol{\theta}$ ), onde cada equação está em função dos três parâmetros. Dessa forma não conseguimos reproduzir o feito na Seção 3.3.3 para obtenção dos parâmetros.

[Ankenman, Nelson e Staum \(2010\)](#) utilizam para o processo de obtenção dos parâmetros algum tipo de algoritmo não linear de otimização aplicado diretamente sobre a função (4.13). Nesse trabalho tomamos um caminho alternativo, utilizando o que foi proposto para o Kriging determinístico. Dessa forma, será utilizado um algoritmo metaheurístico onde os parâmetros a serem ajustados serão  $\sigma_z^2$  e  $\boldsymbol{\theta}$  e o objetivo será minimizar o negativo da função (4.13). Para o cálculo de  $\mu$

utilizamos a sua dependência aos outros parâmetros, como apresentado pela equação (B.6).

De posse dos estimadores  $\hat{\mu}$ ,  $\hat{\sigma}_z^2$  e  $\hat{\boldsymbol{\theta}}$  temos que o preditor é dado pelo metamodelo

$$\begin{aligned}\hat{Y}(\mathbf{d}^+) &= \hat{\mu} + \hat{\sigma}_z^2 \mathbf{h}^T \left[ \hat{\sigma}_z^2 \boldsymbol{\Psi} + \hat{\boldsymbol{\Sigma}}_\varepsilon \right]^{-1} (\bar{\mathbf{y}} - \hat{\mu} \mathbf{1}) \\ &= \hat{\mu} + \hat{\sigma}_z^2 \mathbf{h}^T \hat{\boldsymbol{\Sigma}}^{-1} (\bar{\mathbf{y}} - \hat{\mu} \mathbf{1})\end{aligned}\quad (4.15)$$

onde  $\mathbf{h}$  é o vetor  $n \times 1$  com as correlações espaciais entre os pontos de suporte e o ponto  $\mathbf{d}^+$  e  $\hat{\boldsymbol{\Sigma}} = \hat{\sigma}_z^2 \boldsymbol{\Psi} + \hat{\boldsymbol{\Sigma}}_\varepsilon$ . O erro MSE ocorrido na predição é dado por

$$\begin{aligned}s^2(\mathbf{d}^+) &= \hat{\sigma}_z^2 - (\hat{\sigma}_z^2)^2 \mathbf{h}^T \left[ \hat{\sigma}_z^2 \boldsymbol{\Psi} + \hat{\boldsymbol{\Sigma}}_\varepsilon \right]^{-1} \mathbf{h} + \frac{\Delta^2}{\mathbf{1}^T \left[ \hat{\sigma}_z^2 \boldsymbol{\Psi} + \hat{\boldsymbol{\Sigma}}_\varepsilon \right]^{-1} \mathbf{1}} \\ &= \hat{\sigma}_z^2 - (\hat{\sigma}_z^2)^2 \mathbf{h}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{h} + \frac{\Delta^2}{\mathbf{1}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{1}},\end{aligned}\quad (4.16)$$

onde  $\Delta = 1 - \mathbf{1}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\sigma}_z^2 \mathbf{h}$ .

Para verificarmos como o SK se comporta no campo estocástico, voltemos à função (4.4) dada por (4.3) no domínio  $\mathcal{D} = [-0.5, 4.5]$  onde o parâmetro aleatório  $X$  segue a distribuição  $\mathcal{N}(0, 0.5)$ . Tomando  $n = 17$  pontos de suporte e utilizando  $n_r = 5$  replicações para a aproximação da função em cada ponto de suporte, a Figura 19 apresenta o modelo substituto construído pelo SK. Na figura da esquerda temos o espaço amostral  $S$ , o gráfico da função  $J(d)$  aproximada via MCI (foram utilizadas também 5 replicações para construção desse gráfico) e o modelo SK. Selecionamos com um retângulo uma região do gráfico da esquerda que foi ampliado e está apresentado no gráfico da direita. Essa ampliação mostra que efetivamente o SK atua como regressor e não como interpolador, mostrando que as curvas  $\hat{Y}(d) \pm s(d)$  não se anulam nos pontos de suporte.

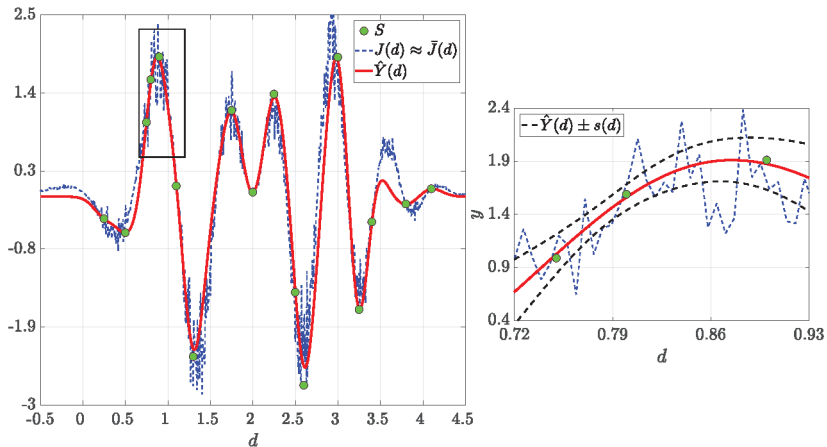


Figura 19 Aproximação via SK.

### 4.3 sEGO PARA FUNÇÕES ESTOCÁSTICAS

Nesta seção vamos efetivamente resolver o problema proposto em (1.2). A abordagem aqui é a mesma que foi utilizada para o Kriging determinístico, o EGO. Então as etapas que foram descritas na Seção 3.4 são consideradas de forma idêntica nessa seção.

Como visto anteriormente, o uso do SK para modelar funções estocásticas ajusta um regressor aos pontos de suporte atuando como um filtro para os ruídos do processo. Dessa forma somos capazes de ver as suavidades e tendências gerais da função. Embora o SK ofereça um bom modelo para a filtragem de ruído, as estimativas do erro no preditor (MSE) não são mais adequadas para uso na escolha de novos pontos de suporte, os IPs (FORRESTER; KEANE; BRESSLOFF, 2006).

Um dos critérios propostos por Jones, Schonlau e Welch (1998) é o EI, Seção 3.4.1, onde o princípio básico é procurar pelo ponto  $\mathbf{d}$  que maximiza a expectativa de melhorarmos  $\hat{Y}$  em relação ao valor mínimo atual  $f_{\min}$ , quando consideramos  $Y(\mathbf{d})$  como sendo uma variável aleatória de média  $\hat{Y}(\mathbf{d})$  e desvio  $s(\mathbf{d})$ . Esse conceito está apresentado matematicamente pelas equações (3.35) e (3.42).

Como provado por [Schonlau \(1997\)](#) o EI convergirá naturalmente para o ótimo global. Porém para que esse teorema seja satisfeito uma condição básica é de que o MSE seja positivo para pontos que não são de suporte e zero para pontos de suporte. Com essas duas condições satisfeitas não há o risco de um ponto de suporte ser selecionado novamente como IP.

Entretanto, quando utilizamos de funções estocásticas com a presença de ruídos ou incertezas, o metamodelo via SK não satisfaz essa condição, já que para pontos de suporte o MSE não será mais nulo (veja a equação (4.11) ou a Figura 19). Dessa forma, se aplicássemos o EI como definido por [Jones, Schonlau e Welch \(1998\)](#), pontos de suporte poderiam ser considerados como IP e facilmente o processo poderia ficar preso em uma exploração local.

[Huang et al. \(2006\)](#) afirmam que a dificuldade de se saber o valor exato da função objetivo acarreta em duas dificuldades para o EI:

- O melhor minimizador atual  $\mathbf{d}^*$  nem sempre será bem definido, pois duas avaliações diferentes em  $\mathbf{d}^*$  poderão resultar em valores diferentes;
- A métrica EI não possui em sua formulação a incerteza associada ao valor mínimo da função  $J_{\min} = J(\mathbf{d}^*)$ .

Para contornar essas deficiências, vários autores vêm tentando estender o EGO diretamente para as simulações estocásticas. Porém, muitas das abordagens assumem que o erro é homogêneo, ou seja, a variância do erro não depende do valor de  $\mathbf{d}$ . As abordagens mais tradicionais nesse contexto podem ser encontradas nos trabalhos de [Huang et al. \(2006\)](#), [Forrester, Keane e Bressloff \(2006\)](#), [Picheny, Wagner e Ginsbourger \(2012\)](#), [Picheny et al. \(2013\)](#), [Picheny, Wagner e Ginsbourger \(2013\)](#), [Picheny e Ginsbourger \(2014\)](#). Para o caso de erro heterogêneo, é possível consultar algumas métricas em [Jalali, Nieuwenhuys e Picheny \(2017\)](#).



Neste trabalho, pretendemos realizar a comparação de algumas das técnicas mais utilizadas na literatura e como elas atuam dentro do modelo de SK proposto para erros heterogêneos. Utilizamos as seguintes métricas:

- ✓ Critério do Percentil Mínimo (*Minimal Quantile Criteria* - MQ): proposto por [Picheny, Wagner e Ginsbourger \(2013\)](#) para ruídos homogêneos e estendido a ruídos heterogêneos por [Jalali, Nieuwenhuys e Picheny \(2017\)](#);
- ✓ Melhora Esperada Aumentada (*Augmented Expected Improvement* - AEI): proposto por [Huang et al. \(2006\)](#) para ruídos homogêneos e estendido a ruídos heterogêneos por [Jalali, Nieuwenhuys e Picheny \(2017\)](#);
- ✓ Percentil de Melhora Esperada (*Expected Quantile Improvement* - EQI): proposto por [Picheny, Wagner e Ginsbourger \(2013\)](#) para ruídos homogêneos e estendido a ruídos heterogêneos por [Jalali, Nieuwenhuys e Picheny \(2017\)](#);
- ✓ Otimização Sequencial de Dois Estágios (*Two-Stage Sequential Optimization* - TSSO): proposto por [Quan et al. \(2013\)](#) para ruídos heterogêneos;
- ✓ Melhora Esperada com Reinterpolação (*Expected Improvement with Reinterpolation* - EIR): proposto por [Forrester, Keane e Bressloff \(2006\)](#) para ruídos homogêneos.

Cabe ressaltar aqui uma importante contribuição deste trabalho. O método EIR até o momento não foi apresentado dentro do contexto de ruídos heterogêneos. Porém, vemos que ele pode possuir grande aplicabilidade nesse cenário, já que é uma extensão natural do EI. Portanto, neste texto é apresentado como podemos realizar os mesmos procedimentos feitos por [Forrester, Keane e Bressloff \(2006\)](#) e acoplar as parcelas do ruído heterogêneo na métrica EIR. Essa atitude se justifica,

pois o EIR irá ajustar o sEGO para trabalhar como realizado no caso determinístico; assim, espera-se que o sEGO atue como o próprio EGO com o EI, que possui um excelente balanço entre exploração local e global.

Cada técnica foi desenvolvida para um determinado fim, logo, para mais detalhes além dos que são apresentados aqui, sugerimos a consulta dos respectivos artigos que as definem. No nosso texto, apresentamos apenas os princípios básicos e formulações.

Por uma questão de notação, utilizaremos  $\mathbf{d}^{n+1}$  para representar a solução do subproblema de otimização da métrica que determina o próximo IP a ser adicionado.

### 4.3.1 MQ

Este critério surge de forma mais natural para processos estocásticos do que o EI para o caso determinístico. Ele foi proposto inicialmente por [Cox e John \(1997\)](#) e consiste em realizar um balanço entre a exploração global (altas variâncias do SK) e exploração local (baixa média preditiva do SK) utilizando de um percentil do valor do SK, ou seja, utilizando uma soma ponderada entre  $\hat{Y}(\mathbf{d})$  e  $s^2(\mathbf{d})$ .

Basicamente o método escolhe como próximo IP o ponto que minimiza ou a média ou um percentil do valor predito pelo SK. Logo, a função objetivo para essa métrica será dada por

$$\text{MQ}(\mathbf{d}) = \hat{Y}(\mathbf{d}) + \Phi^{-1}(\beta)s(\mathbf{d}) \quad (4.17)$$

e o próximo IP será

$$\mathbf{d}^{n+1} = \underset{\mathbf{d} \in \mathcal{D}}{\text{argmin}} \text{MQ}(\mathbf{d}). \quad (4.18)$$

Na equação (4.17) temos que  $\hat{Y}(\mathbf{d})$  é dado por (4.15),  $s(\mathbf{d})$  é o RMSE obtido pela raiz quadrada de (4.16). A função  $\Phi^{-1}$  é a inversa da densidade de probabilidade cumulativa da distribuição normal. [Jalali, Nieuwenhuys e Picheny \(2017\)](#) indica uma probabilidade  $\beta \in (0, 0.5]$ .

Jones (2001) mostrou que essa abordagem é menos efetiva do que o EI para casos determinísticos. Porém Picheny, Wagner e Ginsbourger (2013) a utilizam em suas simulações, já que não é necessária nenhuma modificação específica para o caso de ruídos.

Duas características cabem destaque: primeiro, pode-se ver que essa técnica não necessita de nenhuma informação sobre a variância da parcela intrínseca do modelo em pontos que não são de suporte. Em segundo, vemos que é permitido que um ponto de suporte seja considerado como IP. Caso isso aconteça, novas replicações do valor da função objetivo serão simuladas para esse ponto de suporte e dessa forma a variância do erro neste ponto irá diminuir, o que fará com que  $s(\mathbf{d})$  também diminua.

### 4.3.2 AEI

Esse critério tem por objetivo utilizar uma nova abordagem para o EI no caso determinístico. Para tanto serão levados em consideração as incertezas acerca da avaliação da função e da predição pelo SK. Inicialmente o método foi proposto para lidar com ruídos homogêneos, logo, para nosso intuito, a apresentação do método será realizada de forma a tratar os problemas com ruídos heterogêneos.

Temos que para o EI uma melhora para o valor predito é dado por

$$I = \max\{f_{\min} - Y, 0\},$$

onde o valor de  $f_{\min}$  é tomado como sendo o menor valor da função nos pontos de suporte. Entretanto, para casos estocásticos essa definição não carrega uma informação correta acerca de  $f_{\min}$ , pois seu valor estará sujeito a incertezas. Portanto, uma definição mais natural para o caso estocástico será

$$I = \max\{\hat{Y}(\mathbf{d}^{**}) - Y, 0\}, \quad (4.19)$$

onde  $\mathbf{d}^{**}$  é chamado de Melhor Solução Efetiva (*Effective Best Solution*). O valor de  $\mathbf{d}^{**}$  é definido como sendo o ponto de suporte (ponto já

simulado) que minimiza o valor de (4.17) para  $\beta \in [0.5, 1)$ . Assim, temos que o EI será

$$\mathbb{E}(I(\mathbf{d})) = \left( \hat{Y}(\mathbf{d}^{**}) - \hat{Y}(\mathbf{d}) \right) \Phi \left( \frac{\hat{Y}(\mathbf{d}^{**}) - \hat{Y}(\mathbf{d})}{s(\mathbf{d})} \right) + s(\mathbf{d}) \phi \left( \frac{\hat{Y}(\mathbf{d}^{**}) - \hat{Y}(\mathbf{d})}{s(\mathbf{d})} \right). \quad (4.20)$$

Notamos que a equação anterior não se anula quando  $\mathbf{d} = \mathbf{d}^{**}$ , já que  $s(\mathbf{d}^{**}) > 0$  na segunda parcela da soma. Essa é uma excelente qualidade na medida para o EI, pois permite que pontos de suporte possam ser escolhidos como IP; dessa forma, novas replicações poderão ser adicionadas ao valor da função, diminuindo a variância do erro. Temos também que (4.20) se reduz a (3.42) na ausência de ruídos,  $\Sigma_\varepsilon = \mathbf{0}$ .

Ao mesmo tempo que novas replicações são positivas para poder diminuir a variância do erro cometido na aproximação, deve-se evitar que a métrica fique presa somente a pontos de suporte. Dessa forma definimos o AEI como sendo

$$\text{AEI}(\mathbf{d}) = \mathbb{E}(I(\mathbf{d})) \left( 1 - \frac{\bar{\sigma}(\mathbf{d})}{\sqrt{s^2(\mathbf{d}) + \bar{\sigma}^2(\mathbf{d})}} \right), \quad (4.21)$$

onde  $\mathbb{E}(I(\mathbf{d}))$  é definido por (4.20),  $\bar{\sigma}^2(\mathbf{d})$  é a variância da aproximação via MCI (4.2) e  $s^2(\mathbf{d})$  é o MSE dado por (4.16). Dessa forma, o próximo IP será definido como

$$\mathbf{d}^{n+1} = \underset{\mathbf{d} \in \mathcal{D}}{\operatorname{argmax}} \text{AEI}(\mathbf{d}). \quad (4.22)$$

A segunda parcela da métrica AEI (4.21) atua como uma penalização para o valor do novo EI, prevenindo que a métrica fique presa em um determinado ponto  $\mathbf{d}$ . Teoricamente isso acontece pois, com novas replicações,  $s^2(\mathbf{d})$  tenderá a zero, fazendo com que o AEI também tenda a zero.

Um detalhe importante sobre a métrica AEI para o caso heterogêneo é a necessidade de se saber qual a variância intrínseca  $\bar{\sigma}^2(\mathbf{d})$  no ponto  $\mathbf{d}$ . Supondo que  $\mathbf{d}$  não é um ponto de suporte, então esse valor não foi estimado por (4.2). Logo, como não temos uma lei explícita para  $\bar{\sigma}^2(\mathbf{d})$ , deveremos utilizar outra forma de se determinar o valor dessa variância. Como mostrado na Seção 4.4.1, este valor pertence ao contradomínio de uma função limitada e pode ser estimado pela abordagem adaptativa proposta por Carraro (2017).

### 4.3.3 EQI

A principal ideia por trás da métrica EQI é que, para funções estocásticas, o preditor SK da função pode ser mais próximo do valor real do que os dados originais. Logo, uma melhora no valor predito deve ser dada levando em consideração o efeito dessa nova observação inclusa em um modelo futuro.

Ao invés de considerarmos as observações com ruídos, o mais direto a se fazer é utilizar o percentil do SK tomado por (4.17), aqui reescrito como

$$q_n(\mathbf{d}) = \hat{Y}_n(\mathbf{d}) + \Phi^{-1}(\beta)s_n(\mathbf{d}), \quad (4.23)$$

onde o subíndice  $n$  se refere ao modelo atual com  $n$  pontos e  $\beta \in [0.5, 1)$ . Dessa forma, podemos definir que uma melhora para a avaliação do ponto será dada por

$$I(\mathbf{d}) = \max\{q_{\min} - q_{n+1}(\mathbf{d}), 0\}, \quad (4.24)$$

onde  $q_{\min}$  é o menor valor que  $q_n(\mathbf{d})$  assume somente nos  $n$  pontos de suporte e  $q_{n+1}(\mathbf{d})$  é o percentil do SK (4.23) atualizado com o novo ponto  $\mathbf{d}$  na amostra.

A métrica EQI é definida como

$$\text{EQI}(\mathbf{d}) = \mathbb{E} \left[ I(\mathbf{d}) | Y_n(\mathbf{d}^{(i)}) = \bar{y}^{(i)}, i = 1, 2, \dots, n \right], \quad (4.25)$$

ou seja, é a esperança de melhora condicionada às  $n$  observações já realizadas e sob o fato de que com uma nova medida sendo realizada no

ponto  $\mathbf{d}$ , o valor do modelo  $Y(\mathbf{d})$  a partir das  $n$  observações será uma variável aleatória normal com média  $\hat{Y}_n(\mathbf{d})$  e variância  $s_n^2(\mathbf{d}) + \bar{\sigma}^2(\mathbf{d})$ .

Picheny et al. (2013) mostraram que a distribuição condicionada de  $q_{n+1}(\mathbf{d})$  é analiticamente tratável, e que a métrica EQI é dada por

$$\text{EQI}(\mathbf{d}) = \left( q_{\min} - \hat{Y}_q(\mathbf{d}) \right) \Phi \left( \frac{q_{\min} - \hat{Y}_q(\mathbf{d})}{s_q(\mathbf{d})} \right) + s_q(\mathbf{d}) \phi \left( \frac{q_{\min} - \hat{Y}_q(\mathbf{d})}{s_q(\mathbf{d})} \right), \quad (4.26)$$

onde  $\hat{Y}_q(\mathbf{d})$  e  $s_q(\mathbf{d})$  serão a média e o desvio padrão, respectivamente, do futuro percentil  $q_{n+1}(\mathbf{d})$ . Na prática teremos que

$$\hat{Y}_q(\mathbf{d}) = \hat{Y}_n(\mathbf{d}) + \Phi^{-1}(\beta) \sqrt{\frac{\bar{\sigma}^2(\mathbf{d}) s_n^2(\mathbf{d})}{\bar{\sigma}^2(\mathbf{d}) + s_n^2(\mathbf{d})}} \quad (4.27)$$

$$s_q(\mathbf{d}) = \frac{s_n^2(\mathbf{d})}{\sqrt{\bar{\sigma}^2(\mathbf{d}) + s_n^2(\mathbf{d})}}. \quad (4.28)$$

Sendo assim, o próximo IP será definido como

$$\mathbf{d}^{n+1} = \operatorname{argmax}_{\mathbf{d} \in \mathcal{D}} \text{EQI}(\mathbf{d}). \quad (4.29)$$

Pode-se ver que a métrica EQI também permite selecionarmos pontos de suporte como IP, sendo possível novas replicações. No entanto, Picheny et al. (2013) vão além e permitem que, após a adição do IP e do reajuste do modelo, uma nova etapa chamada de *Replication Step* entre em ação. Nela, é verificado se novas replicações no IP que foi encontrado irão causar alguma evolução no valor do EQI. Se assim o causar, então ele adiciona essa quantidade de replicação ao valor do ponto e atualiza o valor da função, da variância do ruído e reajusta o modelo. Para nossas simulações, decidimos por pular o *Replication Step*, já que a quantidade de replicações de um ponto será decidido pela variância do MCI, como explanado na próxima seção.

Por fim, vemos que a métrica EQI também é dependente do valor da variância em pontos que não são de suporte  $\bar{\sigma}^2(\mathbf{d})$ . Dessa forma, iremos agir como explicado para o caso AEI.

#### 4.3.4 TSSO

O TSSO é um algoritmo que consiste em duas etapas. Após a criação do espaço amostral inicial e do primeiro ajuste via SK para os dados, cada iteração subsequente será composta por uma Etapa de Pesquisa (*Search Stage*) seguido por uma Etapa de Alocação do recurso computacional (*Allocation Stage*).

A primeira etapa do algoritmo utiliza a Melhora Esperada Modificada (*Modified Expected Improvement* - MEI) como uma métrica para a seleção de novos IPs. Na segunda etapa, a técnica Alocação Ótima de Recurso Computacional (*Optimal Computing Budget Allocation* - OCBA) é aplicada para distribuir um número adicional de replicações do valor da função objetivo entre os pontos de suporte. Especificamente, a *Search Stage* do algoritmo é responsável por identificar potenciais pontos ótimos globais, enquanto o *Allocation Stage* busca reduzir a incerteza devido à variabilidade aleatória nos pontos de suporte, com os objetivos de melhorar o ajuste do modelo em regiões que contêm mínimos locais e, eventualmente, selecionar corretamente o ponto ótimo global.

Neste trabalho é utilizada uma versão modificada desse algoritmo. Utilizamos para a primeira etapa (*Search Stage*) o mesmo que sugerem [Quan et al. \(2013\)](#). Porém a segunda etapa (*Allocation Stage*) é diferente da abordagem original. Ao invés de se utilizar o OCBA, utilizamos a pesquisa adaptativa proposta por [Carraro et al. \(2019\)](#) e apresentada na Seção 4.4. Dessa forma, o método ainda continua sendo em duas fases e com as mesmas premissas do algoritmo original.

Para a etapa de pesquisa utilizamos uma modificação para o EI, como proposto por [Quan et al. \(2013\)](#), onde tomamos como melhora

$$I(\mathbf{d}) = \max\{\hat{Y}_{\min} - Y(\mathbf{d}), 0\}, \quad (4.30)$$

sendo  $\hat{Y}_{\min}$  a predição pelo SK (4.15) no ponto de suporte que possui a menor aproximação via MCI da função. O termo  $Y(\mathbf{d})$  será a repre-

sentação de uma variável aleatória normal definida pela média  $\hat{Y}(\mathbf{d})$  obtida pelo preditor do SK (4.15) e variância  $s^2(\mathbf{d})$  dada pelo MSE do Kriging determinístico (3.33). Assim, a métrica MEI será dada por

$$\begin{aligned} \text{MEI}(\mathbf{d}) &= \mathbb{E}(I(\mathbf{d})) \\ &= \left( \hat{Y}_{\min} - Y(\mathbf{d}) \right) \Phi \left( \frac{\hat{Y}_{\min} - Y(\mathbf{d})}{s(\mathbf{d})} \right) + \\ &\quad s(\mathbf{d}) \phi \left( \frac{\hat{Y}_{\min} - Y(\mathbf{d})}{s(\mathbf{d})} \right), \end{aligned} \quad (4.31)$$

e o próximo IP será definido como

$$\mathbf{d}^{n+1} = \underset{\mathbf{d} \in \mathcal{D}}{\text{argmax}} \text{MEI}(\mathbf{d}). \quad (4.32)$$

O uso do preditor SK para a média da variável aleatória  $Y(\mathbf{d})$  é uma escolha razoável, já que fornece uma previsão imparcial do valor da função, dada a natureza heterogênea do ruído. O que diferencia o MEI dos critérios EI anteriores é o tratamento da incerteza do preditor. Como a segunda etapa é destinada a refinar o ruído intrínseco nas avaliações dos IPs, somente a variância  $s^2(\mathbf{d})$  do caso determinístico (3.33) é utilizada na etapa de pesquisa. Isso permite que a pesquisa de IPs se concentre em regiões promissoras com altos valores de melhora e que reduzam a incerteza extrínseca do metamodelo. Além disso, ignorando a incerteza intrínseca do preditor causada pela variabilidade aleatória, a função MEI assume que as observações são feitas com precisão infinita, de modo que o mesmo ponto nunca é selecionado novamente. Isso permite que o algoritmo escape rapidamente de um ótimo local e se aproxime do mesmo comportamento do EI original e de seu balanço entre exploração global e exploração local.

#### 4.3.5 EIR

O método será chamado de *Expected Improvement with Reinterpolation* (EIR), já que ele é baseado no critério EI do caso determinístico.



Na verdade, o método propõe que, ao invés de se modificar o EI para casos estocásticos, utilizemos em conjunto o SK e o Kriging determinístico. A técnica de reinterpolação foi proposta por [Forrester, Keane e Bressloff \(2006\)](#) para casos onde o ruído homogêneo era considerado. Até o momento, não se teve uma análise de seu desempenho para o caso de ruído heterogêneo, o que é desenvolvido neste trabalho.

Basicamente, o método utiliza as matrizes de covariância, os parâmetros de ajuste e os pontos de suporte do modelo de regressão para construir por completo o metamodelo interpolador. Após a construção, as predições calculadas pelo SK nos pontos de suporte serão utilizadas para construção de um novo modelo em Kriging, agora determinístico, já que as predições virão livres do erro intrínseco. Como este último modelo é livre de ruído, o EI clássico poderá ser utilizado como uma métrica para obtenção de novos IPs.

Seja

$$\hat{\mathbf{Y}}_r = \left\{ \hat{Y}_r^{(1)}, \hat{Y}_r^{(2)}, \dots, \hat{Y}_r^{(n)} \right\}^T \quad (4.33)$$

o vetor formado pelas predições do SK (4.15) para os  $n$  pontos de suporte, onde o subíndice  $r$  identifica a palavra regressor. Temos que o preditor do Kriging determinístico baseado na distribuição do vetor aleatório (4.33) será dado por (3.31) e reescrito como

$$\hat{Y}(\mathbf{d}) = \hat{\mu} + \mathbf{h}^T \mathbf{\Psi}^{-1} (\hat{\mathbf{Y}}_r - \mathbf{1}\hat{\mu}), \quad (4.34)$$

onde

$$\hat{\mu} = \frac{\mathbf{1}^T \mathbf{\Psi}^{-1} \hat{\mathbf{Y}}_r}{\mathbf{1}^T \mathbf{\Psi}^{-1} \mathbf{1}}. \quad (4.35)$$

Temos que  $\mathbf{\Psi}$  e  $\mathbf{h}$  continuam os mesmos como definidos por (3.11) e (3.21), respectivamente, e os parâmetros de ajuste não precisarão ser otimizados novamente.

Para mostrarmos que esse novo preditor atua como um interpolador para os dados obtidos do SK, devemos mostrar que  $\hat{Y}_r^{(i)}$  e  $\hat{Y}(\mathbf{d}^{(i)})$  de (4.34) são idênticos. Antes, vamos mostrar que o parâmetro  $\hat{\mu}$  é igual a  $\hat{\mu}_r$ , onde  $\hat{\mu}_r$  é o parâmetro ajustado pelo SK e solução da equação

(B.6) apresentada no Apêndice B. Substituindo todos preditores do SK (4.15) para os pontos de suporte em (4.33) temos que o vetor de predições é dado por

$$\hat{\mathbf{Y}}_r = \mathbf{1}\hat{\mu}_r + \hat{\sigma}_z^2 \mathbf{\Psi} \hat{\mathbf{\Sigma}}^{-1} (\bar{\mathbf{y}} - \mathbf{1}\hat{\mu}_r). \quad (4.36)$$

Substituindo essa nova expressão em (4.35) temos que

$$\begin{aligned} \hat{\mu} &= \frac{\mathbf{1}^T \mathbf{\Psi}^{-1} (\mathbf{1}\hat{\mu}_r + \hat{\sigma}_z^2 \mathbf{\Psi} \hat{\mathbf{\Sigma}}^{-1} (\bar{\mathbf{y}} - \mathbf{1}\hat{\mu}_r))}{\mathbf{1}^T \mathbf{\Psi}^{-1} \mathbf{1}} \\ &= \frac{\mathbf{1}^T \mathbf{\Psi}^{-1} \mathbf{1}\hat{\mu}_r + \hat{\sigma}_z^2 \mathbf{1}^T \mathbf{\Psi}^{-1} \mathbf{\Psi} \hat{\mathbf{\Sigma}}^{-1} (\bar{\mathbf{y}} - \mathbf{1}\hat{\mu}_r)}{\mathbf{1}^T \mathbf{\Psi}^{-1} \mathbf{1}} \\ &= \hat{\mu}_r + \hat{\sigma}_z^2 \left[ \frac{\mathbf{1}^T \hat{\mathbf{\Sigma}}^{-1} \bar{\mathbf{y}} - \mathbf{1}^T \hat{\mathbf{\Sigma}}^{-1} \mathbf{1}\hat{\mu}_r}{\mathbf{1}^T \mathbf{\Psi}^{-1} \mathbf{1}} \right], \end{aligned}$$

e de (B.6) podemos ver que

$$\hat{\mu}_r = \frac{\mathbf{1}^T \mathbf{\Sigma}^{-1} \bar{\mathbf{y}}}{\mathbf{1}^T \mathbf{\Sigma}^{-1} \mathbf{1}} \quad \rightarrow \quad \mathbf{1}^T \mathbf{\Sigma}^{-1} \mathbf{1}\hat{\mu}_r = \mathbf{1}^T \mathbf{\Sigma}^{-1} \bar{\mathbf{y}},$$

e dessa forma, encontramos que

$$\begin{aligned} \hat{\mu} &= \hat{\mu}_r + \hat{\sigma}_z^2 \left[ \frac{\mathbf{1}^T \hat{\mathbf{\Sigma}}^{-1} \bar{\mathbf{y}} - \mathbf{1}^T \hat{\mathbf{\Sigma}}^{-1} \mathbf{1}\hat{\mu}_r}{\mathbf{1}^T \mathbf{\Psi}^{-1} \mathbf{1}} \right] \\ &= \hat{\mu}_r + \hat{\sigma}_z^2 \left[ \frac{\mathbf{1}^T \hat{\mathbf{\Sigma}}^{-1} \bar{\mathbf{y}} - \mathbf{1}^T \hat{\mathbf{\Sigma}}^{-1} \bar{\mathbf{y}}}{\mathbf{1}^T \mathbf{\Psi}^{-1} \mathbf{1}} \right] \\ &= \hat{\mu}_r + \hat{\sigma}_z^2 \cdot 0 \\ &= \hat{\mu}_r. \end{aligned} \quad (4.37)$$

Portanto, vemos que a média do modelo interpolador criado a partir das predições estocásticas, é idêntico à média do modelo regressor SK, não sendo necessário um novo cálculo para o parâmetro  $\hat{\mu}$ . Substituindo agora (4.36) e (4.37) em (4.34) temos

$$\begin{aligned} \hat{\mathbf{Y}}(\mathbf{d}) &= \hat{\mu}_r + \mathbf{h}^T \mathbf{\Psi}^{-1} \left[ \mathbf{1}\hat{\mu}_r + \hat{\sigma}_z^2 \mathbf{\Psi} \hat{\mathbf{\Sigma}}^{-1} (\bar{\mathbf{y}} - \mathbf{1}\hat{\mu}_r) - \mathbf{1}\hat{\mu}_r \right] \\ &= \hat{\mu}_r + \hat{\sigma}_z^2 \mathbf{h}^T \mathbf{\Psi}^{-1} \mathbf{\Psi} \hat{\mathbf{\Sigma}}^{-1} (\bar{\mathbf{y}} - \mathbf{1}\hat{\mu}_r) \end{aligned}$$

$$\begin{aligned}
&= \hat{\mu}_r + \hat{\sigma}_z^2 \mathbf{h}^T \hat{\Sigma}^{-1} (\bar{\mathbf{y}} - \mathbf{1}\hat{\mu}_r) \\
&= \hat{Y}_r(\mathbf{d}).
\end{aligned} \tag{4.38}$$

Assim, concluímos que o preditor obtido em (4.34) não só interpola os dados do SK, como na verdade é idêntico ao preditor do SK.

Assim sendo, como poderemos utilizar o próprio preditor do SK, a reinterpolação precisará somente do valor da nova variância espacial da correlação entre os pontos de suporte. Podemos ver que

$$\begin{aligned}
\hat{Y}_r - \mathbf{1}\hat{\mu}_r &= \mathbf{1}\hat{\mu}_r + \hat{\sigma}_z^2 \Psi \hat{\Sigma}^{-1} (\bar{\mathbf{y}} - \mathbf{1}\hat{\mu}_r) - \mathbf{1}\hat{\mu}_r \\
&= \hat{\sigma}_z^2 \Psi \hat{\Sigma}^{-1} (\bar{\mathbf{y}} - \mathbf{1}\hat{\mu}_r),
\end{aligned}$$

e substituindo (4.36) na estimação (3.18) da variância, temos que

$$\begin{aligned}
\hat{\sigma}_{\text{ri}}^2 &= \frac{(\hat{Y}_r - \mathbf{1}\hat{\mu}_r)^T \Psi^{-1} (\hat{Y}_r - \mathbf{1}\hat{\mu}_r)}{n} \\
&= \frac{\left[ \hat{\sigma}_z^2 \Psi \hat{\Sigma}^{-1} (\bar{\mathbf{y}} - \mathbf{1}\hat{\mu}_r) \right]^T \Psi^{-1} \left[ \hat{\sigma}_z^2 \Psi \hat{\Sigma}^{-1} (\bar{\mathbf{y}} - \mathbf{1}\hat{\mu}_r) \right]}{n} \\
&= (\hat{\sigma}_z^2)^2 \left[ \frac{(\bar{\mathbf{y}} - \mathbf{1}\hat{\mu}_r)^T \hat{\Sigma}^{-1} \Psi \Psi^{-1} \Psi \hat{\Sigma}^{-1} (\bar{\mathbf{y}} - \mathbf{1}\hat{\mu}_r)}{n} \right] \\
&= \frac{(\hat{\sigma}_z^2)^2}{n} \left[ (\bar{\mathbf{y}} - \mathbf{1}\hat{\mu}_r)^T \hat{\Sigma}^{-1} \Psi \hat{\Sigma}^{-1} (\bar{\mathbf{y}} - \mathbf{1}\hat{\mu}_r) \right].
\end{aligned} \tag{4.39}$$

Portanto, para obtermos o MSE para a reinterpolação, basta que  $\hat{\sigma}^2$  seja substituída por (4.39) em (3.33). Dessa forma, temos que o erro ocorrido na predição irá ser reduzido a zero nos pontos de suporte, e isso garante a otimização global via EI pela prova de [Schonlau \(1997\)](#).

Para concluir, de posse do preditor do SK (idêntico para a reinterpolação) e do MSE que ocorre na predição,  $s^2(\mathbf{d})$ , com o uso de (4.39), podemos definir uma variável aleatória que dependa do preditor e da variância obtida no MSE, e de forma análoga ao feito na Seção 3.4.1, construir o EI para termos a métrica do EIR. Um novo IP,  $\mathbf{d}^{n+1}$ , será determinado quando maximizarmos a métrica EIR.

#### 4.4 VARIÂNCIA ADAPTATIVA

No desenvolvimento do SK na Seção 4.2 foi possível estender o conceito do Kriging interpolador para o regressor, agregando o valor da variância obtida pelo MCI. Vimos também que um parâmetro de qualidade da aproximação é o número de replicações  $n_r$  do valor da função. Dessa forma, é possível controlar a matriz de covariância  $\Sigma_\varepsilon$  por meio dessa quantidade.

Podemos ver em [Jalali, Nieuwenhuys e Picheny \(2017\)](#) e [Carraro et al. \(2019\)](#) que um fator de extrema importância para o sucesso de uma boa aproximação e conseqüentemente uma boa otimização, é o correto ajuste de  $n_r$ . Um valor baixo de  $n_r$  irá acarretar em uma má estimação do valor da função e em um valor alto de variância do erro, isso irá gerar um modelo pobre para os dados. Por outro lado, ajustar um alto valor para  $n_r$  gerará uma boa aproximação do valor da função, porém poderemos reduzir a variância ao ponto do SK se parecer com um interpolador.

Uma alternativa mais elegante para a determinação da avaliação da função poderá ser ajustar uma variância alvo  $\bar{\sigma}_{\text{alvo}}^2$  e iteradamente avaliar a função objetivo até que se atinja uma variância menor ou igual à variância alvo. Essa abordagem também nos permite manter um controle sobre a grandeza da matriz de covariância, não deixando seus elementos serem grandes ou pequenos.

No entanto, como elucidado por [Carraro et al. \(2019\)](#), devemos tomar muito cuidado com o ajuste da variância alvo, levando em consideração duas características:

1. Se a variância alvo é ajustada muito alta, o erro associado poderá levar a uma aproximação pobre da integral (1.1);
2. Se a variância alvo tender a zero, o erro também tenderá e levará o modelo ao caso determinístico. Porém, isso irá gerar um alto custo computacional.

Outro detalhe visto por Carraro et al. (2019), é que o ajuste de um alto valor para  $\bar{\sigma}_{\text{alvo}}^2$  pode comprometer a otimização via sEGO. Utilizando o AEI, os autores mostraram que o algoritmo pode ficar preso em um determinado ponto de suporte ou em uma determinada região onde o valor MSE associado é muito alto. Esse mesmo comportamento será encontrado nas métricas MQ e EQI, pois estes também permitem que pontos de suporte sejam escolhidos como IPs. Para as métricas TSSO e EIR essa estagnação do algoritmo não é esperada, pois suas métricas não permitem a escolha de pontos de suporte como IPs, mas mesmo assim a otimização ficará prejudicada pela baixa qualidade do metamodelo.

Como o nosso foco está na otimização de (1.2), então deve ser utilizada uma forma inteligente de escolha da variância alvo. Essa escolha leva em consideração que, quando a variância alvo for alta, então o algoritmo pode se tornar extremamente local devido as altas incertezas. Se for utilizado um valor baixo, então forçaremos um alto número de replicações  $n_r$ , prejudicando o custo computacional de todo o processo. Levando em consideração essas características, temos o algoritmo que será chamado de sEGO Adaptativo.

A proposta de Carraro et al. (2019) consiste de uma seleção com base em uma variância adaptativa, que heurísticamente é capaz de adaptar a seleção da variância alvo para cada situação em que se encontra o algoritmo de otimização. A ideia principal é seguir o princípio de pesquisa global e local do EGO. Esse processo adaptativo começa tomando  $\bar{\sigma}_{\text{alvo}}^2 = \bar{\sigma}_0^2$ , sendo  $\bar{\sigma}_0^2$  uma variância inicial definida pelo usuário. Sugere-se que esse valor não seja pequeno, assim, durante a exploração do sEGO adaptativo, os IPs iniciais serão avaliados somente com algumas replicações no MCI, evitando um gasto computacional excessivo logo no começo do algoritmo. Na sequência, a depender do IP escolhido pelo sEGO adaptativo, o método gradativamente diminui o valor da variância alvo, fazendo ela assumir  $\bar{\sigma}_{\text{alvo}}^2 = \bar{\sigma}_{\text{adap}}^2$  para os IPs que deverão ser amostrados. Espera-se que com essa diminuição o

algoritmo dê certa ênfase à pesquisa local de regiões promissoras, mas não fique aprisionado nelas.

A proposta feita é que a diminuição do valor de  $\bar{\sigma}_{\text{alvo}}^2$  se dê tomando  $\bar{\sigma}_{\text{adap}}^2$  de forma exponencial. Para tanto, utilizamos uma função exponencial parametrizada pelo número de dimensões do problema  $k$  e pelo número de pontos de suporte que se encontram a uma curta distância do próximo IP, denotado por  $n_c$ .

Suponha que  $\mathbf{d}^{n+1}$  é o ponto que resolve algumas das métricas de adição de IPs da Seção 4.3, então teremos que

$$n_c = \sum_{i=1}^n U(\mathbf{d}^{n+1}, \mathbf{d}^{(i)}), \quad (4.40)$$

onde

$$U(\mathbf{d}^{n+1}, \mathbf{d}^{(i)}) = \begin{cases} 1, & \text{se } \|\mathbf{d}^{n+1} - \mathbf{d}^{(i)}\|_{\infty} \leq r_{hc} \\ 0, & \text{se } \|\mathbf{d}^{n+1} - \mathbf{d}^{(i)}\|_{\infty} > r_{hc} \end{cases} \quad (4.41)$$

sendo  $\|\cdot\|_{\infty}$  a representação para a norma infinita de um vetor. O parâmetro  $r_{hc}$  será parte da proposta adaptativa pois representará a distância considerada entre o IP e os pontos de suporte. A escolha da norma infinita como medida representa um hipercubo centralizado em  $\mathbf{d}^{n+1}$ , cuja distância entre o IP e a face do hipercubo é  $r_{hc}$ .

Então temos que, quando um IP é localizado em uma região sem pontos de suporte, a variância adaptativa será a variância inicial, ou seja,  $\bar{\sigma}_{\text{adap}}^2 = \bar{\sigma}_0^2$ . Logo, estes IPs serão pontos exploratórios e não precisarão de uma alta precisão na sua avaliação. Agora se o IP for localizado em uma região que contenha outros pontos de suporte, uma variância alvo menor é aplicada para o cálculo da função. Assim, novos aglomerados de pontos estarão se formando; logo, é inteligente o algoritmo possuir uma exploração mais local nessas regiões, resultando no refinamento da avaliação da função.

A expressão proposta por Carraro (2017) para o cálculo da

variância adaptativa é

$$\bar{\sigma}_{\text{adap}}^2 = \bar{\sigma}_0^2 \exp \left[ 0.01 \cdot k \cdot n_c - 0.5(1 + k + n_c) \right]. \quad (4.42)$$

A justificativa da escolha dessa função foram as seguintes:

- A queda exponencial do valor da variância é a mais natural, considerando que quanto maior o número de avaliações da função, menor será o erro na aproximação. Essa consideração também faz com que a variância alvo vá decrescendo suavemente, conforme os pontos forem se agrupando. Se, por exemplo, uma abordagem linear fosse utilizada, conforme o valor de  $n_c$  crescesse, a variância adaptativa poderia decrescer abruptamente, fazendo com que o algoritmo ficasse paralisado em um certo aglomerado de pontos;
- A taxa de queda da variância é proporcional à dimensão do problema. Para problemas de baixa dimensão, o espaço de busca é relativamente menor e conseqüentemente os IPs terão uma maior facilidade para se agrupar. Dessa forma a diminuição da variância não pode acontecer rapidamente gastando todo o orçamento computacional. Para dimensões maiores, os agrupamentos não acontecerão tão rápido. Dessa forma a taxa de queda deverá ser um pouco maior para garantir boas avaliações em pontos mais excluídos.

Também é necessário que a variância adaptativa não assuma valores extremamente baixos, fazendo a aproximação via MCI se tornar impraticável. Portanto, é ideal que

$$\bar{\sigma}_{\text{min}}^2 \leq \bar{\sigma}_{\text{adap}}^2 \leq \bar{\sigma}_0^2, \quad (4.43)$$

onde  $\bar{\sigma}_{\text{min}}^2$  é um limite inferior para a variância adaptativa.

#### 4.4.1 Ajuste da variância adaptativa às métricas do sEGO

A abordagem vista anteriormente, de se ajustar uma variância alvo, nos dá uma ideia de qual será o valor da variância do erro em

todos os pontos do domínio. Se estivermos trabalhando com pontos de suporte, então a variância do erro será o cálculo realizado em (4.2). Agora para pontos que não são de suporte, sabemos que se aquele ponto for ser amostrado e consecutivamente avaliado, então ele será avaliado até alcançar a variância adaptativa ou a variância alvo inicial.

Como consequência, vemos que a variância do erro pertencerá a uma função limitada, cujo supremo será dado pela variância alvo e ínfimo será dado pela variância mínima. Logo, para qualquer ponto  $\mathbf{d}$  no domínio  $\mathcal{D}$ , podemos afirmar que

$$\bar{\sigma}_{\min}^2 \leq \bar{\sigma}^2(\mathbf{d}) \leq \bar{\sigma}_0^2. \quad (4.44)$$

Dessa forma, temos que as expressões (4.21), (4.27) e (4.28) receberão valores reais e positivos para a variância do erro e poderão ser estimadas em pontos que não são de suporte.

Precisamos definir uma função para  $\bar{\sigma}^2(\mathbf{d})$  que carregue toda a noção de adaptatividade proposta anteriormente. Podemos definir que

$$\bar{\sigma}^2(\mathbf{d}) = \begin{cases} \bar{\sigma}_0^2, & \text{se } n_c = 0 \\ \bar{\sigma}_{\text{adap}}^2, & \text{se } n_c > 0 \end{cases}. \quad (4.45)$$

Vemos que definida dessa forma,  $\bar{\sigma}^2(\mathbf{d})$  satisfaz a relação (4.44) já que  $\bar{\sigma}_{\min}^2 \leq \bar{\sigma}_{\text{adap}}^2$ .

Podemos justificar a função (4.45) observando os seguintes fatos:

- Suponha que  $n_c = 0$  para qualquer  $\mathbf{d}$ , amostral ou não, que fosse ser adicionado como próximo ponto de suporte. Por não estar próximo de nenhum outro ponto, então não seria necessário que o algoritmo refinasse o valor da sua nova avaliação. Logo, vemos que a variância do erro poderá ser, no máximo, a variância alvo inicial. Portanto, podemos deixar que  $\bar{\sigma}^2(\mathbf{d})$  seja igual a  $\bar{\sigma}_0^2$ ;
- Supondo agora que  $n_c > 0$  para qualquer  $\mathbf{d}$ , amostral ou não, então a aproximação do valor da função nesse possível próximo



ponto de suporte deverá ser mais refinado pelo algoritmo. Assim, o ponto será avaliado até que a variância do erro seja menor ou igual ao valor da variância alvo. Portanto, podemos fazer que  $\bar{\sigma}^2(\mathbf{d})$  seja igual a  $\bar{\sigma}_{\text{adap}}^2$ .

Caso um ponto de suporte seja escolhido como próximo IP, é comum não adicionarmos esse ponto no espaço amostral novamente. O que se faz é adicionar novas replicações para o ponto, atualizar o valor da aproximação para a função e consecutivamente atualizar o modelo SK para o mesmo espaço amostral. Porém, como para a variância adaptativa interessa quantos pontos estão próximos do próximo IP, armazenamos uma contagem de quantas vezes um mesmo ponto de suporte será selecionado como IP. Dessa forma, ajustamos  $n_c$  como sendo esse valor durante a pesquisa adaptativa para estes pontos. Isso faz com que, a cada momento onde um mesmo ponto for escolhido como IP, a variância alvo diminua e faça o valor da função se tornar mais refinado nesse ponto.

Para exemplificar a abordagem, voltemos à Figura 19, considerando a mesma função sob as mesmas condições que geraram o referido gráfico. Para o cálculo dos valores amostrais que geraram a aproximação apresentada, foi ajustada uma variância alvo inicial de  $\bar{\sigma}_0^2 = 0.1$ . Suponha que a distância máxima permitida entre IPs e pontos de suporte seja  $r_{hc} = 0.1$ . Vamos considerar que, utilizando a métrica AEI, precisamos da variância do erro no ponto  $d^{n+1} = 1.35$ . Sabendo que existe um ponto de suporte com o valor de  $d^{(i)} = 1.3$ , temos que  $n_c = 1$ . Logo teremos que  $\bar{\sigma}^2(d^{n+1}) = \bar{\sigma}_{\text{adap}}^2 = 0.02254$ . Observe que não é preciso calcular a aproximação via MCI para se determinar o valor da variância do erro. Vamos considerar agora  $d^{n+1} = 3.6$ , também no AEI. Os pontos de suporte perto desse valor são  $d^{(i)} = 3.4$  e  $d^{(j)} = 3.8$ . Portanto, temos que  $n_c = 0$  e consecutivamente  $\bar{\sigma}^2(d^{n+1}) = \bar{\sigma}_0^2 = 0.1$ .

## 4.5 NORMALIZAÇÃO APLICADA AO SK

Até o momento somos capazes de ajustar com qualidade um modelo regressor e com base nas informações do preditor e do MSE, realizar uma otimização global para problemas estocásticos. Vimos que uma das características mais importantes para uma boa aproximação e para o sucesso da otimização, é o correto ajuste da variância alvo no processo de aproximação da função.

Entretanto, um problema natural que surge nos processos estocásticos é a amplitude de resposta da função objetivo. Quando uma função possuir uma extensão (*range*) de contradomínio excessivamente grande, uma variância alvo pequena poderá se tornar inalcançável. Dessa forma, poderíamos ter problemas ao ajustar a primeira aproximação em SK para os dados. Caso a extensão do valor da função seja pequeno, poucas avaliações são necessárias para se obter a variância alvo, logo nesse cenário não espera-se muito trabalho. Iremos aqui fazer uma convenção, escrevendo CD para representar a extensão do contradomínio de uma função, ou seja, CD será a diferença entre o valor máximo e o valor mínimo que uma função poderá assumir.

Para uma ilustração didática de tal dificuldade, definimos uma nova função estocástica (1.1). Utilizamos a função Branin Modificada (FORRESTER; SOBESTER; KEANE, 2008) em um contexto estocástico, definida como  $\psi : \mathcal{D} \times \Omega \rightarrow \mathbb{R}$  e pela lei

$$\begin{aligned} \psi(\mathbf{d}, \mathbf{X}) = & \left( d_2 - \frac{5.1}{4\pi^2} d_1^2 + \frac{5}{\pi} d_1 - 6 \right)^2 \cdot X_1 + 10 \left( 1 - \frac{1}{8\pi} \right) \cos(d_1) \cdot X_2 \\ & + 10 + 5D_1, \end{aligned} \tag{4.46}$$

onde  $\mathbf{d} \in \mathcal{D} = [-5, 10] \times [0, 15]$ ,  $\mathbf{X} \in \Omega$ . Vamos definir que  $\Omega$  será o espaço formado por todos os pares de combinações possíveis entre os valores das variáveis aleatórias  $X_1 \sim \mathcal{N}(1, 0.05)$  e  $X_2 \sim \mathcal{N}(1, 0.05)$ . A saber, o minimizador determinístico dessa função é  $\mathbf{d}^* = \{-3.68929, 13.62999\}^T$  e o contradomínio é  $[-16.6440, 283.1291]$ , apresentando  $CD_\psi = 299.7731$ .

Vamos supor que queiramos avaliar  $\psi(\mathbf{d}, \mathbf{X})$  para  $\mathbf{d} = \{-3, 14\}^T$  a uma variância inicial  $\bar{\sigma}_0^2 = 0.01$ . Realizando uma simulação estocástica, obtemos que com  $n_r = 12$  replicações, a variância da aproximação via MCI é dada por  $\bar{\sigma}^2 = 0.00956$ . O valor aproximado foi  $\bar{J}(\mathbf{d}) = -10.5458$  e o valor original é  $J(\mathbf{d}) = -10.2513$ . Vemos que nesse caso o número de replicações é completamente tratável por qualquer algoritmo.

Como o sEGO adaptativo é um algoritmo de exploração global, nem sempre iremos avaliar a função objetivo em pontos próximos do minimizador (parágrafo anterior). A valer, até que encontremos a bacia com o possível mínimo, o algoritmo tenderá a explorar locais com alto valor do MSE. Suponha que em uma dessas pesquisas, um IP foi selecionado como  $\mathbf{d} = \{-5, 0\}^T$  e que nenhum ponto de suporte esteja perto desse ponto. Então precisaremos avaliar a função  $\psi(\mathbf{d}, \mathbf{X})$  até que atinjamos a variância alvo de  $\bar{\sigma}_0^2 = 0.01$ . Nesse cenário seriam necessárias  $n_r = 21\,769$  replicações até que consigamos  $\bar{\sigma}^2 = 0.0099998$ . A avaliação da função é  $\bar{J}(\mathbf{d}) = 283.1040$  enquanto o valor original é  $J(\mathbf{d}) = 283.1291$ . Podemos ver que o número de replicações aumentou consideravelmente. Nesse caso, o algoritmo já dependeria de uma grande reserva de avaliações da função.

Indo um pouco além, suponha que o mesmo IP,  $\mathbf{d} = \{-5, 0\}^T$ , fosse selecionado, porém agora ele será pertencente a uma região que já possui outros três pontos agrupados. Nessa situação, o algoritmo está refinando uma determinada região do espaço de busca. Como  $n_c = 3$ , então a variância adaptativa entra em vigor, e deveremos avaliar a função até atingirmos  $\bar{\sigma}_{\text{adap}}^2 = 0.0005287$ . Para tal situação, o algoritmo gasta  $n_r = 413\,945$  replicações alcançando a variância de  $\bar{\sigma}^2 = 0.0005286995$ , com a função valendo  $\bar{J}(\mathbf{d}) = 283.1466$ . Observe a imensa diferença entre o número de replicações para dois cenários muito comuns ao utilizarmos o sEGO adaptativo.

Nesse último caso, teríamos um cenário digno de ser chamado de intratável, já que o tempo computacional para se avaliar  $n_r$  seria exorbitante (claro, a depender de cada problema!). Outro fator impor-

tante é o orçamento computacional disponível para se avaliar a função nos IPs. Por exemplo, no trabalho de [Jalali, Nieuwenhuys e Picheny \(2017\)](#) os autores fixam uma quantidade máxima de 2750 avaliações para a busca de IPs. Com somente essa quantidade, na última situação descrita acima, conseguiríamos aproximar a função até  $\bar{\sigma}^2 = 0.0793$  e com um valor aproximado de  $\bar{J}(\mathbf{d}) = 283.1414$ . Podemos ver que o valor da função até se aproxima do valor original, no entanto, a forma lenta de diminuição da variância está associada ao alto CD dos valores da função. Dentro das 2750 avaliações, obtivemos as simulações 236.0041 e 334.3428 como valores mínimo e máximo, respectivamente, para a função.

#### 4.5.1 A normalização via tunelamento estocástico

O que conseguimos inferir das situações delineadas anteriormente, é que o algoritmo tem que ser capaz de trabalhar de forma inteligente as avaliações das funções. É preciso um tratamento eficaz para os diversos valores assumidos pela função, para que os mesmos não consumam de uma única vez o orçamento computacional.

Para realizar tal tratamento, realizamos neste trabalho uma normalização do valor da função, para que suas avaliações possuam um CD menor que o original. Conseqüentemente, a variabilidade dos valores obtidos pelo MCI é menor e com isso conseguimos reduzir a variância do erro na aproximação. Essa normalização será uma transformação não linear que irá receber o valor original da função e retornar um valor menor para a mesma avaliação. Porém, garantindo as mesmas nuances do ruído original.

A métrica de normalização escolhida para esse trabalho é uma adaptação do tunelamento estocástico realizado por [Wenzel e Hamacher \(1999\)](#). Essa métrica foi desenvolvida pelos autores para a correção de um problema que surge dentro do algoritmo Recozimento Simulado (*Simulated Annealing* - SA). O intuito é corrigir o chamado Problema de Congelamento (*Freezing Problem*) causado quando dois mínimos locais

próximos possuem uma distância menor do que o valor da função nas regiões que os separam. A transformação proposta é

$$F(\mathbf{d}) = 1 - \exp[-\gamma(J(\mathbf{d}) - J_0)], \quad (4.47)$$

onde  $\gamma$  é um real positivo chamado Constante de Tunelamento,  $J(\mathbf{d})$  é o valor original da função objetivo e  $J_0$  é o valor mínimo atual encontrado pelo SA.

Essa transformação preserva as localizações de todos os mínimos realizando um mapeamento de toda a função objetivo para o intervalo  $[0, 1]$ , partindo do valor de  $J_0$  até o máximo da função. O grau de declividade entre os vales da função objetivo é controlado pela constante de tunelamento  $\gamma$ . Conforme o ajuste contínuo da função objetivo é realizada em torno de  $J_0$ , as características irrelevantes da função objetivo vão sendo eliminadas, fazendo com que o algoritmo do SA não fique preso em mínimos locais.

Para ilustração da transformação, consideremos a função determinística proposta em (3.34). A Figura 20(a) apresenta o gráfico e o contradomínio original da função  $[-2.3199, 2.1271]$  com  $CD_\psi = 4.4470$ .

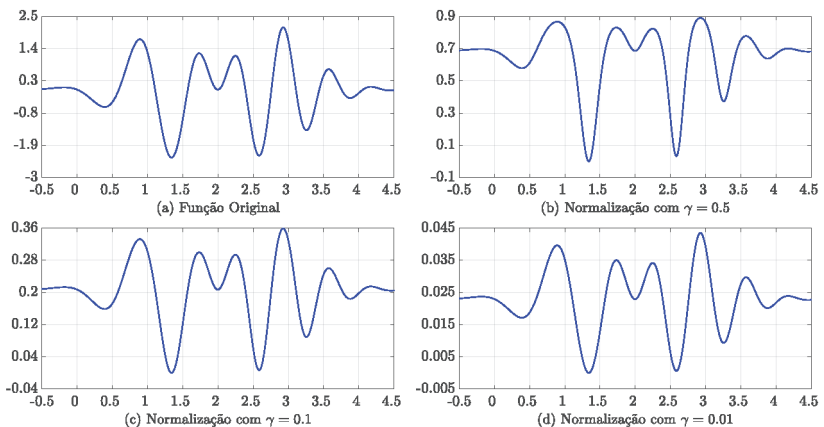


Figura 20 Aplicação da normalização.

Aplicando a transformação (4.47) com  $J_0 = -2.3199$  e para os valores de  $\gamma = 0.5, 0.1, 0.01$ , obtemos respectivamente os gráficos 20(b), 20(c) e 20(d). Podemos observar que os CDs da função diminuem expressivamente conforme diminuímos  $\gamma$ , onde para  $\gamma = 0.01$  obtemos  $CD_{\psi}^{\text{norm}} = 0.0435$ . Observamos também que, quando a constante de tunelamento assume os valores 0.1 e 0.01, é mantida inalterada a forma da função original. Para  $\gamma = 0.5$  nas regiões dos máximos observamos uma leve alteração.

Uma situação oposta acontece conforme aumentamos o valor de  $\gamma$ . As Figuras 21(b), 21(c) e 21(d) apresentam, respectivamente, os gráficos para os valores  $\gamma = 1, 5, 10$ . Podemos ver como a normalização modifica a função. Primeiro o CD da função fica limitado ao intervalo  $[0, 1]$  pois podemos observar que: quando avaliada no minimizador,  $J(\mathbf{d}) = J_0$ , e conseqüentemente  $F(\mathbf{d}) = 0$  para qualquer  $\gamma$ ; agora para os demais pontos,  $J(\mathbf{d}) > J_0$ , e conforme  $\gamma \rightarrow \infty$  temos que  $F(\mathbf{d}) \rightarrow 1$ . Logo, vemos que a normalização é capaz de ir removendo as regiões onde possivelmente o mínimo global não estará. Essa é a principal característica que evita o problema de congelamento do SA, já que diminui-se as regiões com vales e facilmente o algoritmo é capaz de realizar a transição de um vale para o outro.

O nosso objetivo não é evitar os vários vales que a função objetivo pode possuir, já que a premissa do sEGO adaptativo é que ele consegue varrer as regiões com altas incertezas da função e encontrar a bacia do mínimo global. Estamos mais interessados no que acontece quando aplicamos a transformação conforme Figura 20, já que poderemos diminuir o CD de aplicação da função.

Para podermos visualizar como se comporta essa transformação não linear como uma normalização para a função estocástica, vamos voltar à função (4.3) dada por (4.4), porém vamos aumentar a variabilidade do parâmetro estocástico, fazendo  $X \sim \mathcal{N}(1, 2)$ . A Figura 22(a) apresenta a função original e suas aproximações sem a normalização para  $n_r = 5, 10$  e 50. Podemos ver como uma variabilidade maior do

parâmetro estocástico prejudica a aproximação da função, onde para  $n_r = 5$  perdemos a suavidade em algumas regiões. Com  $n_r = 50$  a suavidade se estabiliza.

A Figura 22(b) apresenta a função normalizada, tanto para o caso sem ruído, como para os casos com ruído estocástico. Observamos que a normalização é bem sucedida, uma vez que consegue reduzir o CD da função e manter os ruídos aplicados sem alterar a forma final da função. Podemos observar que alguns valores na função normalizada foram menores do que zero. Esse comportamento não acontecerá caso a função seja determinística, já que o valor mínimo será sempre igual a  $J_0$ . Porém, para casos estocásticos essa afirmação não é válida, uma vez que podemos obter  $\bar{J}(\mathbf{d})$  menor do que  $J_0$  para algum ponto do domínio (não necessariamente o minimizador).

O fato de surgirem valores negativos apontam uma fraqueza da normalização para os casos estocásticos. Para circundar essa negatividade, o ajuste dos parâmetros  $\gamma$  e  $J_0$  deverão ser realizados com cautela. Por exemplo, suponha que  $\gamma = 1$  e  $J_0 = -2.3199$  para os casos analisados anteriormente. Para o parâmetro estocástico, sabemos que

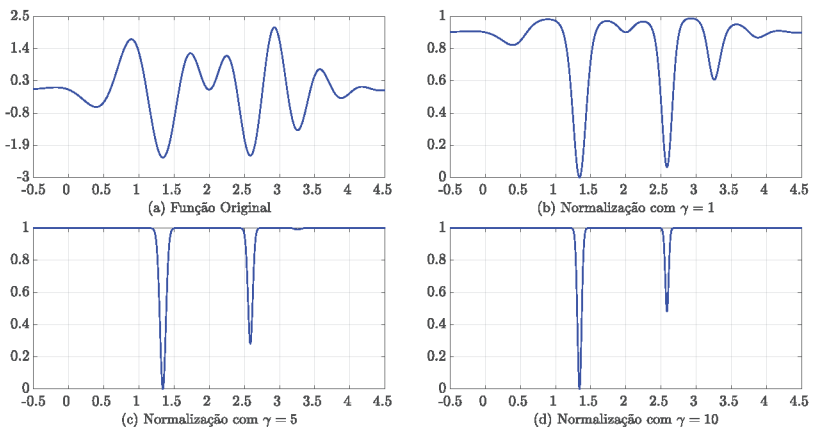
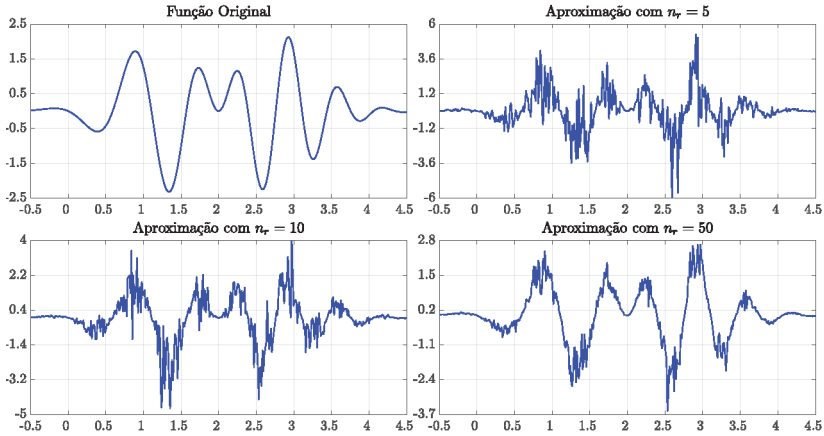


Figura 21 Modificação da função original via normalização.



(a) Função Original.

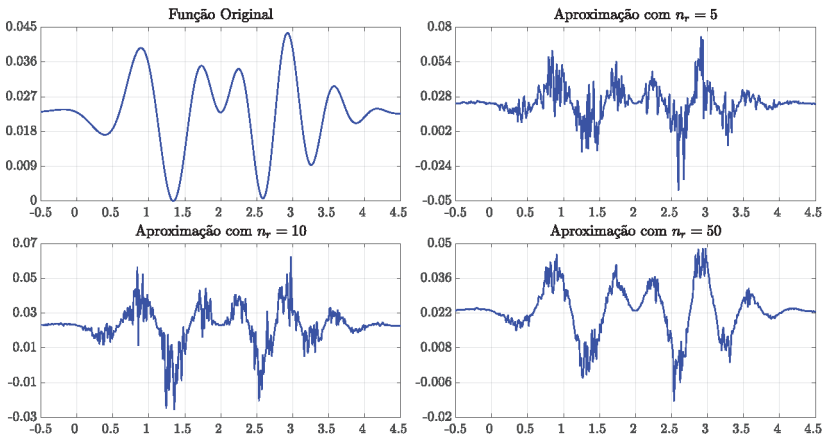
(b) Função Normalizada com  $\gamma = 0.01$ .

Figura 22 Função estocástica e sua normalização.

$P(X = 5) \approx 3\%$ . Essa probabilidade pode parecer baixa, porém ela é capaz de alterar nossa pesquisa de forma expressiva. Assim, corremos o risco de nos depararmos com uma simulação no minimizador onde  $J(d) = -11.5995$  (observe a definição da função em (4.3)) e consequentemente  $F(d) = -10716.1442$ . Situação muito diferente é obtida



quando  $\gamma = 0.01$  e consecutivamente  $F(d) = -0.0972378$ . Mesmo negativo, o valor da função com  $\gamma = 0.01$  é a que nós procuramos com a normalização.

Outra alternativa que temos para evitar tal problema é ajustar o valor de  $J_0$  incluindo a possível variabilidade do valor da função. Por exemplo, se tomássemos  $J_0 = -2.3199 \cdot 7 = -16.2393$  teríamos que na situação anterior  $F(d) = 0.9903$  para  $\gamma = 1$  e  $F(d) = 0.0453$  para  $\gamma = 0.01$ . A escolha de multiplicar por 7 o valor mínimo é duas: primeiro que o ruído estocástico é multiplicativo, ou seja, o parâmetro estocástico multiplica o valor da função determinística; e em segundo é que a probabilidade de  $P(X > 7) \approx 0.135\%$ , logo temos uma probabilidade menor que 1% de obtermos uma simulação inferior a 7 vezes o valor mínimo. Essas são duas características que podem guiar a escolha de  $J_0$ .

Outro detalhe que fica inferido das situações anteriores é que o valor de  $J_0$  não será dinâmico igual ao aplicado no SA. Para a aplicação da normalização ao sEGO adaptativo utilizaremos o  $J_0$  estático e igual ao mínimo determinístico da função. Quando esse mínimo for desconhecido, deveremos fazer uma análise prévia do contradomínio da função para que possamos pelo menos prever a grandeza do valor mínimo. Determinada a grandeza, julgou-se eficaz setar  $J_0$  uma grandeza a menos do valor mínimo. Para exemplificar, suponha que desconhecemos o mínimo determinístico, porém é esperado que o mínimo seja algo da grandeza de  $10^2$ . Então, sugere-se escolher  $J_0$  com a grandeza  $10^1$ .

#### 4.5.2 Influência da normalização no modelo SK

Uma consequência direta da normalização é a queda da variância obtida na aproximação do valor da função. Por exemplo, voltando ao último caso analisado no exemplo da função Branin Modificada, suponha que esta função tenha sido normalizada com  $\gamma = 0.01$  e  $J_0 = -16.644022$ . Temos que, no caso onde  $\mathbf{d} = \{-5, 0\}^T$  com  $n_c = 3$ , em apenas  $n_r = 2$  simulações obtemos a variância igual a  $\bar{\sigma}(\mathbf{d}) = 9.4991 \times 10^{-8}$ , que

é menor do que a variância adaptativa proposta. Com somente duas replicações do valor da função, obtemos a média igual a  $\bar{J}(\mathbf{d}) = 272.7396$ , que é levemente diferente da média original determinística. Entretanto, tal diferença não prejudica o processo de aproximação via o SK, já que este recolhe as tendências dos pontos com ruído, ou de otimização, pois caso o sEGO adaptativo revisite essa região, novas replicações podem ser adicionadas ou novos pontos nesta região serão adicionados. No primeiro caso, teremos que novas replicações farão o valor médio da função se aproximar do valor real. Já no segundo caso, teremos que  $n_c > 0$ , então a variância adaptativa entrará em ação aumentando o número de replicações da função para pontos nestas regiões.

Essa influência da normalização na covariância dos modelos SK, faz com que rapidamente o modelo em SK tenda a um interpolador. Porém, mostramos no restante do texto que essa influência não traz um comportamento negativo para o sEGO adaptativo. A diminuição da variância com a normalização permite que utilizemos mais do recurso computacional para podermos adicionar mais IPs. Com a adição maior de IPs, é esperado que o processo de otimização seja mais confiável, pois ou estaremos fazendo uma busca exploratória do domínio, ou estaremos fazendo o refinamento do melhor ponto obtido.

Para ilustração, vamos retornar à função (4.3), porém com  $X \sim \mathcal{N}(1, 0.5)$  e normalizada com  $\gamma = 0.01$  e  $J_0 = -2.3199$ . Suponha que tenhamos  $n = 30$  pontos de suporte avaliados com  $n_r = 2$ . Logo, um modelo inicial SK com a normalização sobre a função é apresentado pela Figura 23. Na simulação que foi utilizada para a geração dos gráficos apresentados, o contradomínio da função original foi de  $[-3.7323, 3.5584]$  apresentando  $CD_\psi = 7.2907$ . Utilizando a normalização o contradomínio foi reduzido à  $[-0.0142, 0.0571]$  com  $CD_\psi^{\text{norm}} = 0.0713$ . Podemos observar que o valor negativo foi gerado por uma aproximação via MCI que foi menor que  $J_0$ , porém balanceamos o valor alto do ruído ajustando um valor baixo para  $\gamma$ .

Podemos visualizar a influência direta da normalização no MSE

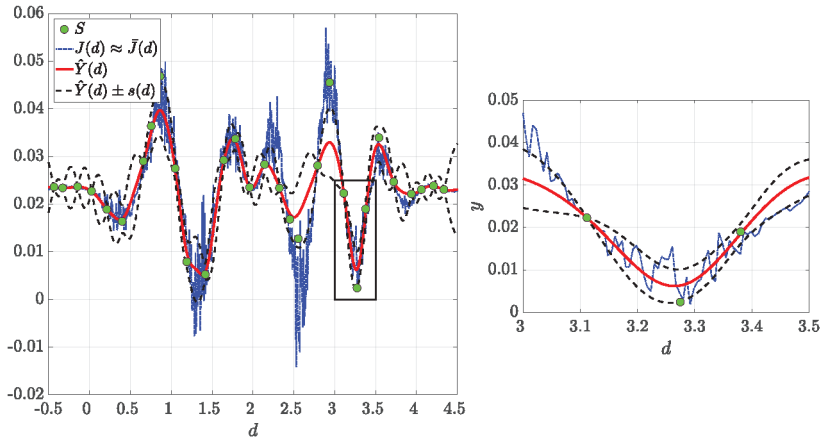


Figura 23 Aproximação via SK para função 1D normalizada.

da aproximação pelo gráfico à direita da Figura 23. Nela, ampliamos a região apresentada pelo retângulo que aparece na figura à esquerda para que possamos ver melhor a influência da normalização sobre a matriz de covariância do SK. Vemos que, mesmo a variância do erro da aproximação via MCI sendo pequena, isso não altera o comportamento de regressor do SK, mantendo as suas principais características.

#### 4.6 CARACTERÍSTICAS DAS MÉTRICAS DE ADIÇÃO DE IPs

Nesta seção ilustramos as principais características e comportamento das métricas de adição de IPs, por meio da função estocástica em uma dimensão que vem sendo apresentada. Utilizaremos o SEGO adaptativo para cada uma das métricas, levando em consideração a abordagem adaptativa, bem como a normalização. Nosso intuito é trazer uma clareza visual e algorítmica de como cada métrica atua. Na literatura atual (FORRESTER; KEANE; BRESSLOFF, 2006; HUANG et al., 2006; PICHENY; WAGNER; GINSBOURGER, 2012; PICHENY et al., 2013; QUAN et al., 2013; JALALI; NIEUWENHUYSE; PICHENY, 2017) não existe um estudo do comportamento geométrico de cada métrica; desta

forma, trazemos essa análise inédita, focando em como determinada métrica se põe perante a exploração local versus a exploração global.

Utilizamos a função (4.3) mantendo  $X \sim \mathcal{N}(1, 0.5)$  e normalizada com  $\gamma = 0.01$  e  $J_0 = -2.3199$ . Para a amostra inicial utilizamos  $n = 10$  pontos de suporte, gerados pela função `lhsdesign` do software `MatLab®` (MATHWORKS, 2017). Cada ponto de suporte inicial foi replicado  $n_r = 2$  vezes para a primeira aproximação do SK. É imposto o número máximo de 100 avaliações da função, incluindo as replicações utilizadas no espaço amostral, bem como as avaliações necessárias para cada IP adicionado. Como variância alvo inicial utilizamos  $\bar{\sigma}_0^2 = 0.01$ .

#### 4.6.1 Análise do MQ

A primeira métrica analisada é a MQ, cujo resultado após a execução completa do sEGO adaptativo pode ser conferida na Figura 24. Foram adicionados 18 IPs ao término das 100 avaliações, que foram representados por uma estrela de seis pontas. Como podemos ver, a métrica MQ se comportou de maneira puramente local. Como essa técnica minimiza a predição do SK penalizada pelo valor do RMSE,

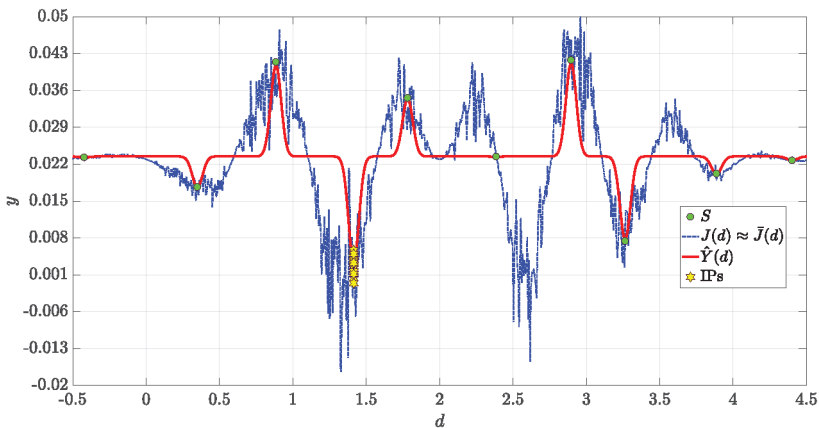


Figura 24 sEGO utilizando a métrica MQ.

então a menor predição associada ao menor RMSE normalmente será escolhida como IP.

Nesse caso em específico, a região que foi refinada pelo método foi escolhida por influência de um ponto de suporte que possuímos nessa região. Esse ponto de suporte possui um baixo valor de RMSE e é o menor dos valores preditos. Logo, como o método MQ permite replicações no mesmo ponto, esse ponto de suporte é então selecionado em todas as próximas adições de IPs. Deixamos marcados com a estrela todos os IPs simplesmente por motivos didáticos. Lembre-se que, na prática, os pontos de suporte que são selecionados como IPs não são repetidos no espaço amostral.

Outro fator complicador para o sEGO adaptativo ter sido convertido a uma busca local nesse caso, foi a pobre aproximação realizada pelo SK. Podemos ver que o SK quase atuou como interpolador, não tendo captado as tendências da função e simplesmente passando próximo aos valores da função. Esse comportamento é derivado do baixo valor da correlação entre os pontos de suporte. Como temos pontos distantes entre si varrendo todo o espaço de busca, não conseguimos encontrar uma boa correlação para os pontos; logo, nossa matriz de correlação espacial é próxima da matriz identidade. A título de curiosidade, a matriz resultante da subtração entre a matriz de correlação para os 10 pontos de suporte e a matriz identidade de ordem 10, é tal que sua norma Euclidiana é  $8.4377 \times 10^{-15}$ .

Podemos concluir que a técnica MQ é extremamente sensível ao espaço amostral inicial. Um espaço amostral inicial ruim, com pontos muito distantes ou com variâncias do erro muito baixas, farão com que o método seja puramente local. Portanto, é essencial uma correta escolha do espaço inicial de modo a obtermos pontos bem correlacionados e que façam uma boa dispersão dos dados, para que tenhamos uma maior probabilidade de encontrar a região com a bacia do mínimo global. No entanto, por possuir um poder alto de refinamento, a técnica MQ pode se sobressair perante as demais quando consideramos como comparação

o valor mínimo da função.

#### 4.6.2 Análise do AEI

Para a métrica AEI temos o resultado apresentado na Figura 25. Foram adicionados 17 IPs ao término da execução do sEGO adaptativo. Diferentemente do MQ, o AEI não foi puramente local, podemos observar que dos 17 IPs, três foram colocados em regiões exploratórias. Porém 14 IPs foram colocados na região do mínimo, realizando uma excelente modelagem da função naquela região.

Tal comportamento é explicado pela própria lei do AEI. Como vimos, sua definição permite a replicação de pontos de suporte; logo, pontos que ainda possam ser melhorados serão adicionados até que a parte de penalização do AEI entre em vigor e faça tal ponto de suporte ser desconsiderado. Isso se justifica pela adição dos 14 IPs na região do mínimo global. Como aqui nem todos os 14 IPs foram pontos de suporte, a variância adaptativa toma lugar e mais avaliações vão ser gastas para se avaliar a função. Isso justifica a boa qualidade da aproximação do SK na região do mínimo, bem como explica porquê o AEI possui um IP a menos que o MQ.

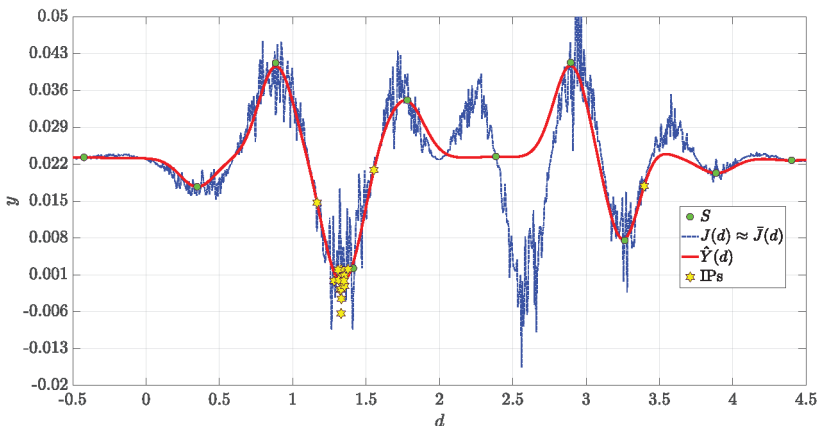


Figura 25 sEGO utilizando a métrica AEI.

Vemos também que, diferentemente do MQ, o modelo SK final já possui algumas das tendências da função original. Isso é consequência de conseguirmos ao longo do sEGO adaptativo, montar um espaço amostral com maior correlação entre os pontos, já que na região de mínimo vários pontos diferentes fazem parte do espaço amostral. Isso corrobora a qualidade do AEI em balancear pesquisa global com pesquisa local, mesmo a pesquisa local ainda sendo expressiva.

### 4.6.3 Análise do EQI

A Figura 26 ilustra o sEGO adaptativo utilizando a métrica EQI. Como também é de se esperar, o EQI possui um comportamento muito parecido com o MQ. Isso é decorrente de fazermos a suposição de que o modelo SK é mais confiável que as próprias aproximações da função, e de tomarmos como base um percentil da predição futura do modelo, levando em consideração o ponto em análise. Essas abordagens é que fazem o EQI se tornar uma versão mais sofisticada do MQ.

Foram adicionados com o EQI 17 IPs, o igualando ao AEI. O diferencial do EQI para o MQ é que, ao levarmos em consideração que

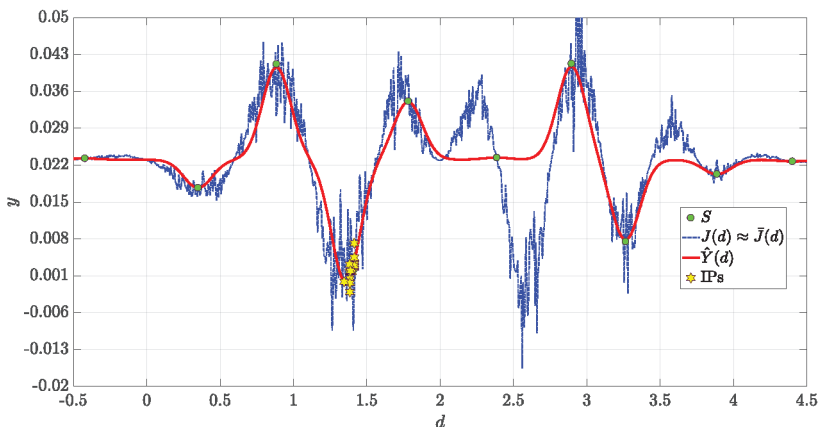


Figura 26 sEGO utilizando a métrica EQI.

um modelo futuro SK é melhor do que a própria avaliação da função, tiramos de cena a grande incerteza que pode ocorrer na avaliação da função. Essa incerteza é substituída pela variância adaptativa durante o processo adaptativo. Isso faz com que o percentil da predição não seja tão penalizado pelo RMSE, e consecutivamente faz com que o EQI seja capaz de visualizar pontos que realmente possuam uma melhora significativa para seu valor.

Como dito anteriormente, o EQI também permite replicações para pontos de suporte, logo teremos a oportunidade de refinarmos o valor de uma região. Essa configuração é vista na Figura 26 onde foi realizada somente uma busca local. Porém, diferentemente do MQ, podemos ver que, por não termos replicado somente pontos de suporte, conseguimos uma maior correlação espacial entre os pontos, fazendo com que a tendência da função original na região do mínimo fosse alcançada pelo SK. Esse comportamento consecutivamente melhora o valor mínimo já obtido pelo modelo.

Outra característica que cabe destaque é o cuidado ao definir o espaço amostral inicial. Se trabalharmos com espaços que prezem por pontos extremamente afastados, a baixa correlação fará com que o EQI fique facilmente preso em um ponto de suporte, até a variância do erro ser baixa o suficiente para a métrica escapar daquela região e começar a explorar o domínio. Nesse contexto, muitas avaliações da função seriam gastas somente nessa busca local inicial.

#### 4.6.4 Análise do TSSO

A métrica TSSO foi aplicada no sEGO adaptativo e a visualização do resultado pode ser conferido na Figura 27. Com essa métrica foram adicionados 39 IPs. Podemos verificar que o TSSO possui o comportamento previsto. Nenhum ponto de suporte é selecionado como IP já que o MEI leva em consideração o RMSE do Kriging determinístico.

Vemos uma excelente combinação entre exploração local e exploração global, onde as bacias de mínimos foram encontradas e bem



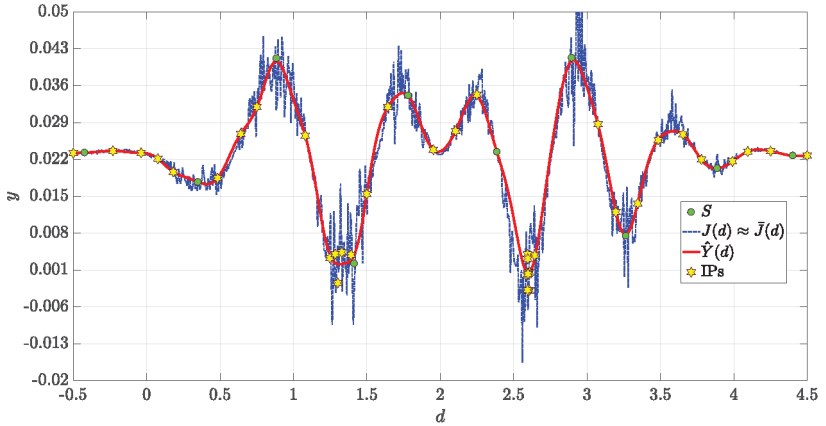


Figura 27 sEGO utilizando a métrica TSSO.

refinadas pelo modelo. Podemos ver que a capacidade de exploração global é capaz de aumentar a correlação entre os pontos de suporte, fazendo com que o modelo SK consiga captar todas as nuances da função original. Se atentarmos para o fato de que o espaço amostral é idêntico em todas as métricas vistas até agora, podemos ver a potencialidade do TSSO em realmente substituir a função original.

Claramente, esse poder exploratório pode prejudicar a procura local se o método der uma ênfase muito grande à exploração global. Porém, é possível ver que o TSSO foi capaz de fazer uma excelente procura local, isso graças ao preditor SK ser utilizado no MEI. Com o fato de levarmos em consideração o preditor estocástico para construção do MEI, é disponibilizado à métrica a potencialidade de replicar pontos de suporte. Tais pontos não são replicados graças à nulidade do RMSE, porém podem levar o método para as regiões com um alta potencialidade de valores melhores.

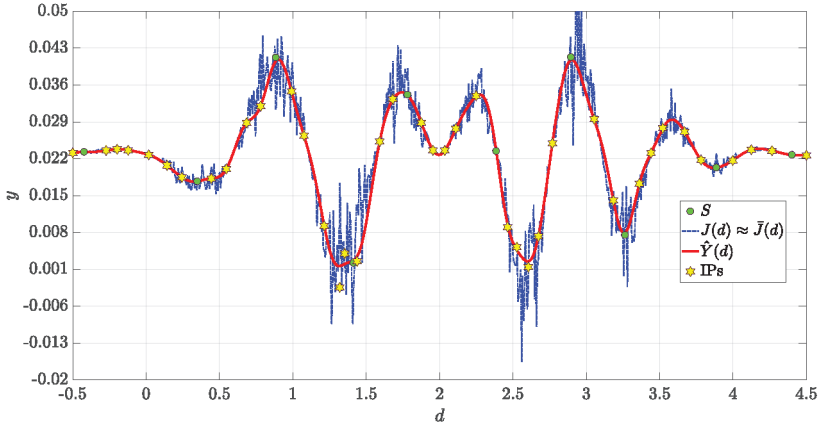


Figura 28 sEGO utilizando a métrica EIR.

#### 4.6.5 Análise do EIR

Por fim trazemos a representação gráfica do sEGO adaptativo utilizando o EIR como métrica na Figura 28. Dessa vez conseguimos adicionar 40 IPs com a técnica. Vemos que o comportamento do EIR também saiu como em sua definição. Nenhum ponto de suporte é selecionado como IP; dessa forma, não há a possibilidade de o método realizar replicações.

Podemos ver que o EIR também foi capaz de realizar uma excelente substituição do modelo, realizando uma grande exploração global do domínio de busca. Porém, se comparado ao TSSO, o que deixa a desejar no EIR é a busca local. Por consequência direta da reinterpolação, nesse método não é dada muita ênfase à busca local. Ao se utilizar o modelo reinterpolado, a variância espacial da matriz de correlação é substituída pela variância da reinterpolação (equação (4.39)). Essa substituição provoca uma certa inflação no valor da variância espacial, e que consecutivamente aumenta o RMSE do modelo SK (e também do Kriging determinístico como apresentado por (4.38)).

Como o EI preza por pontos que trazem grandes melhoras ou

grandes variâncias, após a localização de uma certa região de mínimo, a melhora dos valores da função nos pontos dessa região não serão superiores às variâncias em regiões ainda não exploradas. Assim, o método perfaz um IP exploratório, ao invés de refinar o valor em uma certa região do espaço.

Obviamente, por termos o RMSE nulo nos pontos de suporte, temos que o teorema de [Schonlau \(1997\)](#) se aplica e somos capazes de garantir a otimização da função objetivo. Porém, o não refinamento do EIR pode prejudicar a escolha do valor mínimo, já que deverão ser gastos inúmeras avaliações até obtermos um novo ponto em uma mesma região já explorada.

Porém, dentro das técnicas apresentadas, o EIR apresenta um excelente resultado pós-execução do sEGO adaptativo. Nela vemos o menor valor RMSE para o problema. Dessa forma, temos que os pontos avaliados são discrepantes dos valores originais apenas pelo erro cometido na aproximação da função, e representados pela variância do erro na aproximação.



## 5 RESULTADOS NUMÉRICOS

Para podermos testar as várias alternativas que o sEGO adaptativo nos fornece para solucionarmos o problema proposto em (1.2), será utilizado um conjunto de funções presentes na literatura acerca do tema. Cada uma dessas funções possui uma representação determinística. Logo, para o tratamento estocástico, serão necessários alguns ajustes sobre as funções. A transformação da função determinística em estocástica nos dá a possibilidade de sabermos qual o mínimo analítico que será objetivo do sEGO adaptativo. Essa informação facilita a comparação entre o resultado obtido pelo sEGO adaptativo e o resultado obtido na literatura.

No meio científico, para a execução do sEGO ou do EGO, é comum o uso de alguns softwares, entre eles temos o `DiceKriging` e o `DiceOptim` implementados por [Roustant, Ginsbourger e Deville \(2010\)](#), ambos implementados na linguagem R. Apesar destes algoritmos serem bem robustos e disponibilizarem uma gama de parâmetros ajustáveis, para que alcancemos a flexibilidade necessária para nosso trabalho e a maior confiabilidade nos resultados obtidos, julgou-se mais conveniente a criação de um novo algoritmo para lidar com o sEGO adaptativo. Além de nos proporcionar a liberdade necessária para as simulações, a implementação do código pelo autor trouxe lucidez a partes da teoria que seriam puramente abstratas. Pode-se afirmar que grande parte das compreensões adquiridas pelo autor sobre a forma como o metamodelo realiza as substituições e como o sEGO adaptativo se comporta, só foram possíveis graças à essa implementação.

Escolheu-se como base para a implementação o software comercial `MatLab`<sup>®</sup> ([MATHWORKS, 2017](#)) associada à linguagem de programação C++. Essa escolha foi baseada no conhecimento computacional do autor e seus orientadores, de forma a extrair a maior confiabilidade possível para os dados. Foi realizado um intenso estudo das funções do software, de forma a extrairmos a maior facilidade e operacionalidade do programa.

Para maior intelecção do método, este capítulo de soluções possui na sua seção 5.1 as funções que foram utilizadas, bem como todas as informações que as circundam. Na seção 5.2 são apresentados os parâmetros fixos do algoritmo para todos os problemas e que tornam funcional a implementação. Os resultados são apresentados a partir da Seção 5.3.

## 5.1 FUNÇÕES UTILIZADAS

### 5.1.1 Função 1D-1

Essa função é utilizada seguindo o trabalho de Carraro et al. (2019). Seja a função  $\psi : \mathcal{D} \times \Omega \rightarrow \mathbb{R}$  definida por

$$\psi(d, X) = -(1.4 - 3d) \text{sen}(18d) \cdot X, \quad (5.1)$$

onde  $d \in \mathcal{D} = [0, 1.2]$  é o domínio de busca e  $\Omega$  é o espaço unidimensional formado pela variável aleatória  $X \sim \mathcal{N}(1, \sigma_x)$ . Essa função é analisada para  $\sigma_x = 0.2$  e  $\sigma_x = 0.3$ .

A função determinística  $\psi(d, 1)$  possui quatro mínimos e quatro máximos locais em  $\mathcal{D}$  com  $\text{CD}_\psi = 3.5$ . O minimizador global se encontra em  $d^* = 0.966086$  com o valor mínimo de  $y_{\min} = -1.489072$ . Como o ruído aplicado é resultado da multiplicação da função pela variável aleatória, espera-se que qualquer versão estocástica dessa função também possua o mesmo mínimo.

Para a normalização dessa função são utilizados os valores fixos  $\gamma = 0.01$  e  $J_0 = -1.489072$ .

### 5.1.2 Função 1D-2

A segunda função em uma dimensão é a mesma função que foi utilizada no Capítulo 4 e pode ser conferida em (4.3). Para recordar, a definimos como  $\psi : \mathcal{D} \times \Omega \rightarrow \mathbb{R}$  dada pela lei

$$\psi(d, X) = \left[ (2d - 4) \exp[-(d^2 - 4d + 3)] \text{sen}(0.7d^2 + 4.9d) \right] \cdot X, \quad (5.2)$$

onde  $d \in \mathcal{D} = [-0.5, 4.5]$  é o domínio de busca e  $\Omega$  é o espaço unidimensional formado pela variável aleatória  $X \sim \mathcal{N}(1, \sigma_x)$ . Essa função é analisada para  $\sigma_x = 0.1$  e  $\sigma_x = 0.5$ .

Podemos ver na Figura 20(a) o gráfico da função determinística  $\psi(d, 1)$  a partir de (5.2), e dela vemos quais foram as intenções de sua definição. As características que desejamos dessa função são:

- Ser multimodal, possuindo a mesma quantidade de máximos e mínimos. Mais especificamente, seis mínimos locais, um mínimo global, seis máximos locais e um máximo global;
- Ter um valor mínimo local próximo do global (diferença de 3%), porém com minimizadores local e global em pontos distantes (diferença de 93%);
- Possuir altos ruídos amplificados nas regiões de mínimo. Dessa forma, a diferenciação entre os mínimos, local e global, será dificultada, podendo fazer o sEGO adaptativo cair no vale do mínimo local. Para visualização, pode-se recorrer à Figura 28;
- Qualquer versão estocástica dessa função possuirá o mesmo mínimo global que a versão determinística.

O valor mínimo da função é dado por  $y_{\min} = -2.319964$  no minimizador  $d^* = 1.340306$ . Teremos que para a função determinística  $CD_{\psi} = 4.447046$ . Para aplicarmos a normalização à função, utilizamos os valores fixos  $\gamma = 0.1$  e  $J_0 = -2.319964$ .

### 5.1.3 Função 1D-3

A terceira e última função em uma dimensão é tal que  $\psi : \mathcal{D} \times \Omega \rightarrow \mathbb{R}$  e sua lei será

$$\psi(d, X) = 1 + 0.2dX + \cos(0.3(dX)^2), \quad (5.3)$$

onde  $d \in \mathcal{D} = [-1, 7]$  é o domínio de busca e  $\Omega$  é o espaço unidimensional formado pela variável aleatória  $X \sim \mathcal{N}(1, \sigma_x)$ . O gráfico dessa função

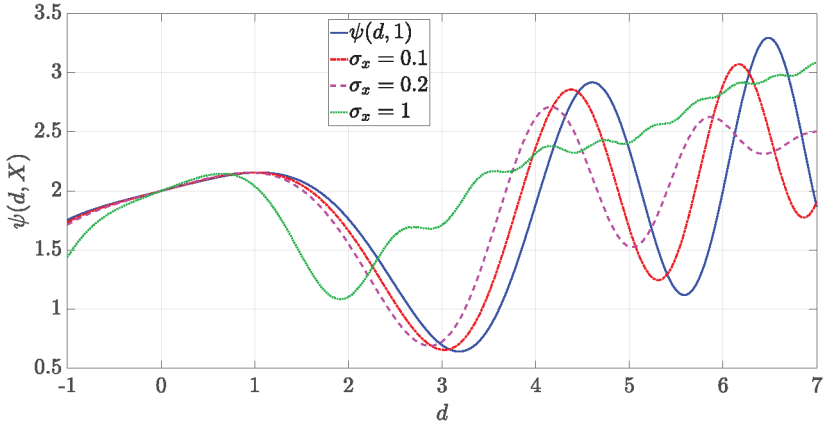


Figura 29 Gráficos determinístico e estocásticos para a função 1D-3.

para os casos onde  $X \sim \mathcal{N}(1, 0)$  (determinístico),  $X \sim \mathcal{N}(1, 0.1)$ ,  $X \sim \mathcal{N}(1, 0.2)$  e  $X \sim \mathcal{N}(1, 1)$  é apresentado na Figura 29. Para obtermos uma construção fidedigna da curva da função nas suas formas estocásticas, foram utilizadas 200 000 replicações da função por ponto.

Podemos observar que o ruído agora é aplicado de forma diferente. Multiplicamos a variável aleatória  $X$  ao valor da variável de projeto  $d$  e depois calculamos o valor da função. Podemos ver que o resultado é uma alteração das tendências da curva. Pode-se observar que para valores pequenos da variância, exemplo  $\sigma_x = 0.1$ , espera-se que o comportamento da função estocástica seja o mesmo da função determinística, possuindo minimizadores e valores mínimos próximos. Porém, conforme aumentamos o valor da variável aleatória, as nuances de (5.3) começam a se distanciar da curva determinística. Isso faz com o que os mínimos possam mudar sua localização. Dessa forma, para cada análise  $\sigma_x$  espera-se um valor mínimo diferente.

Para obtermos uma referência de valor mínimo e de minimizador para a função (5.3), podemos utilizar das condições de otimalidade da função. Sabemos que se  $d^*$  é um minimizador local de  $\psi(d, X)$ , então



devemos ter

$$\left. \frac{d\psi(d, X)}{dd} \right|_{d=d^*} = 0,$$

logo

$$\begin{aligned} 0.2X - 0.6d^* X \operatorname{sen}(0.3(d^* X)^2) &= 0 \\ 0.2X [1 - 3d^* \operatorname{sen}(0.3(d^* X)^2)] &= 0. \end{aligned} \quad (5.4)$$

Podemos ver que existe a probabilidade  $P(X = 0)$  para qualquer  $\sigma_x > 0$ , porém quando efetivamente  $X = 0$  a igualdade (5.4) será verdadeira. Entretanto, desconsiderando os valores nulos da variável aleatória  $X$ , teremos que  $d^*$  deverá satisfazer

$$1 - 3d^* \operatorname{sen}(0.3(d^* X)^2) = 0. \quad (5.5)$$

Como a equação (5.5) não é explícita em  $d^*$  não podemos resolvê-la de modo analítico. Logo, somos forçados a utilizar algum método numérico. Dessa forma, para que possamos obter uma aproximação para a solução, é utilizada a função `fzero` do MatLab<sup>®</sup> para solução de equações e sistemas não lineares. A função utilizada para busca pelo algoritmo é dada por

$$F(d) = 1 - 3d \operatorname{sen}(0.3(dX)^2),$$

cujos pontos iniciais de busca dependerão de  $\sigma_x$ . Por influência da Figura 29, tomamos, respectivamente,  $d_0 = 3, 2.8, 1.7$  para  $\sigma_x = 0.1, 0.2, 1$ . Como a função de busca depende de uma variável aleatória, tomamos o valor da função em  $d$  como sendo aproximadamente a média de 200 000 replicações da função para aquele  $d$  (como realizado para seu gráfico). A função `fzero` foi simulada 10 000 vezes e o minimizador (solução de (5.5)) é dado pela média dos minimizadores obtidos em cada simulação. Os valores obtidos estão apresentados na Tabela 2.

Na Tabela 2 também é apresentado o CD do contradomínio da função estocástica. Podemos ver que, diferentemente dos casos anteriores

o CD da função estocástica diminui. Porém, será apresentado nos resultados que a diminuição do CD não contorna a dificuldade apresentada pelas altas variâncias da variável aleatória.

A função 1D-3 é normalizada utilizando o parâmetro fixo  $\gamma = 0.01$  e os parâmetros  $J_0 = 0.65, 0.67, 1.07$  para  $\sigma_x = 0.1, 0.2, 1$ , respectivamente.

Tabela 2 – Principais características da função 1D-3.

$\sigma_x$	$d^*$	$y_{\min}$	$R_\psi$
0.1	3.023396	0.6552761	2.418663
0.2	2.870912	0.688963	2.025407
1	1.899993	1.084044	2.014521

#### 5.1.4 Função Branin Modificada

Como primeira função em duas dimensões iremos utilizar a já definida função Branin Modificada (4.46). Relembrando, temos  $\psi : \mathcal{D} \times \Omega \rightarrow \mathbb{R}$  dada pela lei

$$\psi(\mathbf{d}, \mathbf{X}) = \left( d_2 - \frac{5.1}{4\pi^2} d_1^2 + \frac{5}{\pi} d_1 - 6 \right)^2 \cdot X_1 + 10 \left( 1 - \frac{1}{8\pi} \right) \cos(d_1) \cdot X_2 + 10 + 5d_1, \quad (5.6)$$

onde  $\mathbf{d} \in \mathcal{D} = [-5, 10] \times [0, 15]$  é o domínio de busca,  $\mathbf{X} \in \Omega$ , onde  $\Omega$  é o espaço bidimensional formado por todos os pares de combinações possíveis entre os valores das variáveis aleatórias  $X_1 \sim \mathcal{N}(1, \sigma_x)$  e  $X_2 \sim \mathcal{N}(1, \sigma_x)$ . Analisamos os casos estocásticos da função fazendo  $\sigma_x = 0.01$  e  $\sigma_x = 0.05$ .

Nesse caso, temos que o ruído é aplicado de forma multiplicativa no valor da função, onde duas parcelas da função Branin Modificada

serão alteradas pelos ruídos. Logo é de se esperar que o minimizador da função estocástica seja igual ao da determinística  $\psi(\mathbf{d}, \mathbf{1})$ . A saber, o minimizador será  $\mathbf{d}^* = \{-3.689285, 13.629987\}^T$  com o valor mínimo de  $y_{\min} = -16.644021$ . Para a função determinística temos que  $\text{CD}_\psi = 299.773117$ .

A normalização é realizada ajustando os parâmetros fixos  $\gamma = 0.01$  e  $J_0 = -16.644021$ .

### 5.1.5 Função Rosenbrock

Essa função é comumente utilizada na literatura sobre otimização. Ela também é conhecida como *Banana's Function*. Seja  $\psi(\mathbf{d}, X) : \mathcal{D} \times \Omega \rightarrow \mathbb{R}$  dada por

$$\psi(\mathbf{d}, X) = \left[ 100(d_2 - d_1^2)^2 + (1 - d_1)^2 \right] \cdot X, \quad (5.7)$$

onde  $\mathbf{d} \in \mathcal{D} = [-5, 10]^2$  é o domínio de busca,  $X \in \Omega$ , onde  $\Omega$  é o espaço unidimensional formado pela variável aleatória  $X \sim \mathcal{N}(1, \sigma_x)$ . São analisados dois casos estocásticos com  $\sigma_x = 0.05$  e  $\sigma_x = 0.1$ .

Nesse caso em específico, temos uma função convexa e que possui uma superfície que assume valores elevados para pontos longe do mínimo global. Entretanto, a medida que o processo de otimização do sEGO adaptativo começa a se mover para a região do mínimo local, o valor da função diminui muitas ordens de grandeza. Essa característica a transforma em um excelente problema para se mostrar o poder da normalização. O minimizador da função determinística  $\psi(\mathbf{d}, \mathbf{1})$  será  $\mathbf{d}^* = \{1, 1\}^T$ , com o valor mínimo de  $y_{\min} = 0$ . Temos que  $\text{CD}_\psi = 1102581$ .

A normalização é executada com os parâmetros  $\gamma = 10^{-4}$  e  $J_0 = 0$ .

### 5.1.6 Função Hartmann 3D

A função estocástica Hartmann 3D é definida por  $\psi(\mathbf{d}, \mathbf{X}) : \mathcal{D} \times \Omega \rightarrow \mathbb{R}$  dada por

$$\psi(\mathbf{d}, \mathbf{X}) = - \sum_{i=1}^4 \left[ \alpha_i \exp \left( - \sum_{j=1}^3 A_{ij} (d_j \cdot X_j - P_{ij})^2 \right) \right], \quad (5.8)$$

onde  $\mathbf{d} \in \mathcal{D} = [0, 1]^3$  é o domínio de busca,  $\mathbf{X} \in \Omega$ , onde  $\Omega$  é o espaço tridimensional formado pelas combinações possíveis entre as variáveis aleatórias  $X_i \sim \mathcal{N}(1, \sigma_x)$ , para  $i = 1, 2, 3$ . Temos os parâmetros auxiliares

$$\boldsymbol{\alpha} = \{1, 1.2, 3, 3.2\}^T \quad \mathbf{A} = \begin{pmatrix} 3.0 & 10 & 30 \\ 0.1 & 10 & 35 \\ 3.0 & 10 & 30 \\ 0.1 & 10 & 35 \end{pmatrix}$$

$$\mathbf{P} = \begin{pmatrix} 0.3689 & 0.1170 & 0.2673 \\ 0.4699 & 0.4387 & 0.7470 \\ 0.1091 & 0.8732 & 0.5547 \\ 0.3810 & 0.5743 & 0.8828 \end{pmatrix}.$$

Para essa função faremos duas análises utilizando  $\sigma_x = 0.05$  e  $\sigma_x = 0.1$ . Como o ruído não é multiplicativo ao valor da função e sim na variável de projeto, temos que o mínimo muda suavemente de valor. Utilizando o mesmo procedimento apresentado na Subseção 5.1.3, encontramos para  $\sigma_x = 0.05$  que o minimizador será

$$\mathbf{d}^* = \{0.059108, 0.556578, 0.843213\},$$

com o valor mínimo de  $y_{\min} = -3.692819$ . Para  $\sigma_x = 0.1$  temos que o minimizador será

$$\mathbf{d}^* = \{0.054683, 0.552648, 0.831554\},$$

com o valor mínimo de  $y_{\min} = -3.294724$ .

Na função determinística temos que  $CD_\psi = 3.872872$ . A normalização é realizada com os parâmetros fixos  $\gamma = 0.05$  e  $J_0 = -3.3$ .

### 5.1.7 Função Colville

Trazemos agora uma função em quatro dimensões conhecida como função Colville. Ela é dada por  $\psi(\mathbf{d}, X) : \mathcal{D} \times \Omega \rightarrow \mathbb{R}$  e representada pela lei

$$\psi(\mathbf{d}, X) = \left[ 100(d_1^2 - d_2)^2 + (d_1 - 1)^2 + (d_3 - 1)^2 + 90(d_3^2 - d_4)^2 + 10.1 \left[ (d_2 - 1)^2 + (d_4 - 1)^2 \right] + 19.8(d_2 - 1)(d_4 - 1) \right] \cdot X, \quad (5.9)$$

onde  $\mathbf{d} \in \mathcal{D} = [-10, 10]^4$  é o domínio de busca,  $X \in \Omega$ , onde  $\Omega$  é o espaço unidimensional formado pela variável aleatória  $X \sim \mathcal{N}(1, \sigma_x)$ . Utilizaremos para análise as variâncias de  $\sigma_x = 0.01$  e  $\sigma_x = 0.05$ .

Basicamente, a escolha da função Colville parte dos mesmos requisitos da função Rosenbrock: possuir regiões afastadas do mínimo onde seu valor é bem maior do que no mínimo. Porém, a função Colville traz algumas complicações a mais, pois se trata de uma função multimodal com um mínimo global, um mínimo local e um máximo local. Os mínimos estão em vales separados pelo vale do máximo local, isso faz com que o ruído próximo à essas áreas seja grande, dificultando a modelagem SK e o sucesso do sEGO adaptativo.

Por adicionarmos o ruído aleatório como multiplicativo ao valor da função, o minimizador estocástico deve ser idêntico ao determinístico  $\psi(\mathbf{d}, 1)$ . Portanto temos que  $\mathbf{d}^* = \{1, 1, 1, 1\}^T$  com  $y_{\min} = 0$ . Temos  $CD_\psi = 2\,304\,082$  para o caso determinístico. A normalização é aplicada utilizando os parâmetros fixos  $\gamma = 10^{-4}$  e  $J_0 = 0$ .

### 5.1.8 Função Hartman 6D

Uma outra versão da função Hartmann pode ser tomada em seis dimensões. Podemos defini-la por  $\psi(\mathbf{d}, \mathbf{X}) : \mathcal{D} \times \Omega \rightarrow \mathbb{R}$  dada por

$$\psi(\mathbf{d}, \mathbf{X}) = - \sum_{i=1}^4 \left[ \alpha_i \exp \left( - \sum_{j=1}^6 A_{ij} (d_j \cdot X_j - P_{ij})^2 \right) \right], \quad (5.10)$$

onde  $\mathbf{d} \in \mathcal{D} = [0, 1]^6$  é o domínio de busca,  $\mathbf{X} \in \Omega$ , onde  $\Omega$  é o espaço de seis dimensões formado pelas combinações possíveis entre as variáveis aleatórias  $X_i \sim \mathcal{N}(1, \sigma_x)$ , para  $i = 1, 2, \dots, 6$ . Temos também os parâmetros auxiliares

$$\boldsymbol{\alpha} = \{1, 1.2, 3, 3.2\}^T \quad \mathbf{A} = \begin{pmatrix} 10 & 3 & 17 & 3.5 & 1.7 & 8 \\ 0.05 & 10 & 17 & 0.1 & 8 & 14 \\ 3 & 3.5 & 1.7 & 10 & 17 & 8 \\ 17 & 8 & 0.05 & 10 & 0.1 & 14 \end{pmatrix}$$

$$\mathbf{P} = \begin{pmatrix} 0.1312 & 0.1696 & 0.5569 & 0.0124 & 0.8283 & 0.5886 \\ 0.2329 & 0.4135 & 0.8307 & 0.3736 & 0.1004 & 0.9991 \\ 0.2348 & 0.1451 & 0.3522 & 0.2883 & 0.3047 & 0.6650 \\ 0.4047 & 0.8828 & 0.8732 & 0.5743 & 0.1091 & 0.0381 \end{pmatrix}.$$

Solucionamos essa função para dois valores alternativos de variância,  $\sigma_x = 0.05$  e  $\sigma_x = 0.1$ . Por termos o ruído aplicado ao valor da variável de projeto, então o valor mínimo deve diferir levemente do valor determinístico. Utilizando o apresentado na Subseção 5.1.3 encontramos que para  $\sigma_x = 0.05$  temos o minimizador

$$\mathbf{d}^* = \{0.201891, 0.153637, 0.467157, 0.274308, 0.307810, 0.652522\}^T,$$

com o valor mínimo de  $y_{\min} = -3.268801$ . Para  $\sigma_x = 0.1$  temos o minimizador

$$\mathbf{d}^* = \{0.218380, 0.147480, 0.451787, 0.271590, 0.307490, 0.652723\}^T,$$

com o valor mínimo de  $y_{\min} = -3.118222$ . O valor da extensão do contradomínio da função determinística  $\psi(\mathbf{d}, \mathbf{1})$  é de  $\text{CD}_\psi = 3.322369$ . A normalização é aplicada com os parâmetros  $\gamma = 0.05$  e  $J_0 = -3.3$ .

### 5.1.9 Função Levy 10D

Por fim apresentamos a função Levy 10D. Essa função tem por mérito trazer o maior desafio ao sEGO adaptativo. Se trata de uma função com múltiplos mínimos e máximos, bem como possui valores altos para pontos distantes do minimizador global. Uma grande dificuldade para qualquer otimizador resolver essa função é vencer os grandes vales entre os mínimos locais para se alcançar um vale cujo valor da função seja menor.

A definiremos como  $\psi(\mathbf{d}, X) : \mathcal{D} \times \Omega \rightarrow \mathbb{R}$  dada pela lei

$$\psi(\mathbf{d}, X) = \left\{ \begin{aligned} &\text{sen}^2(\pi D_1) + \sum_{i=1}^9 \left\{ (D_i - 1)^2 [1 + 10 \text{sen}^2(\pi D_i + 1)] \right\} \\ &+ (D_{10} - 1)^2 [1 + \text{sen}^2(2\pi D_{10})] \end{aligned} \right\} \cdot X, \quad (5.11)$$

onde  $\mathbf{d} \in \mathcal{D} = [-10, 10]^{10}$  é o domínio de busca,  $X \in \Omega$ , onde  $\Omega$  é o espaço unidimensional formado pela variável aleatória  $X \sim \mathcal{N}(1, \sigma_x)$ . Por fim, temos que

$$D_i = 1 + \frac{d_i - 1}{4}.$$

A análise dessa função é dada apenas para  $\sigma_x = 0.01$ .

Como o ruído é multiplicativo ao valor da função, temos que o minimizador para a versão estocástica será idêntico ao da versão determinística. Logo, temos que o minimizador será dado por

$$\mathbf{d}^* = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1\}^T,$$

com o valor mínimo de  $y_{\min} = 0$ . Para  $\psi(\mathbf{d}, 1)$  temos  $\text{CD}_\psi = 733.445281$ . A normalização é executada com  $\gamma = 0.01$  e  $J_0 = 0$ .

---

Definidas as nove funções que são analisadas, temos que no total serão resolvidos dezoito problemas de otimização, se contarmos

as diferentes variações para a variável estocástica. Os problemas são dados por (1.2) onde  $J(\mathbf{d})$  é dado por (1.1). Para a função  $J(\mathbf{d})$  temos que  $\psi(\mathbf{d}, \mathbf{X})$  é dada por uma entre as dez leis apresentadas de (5.1) a (5.3) e de (5.6) a (5.11), enquanto a função peso  $f_{\mathbf{x}}$  assume a função de densidade de probabilidade para o vetor aleatório  $\mathbf{X}$ . Para simplificação, seguimos no restante deste capítulo a nomenclatura apresentada na Tabela 3.



Tabela 3 – Nomenclatura das funções analisadas.

Símbolo	Função	$\sigma_x$	NFE
F1	1D-1	0.2	80
F2		0.3	
F3	1D-2	0.1	150
F4		0.5	
F5	1D-3	0.1	150
F6		0.2	
F7		1	
F8	Branin Modificada	0.01	100
F9		0.05	
F10	Rosenbrock	0.05	150
F11		0.1	
F12	Hartman 3D	0.05	100
F13		0.1	
F14	Colville	0.01	200
F15		0.05	
F16	Hartman 6D	0.05	240
F17		0.1	
F18	Levy 10D	0.01	250

## 5.2 CONSIDERAÇÕES E CONFIGURAÇÕES GERAIS

### 5.2.1 Configurações básicas para execução do sEGO

Nesta seção apresentamos alguns parâmetros e funções que fazem com que o sEGO adaptativo funcione. Nosso intuito é simplesmente dar valores a alguns parâmetros que possam parecer abstratos durante suas definições. Eles são apresentados em tópicos que não necessariamente são correlacionados. Vamos a eles:

1. Para o número de elementos do espaço amostral inicial  $S$ , utilizamos a sugestão apresentada por [Jones, Schonlau e Welch \(1998\)](#) de  $n = 10k$ , onde  $k$  é a dimensão do problema. Somente a função F18 possui  $n = 7k$ ;
2. O espaço amostral inicial é formado por um Hipercubo Latino. Este por sua vez, é determinado pela função `lhsdesign` do MatLab® utilizando seus parâmetros padrões;
3. O número de replicações para cada ponto de suporte inicial é tomado como  $n_r = 2$ , que é o valor mínimo para que consigamos calcular a variância do erro. Essa configuração reserva o orçamento computacional para a etapa de pesquisa adaptativa, onde precisamos de um refinamento maior da função;
4. O critério de parada de execução do algoritmo sEGO adaptativo é o Número de Avaliações da Função (*Number of Funtion Evaluations - NFE*) disponíveis. Na Tabela 3 temos apresentado na última coluna qual o NFE máximo permitido para cada problema analisado;
5. A variância mínima, alvo ou adaptativa, foi considerada igual a  $\sigma_{\min}^2 = 10^{-10}$  para todas as funções;
6. Os parâmetros  $\sigma_z^2$  e  $\theta$  que ajustam o modelo SK assumem somente valores positivos, dessa forma é utilizado em todo o processo potências de 10 ao invés de seu valor original. Ou seja, se por exemplo, o parâmetro  $\sigma_z^2$  precisasse assumir o valor 746.3129,

então tomamos que  $\sigma_z^2 = 2.8729$  e utilizamos  $10^{2.8729}$  como o valor do parâmetro no código;

7. Para a otimização da função de verossimilhança (4.13) (Seção 4.2.3) foi aplicado o algoritmo heurístico PSO (Seção 2.4.3). Utilizamos a função `particleswarm` do próprio MatLab<sup>®</sup>. Para o número de indivíduos no enxame utilizamos  $20k$ . Os limites de busca para todos parâmetros de ajuste foram tomados no intervalo  $[-4, 4]$ . A consideração desse intervalo de busca foi baseada em inúmeras execuções do programa, onde se viu que nenhum parâmetro ficou menor que  $0.0001 = 10^{-4}$  e nem maior do que  $10000 = 10^4$ . Os demais parâmetros básicos do PSO são considerados como os padrões que vem implementados na função `particleswarm`. Essa abordagem será utilizada em todos os dezoito casos analisados;
8. Para a métrica MQ foi utilizado  $\beta = 0.5$ , representando a média do SK. Para encontramos a melhor solução efetiva  $\mathbf{d}^{**}$  no AEI, procuramos em  $S$  qual o ponto de suporte que minimiza a métrica MQ com  $\beta = 0.841345$ . Para o EQI utilizamos  $\beta = 0.9$ ;
9. Para otimização das métricas de adição de IPs também utilizamos o PSO como comentado no item 6. A diferença está no número de indivíduos do enxame, onde nesse caso, utilizamos a regra  $50k$ ;
10. Toda execução do sEGO adaptativo começa com uma determinada variância alvo inicial  $\bar{\sigma}_{\text{alvo}}^2 = \bar{\sigma}_0^2$ . No momento em que um IP é encontrado, sua avaliação é realizada tomando a variância alvo igual a variância adaptativa  $\bar{\sigma}_{\text{alvo}}^2 = \bar{\sigma}_{\text{adap}}^2$ ;
11. No cálculo do  $n_c$  para a variância adaptativa, é necessário o ajuste do parâmetro  $r_{hc}$ . Seguindo o trabalho de Carraro et al. (2019) utilizamos  $r_{hc} = 0.1$ .

### 5.2.2 Considerações gerais acerca das soluções apresentadas

Temos que ao longo de uma execução do sEGO adaptativo, a criação do plano de amostra inicial, as simulações utilizadas para

o cálculo do MCI e os indivíduos utilizados pelo PSO, dependem de variáveis que são aleatórias e não controladas pelo usuário. Assim, é de se esperar que a cada execução do sEGO adaptativo tenhamos resultados diferentes. Logo, não há como verificarmos a boa qualidade do método de otimização, simulando apenas uma única vez o problema.

Uma forma de contornarmos esse problema é apresentarmos os resultados de forma estatística como sugerido por [Gomes et al. \(2018\)](#), utilizando uma amostra de execuções do método. Dessa forma, para cada um dos dezoito problemas analisados, executamos 30 simulações do sEGO adaptativo. Desse conjunto de soluções, apresentamos os resultados de forma estatística utilizando a técnica de Gráficos em Caixa (*BoxPlots* - BP). A Figura 30 apresenta o tipo de BP que é utilizado neste trabalho. Nesse BP temos que o segmento dentro do retângulo representa a mediana das 30 soluções. O retângulo representa todas as soluções entre aquelas que são respectivamente os percentis de 10% (lado inferior) e de 90% (lado superior). Os segmentos pontilhados externos ao retângulo compreendem 99.83% das soluções obtidas do problema. Os pontos vermelhos que por ventura aparecerem acima dos segmentos pontilhados vão representar as piores soluções obtidas pelo sEGO adaptativo entre as 30 simulações. Para considerar a solução como a pior, tomamos a média e o desvio padrão das trinta soluções. Após, consideramos que as trinta soluções fazem parte de uma distribuição normal dada por esta média e desvio. As piores soluções são aquelas que estão a aproximadamente 2.7 desvios da média, compreendendo os pontos fora de 99.3% da distribuição.

Cada simulação é realizada a partir de um espaço amostral inicial  $S$  diferente. Porém, para cada uma das métricas é utilizado o mesmo espaço amostral daquela simulação. Isso garante a melhor comparação entre os resultados obtidos.

Analisamos a qualidade do sEGO adaptativo por meio do valor mínimo obtido em cada simulação. Porém, como o sEGO adaptativo funciona como um algoritmo de preenchimento e não um algoritmo

sequencial, não podemos afirmar que o último IP adicionado será aquele com o menor valor da função objetivo. Em adição, com problemas estocásticos devemos tomar cuidado com o valor final a ser considerado como mínimo, levando em consideração a variação dos parâmetros estocásticos (PICHENY et al., 2013; JALALI; NIEUWENHUYSE; PICHENY, 2017). Dessa forma, são considerados duas alternativas de extração do melhor minimizador:

1. A primeira alternativa, que será chamada de P1, é realizar uma otimização global do metamodelo SK criado com o espaço amostral no fim do sEGO adaptativo. Essa abordagem se justifica, pois, calcular o valor da predição do SK (4.15) é computacionalmente barato, a partir do momento em que já temos o modelo construído, além do que, o SK carrega a informação necessária das variabilidades impostas aos problemas estocásticos. Para tal otimização precisamos de um algoritmo global robusto para pesquisa do ótimo global. Assim, utilizamos o algoritmo SGA (GONÇALVES; LOPEZ; MIGUEL, 2015), como comentado na Seção 2.4.3.
2. A segunda alternativa, nomeada por P2, é escolher como minimizador o ponto que minimiza um percentil de 70% do SK, ou seja, aplicar  $\beta = 0.7$  em (4.17). Essa abordagem também é realizada por Picheny e Ginsbourger (2014) e se justifica por ser a forma mais simples de se extrair um ponto que leve em conta a variabilidade

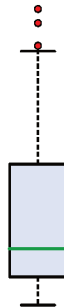


Figura 30 BP para análise estatística.

dos dados, já que somamos à média do valor do SK uma parcela com o erro RMSE da predição. Portanto, o ponto que gerar o menor valor predito ponderado por 70% do RMSE é escolhido como minimizador.

Escolhido um ponto como minimizador, o valor ótimo do problema é obtido simulando um MCI com  $n_r \rightarrow \infty$ . Para tanto, calculamos o MCI no minimizador, utilizando  $n_r = 100\,000$  replicações.

### 5.3 sEGO ADAPTATIVO COM NORMALIZAÇÃO E SEM NORMALIZAÇÃO

A primeira rodada de resultados é utilizada para mostrar o desempenho do sEGO adaptativo utilizando a normalização (4.47) em comparação com a aplicação do sEGO adaptativo sem a normalização. Nessa etapa, além de todos os ajustes comentados anteriormente, utilizamos como variância alvo inicial o valor de  $\bar{\sigma}_0^2 = 0.01$ . Por uma questão de simplificação, são apresentados os resultados obtidos pelo sEGO adaptativo considerando somente o valor mínimo obtido com o minimizador P2. Essa abordagem será justificada na próxima seção.

Na Figura 31 são apresentados os BP obtidos após as trinta simulações para os dezoito casos analisados. Nessa figura os resultados são apresentados para cada uma das cinco métricas, onde no bloco de uma mesma métrica, estão contidos dois BP, o da esquerda (laranja) representa o sEGO adaptativo com a normalização, enquanto o da direita (vermelho) representa o sEGO adaptativo sem a normalização. Os números que aparecem logo abaixo de um BP representam a média de IPs que foram adicionados por aquela determinada métrica. A reta contínua na cor azul representa o objetivo do sEGO adaptativo, ou seja, o valor mínimo da função.

Podemos ver que para as funções 1D, Figuras 31(a) a 31(g), todas as cinco métricas foram capazes de obter pontos cujos valores mínimos são bem próximos do objetivo. Isso acontece tanto para o caso

com e sem normalização. Porém, podemos ver que a grande diferença entre o *sEGO* adaptativo aplicado com e sem a normalização, está na quantidade de IPs adicionados. Podemos ver nestes casos que o *sEGO* adaptativo rodando com a normalização é capaz de adicionar uma quantidade muito superior de IPs utilizando o mesmo orçamento computacional. Dessa forma, o espaço de busca pode ser mais explorado ou refinado, e o valor mínimo obtido com maior qualidade.

Podemos ver nos casos 1D como o CD da função afeta diretamente a quantidade de IPs adicionados. Para as funções F1 e F2 temos que quase nenhuma diferença é vista entre as versões sem e com normalização, já que o CD da função é pequeno e não sofre uma variabilidade muito grande. Dessa forma, pontos de suporte iniciais já podem ser ótimos candidatos a minimizadores. Isso explica o fato de sem a normalização conseguirmos obter valores bem próximos do objetivo. Esse mesmo comportamento é visto para as funções F3 e F4. Porém, para as funções F5, F6 e F7, podemos ver que conforme o CD da função diminui, a quantidade de IPs adicionados sem a normalização tende a aumentar. Isso nos mostra como o CD da função pode influenciar na busca do *sEGO* adaptativo.

Tomando agora os casos mais extremos, vamos para as funções F8, F9, F10 e F11 (funções 2D) onde começamos a ver o excelente resultado da normalização. Essas são funções que possuem um CD bem superior aos casos 1D; logo, temos que sem a normalização, somente no caso F8 (menor variabilidade estocástica e de CD) as métricas não adicionaram somente 1 IP. Para os demais casos, o único IP adicionado é contabilizado, porém, o mesmo foi calculado até atingir o NFE máximo e não a variância alvo.

Agora, com a normalização, temos que a variabilidade das soluções obtidas pelo *sEGO* adaptativo é muito inferior ao caso sem normalização. Isso traz uma robustez muito maior para estes casos. Essa robustez está associada diretamente ao número de IPs adicionados.

Nas funções F12 e F13 temos os casos 3D, que também possuem

resultados parecidos com os anteriores. Novamente por conter um CD não tão grande, os casos sem a normalização também possuem um bom resultado, mas ainda não são comparáveis à aplicação da normalização. Isso pode ser conferido pelas medianas das soluções, onde em todos os casos com a normalização obtivemos medianas menores.

Para as funções F14 e F15 (caso 4D) temos novamente uma função com CD muito alto e multimodal. Nestes casos já somos capazes de ver que com a normalização, até o percentil de 90% das soluções, já é melhor que o menor valor obtido pelo caso sem a normalização. Novamente observamos uma enorme discrepância entre o número de IPs adicionados.

Para as funções em seis dimensões, funções F16 e F17, temos que o fato da função ser bem multimodal traz um pouco de complicações para os dois casos, porém mesmo assim o caso com a normalização foi superior. Por fim temos o caso 10D, com a função F18, onde o caso sem normalização não conseguiu trazer tantos benefícios quanto o caso com normalização. Novamente, tivemos que o percentil de 90% do caso com normalização foi muito superior ao melhor resultado obtido sem a normalização.

Podemos ressaltar que todas as métricas foram capazes de encontrar um excelente minimizador. Portanto, não cabe aqui destaque para qual métrica é superior, isso será feito na próxima seção. Porém, podemos verificar que a métrica EIR é a que cumpre o melhor papel de todos em varrer o domínio de busca, adicionando a maior quantidade de IPs. Somente em quatro dos dezoito problemas é que tivemos outras métricas com a mesma quantidade de IPs adicionados que o EIR, mas nenhuma métrica foi capaz de adicionar mais pontos do que o EIR. Entretanto, o fato de não ser um exímio refinador de soluções, faz com que o EIR as vezes não seja a métrica com a melhor solução entre todas.

Para finalizar, no Apêndice D temos a Tabela 6 na qual constam alguns dos valores estatísticos obtidos pelas cinco métricas do sEGO adaptativo nos problemas F7, F9, F11, F13, F15, F17 e F18. Em cada



problema e para cada métrica, são apresentados o valor da mediana, o melhor ( $\bar{J}_{\min}^{\text{melhor}}$ ) e o pior ( $\bar{J}_{\min}^{\text{pior}}$ ) valores mínimos obtidos pelo sEGO adaptativo.

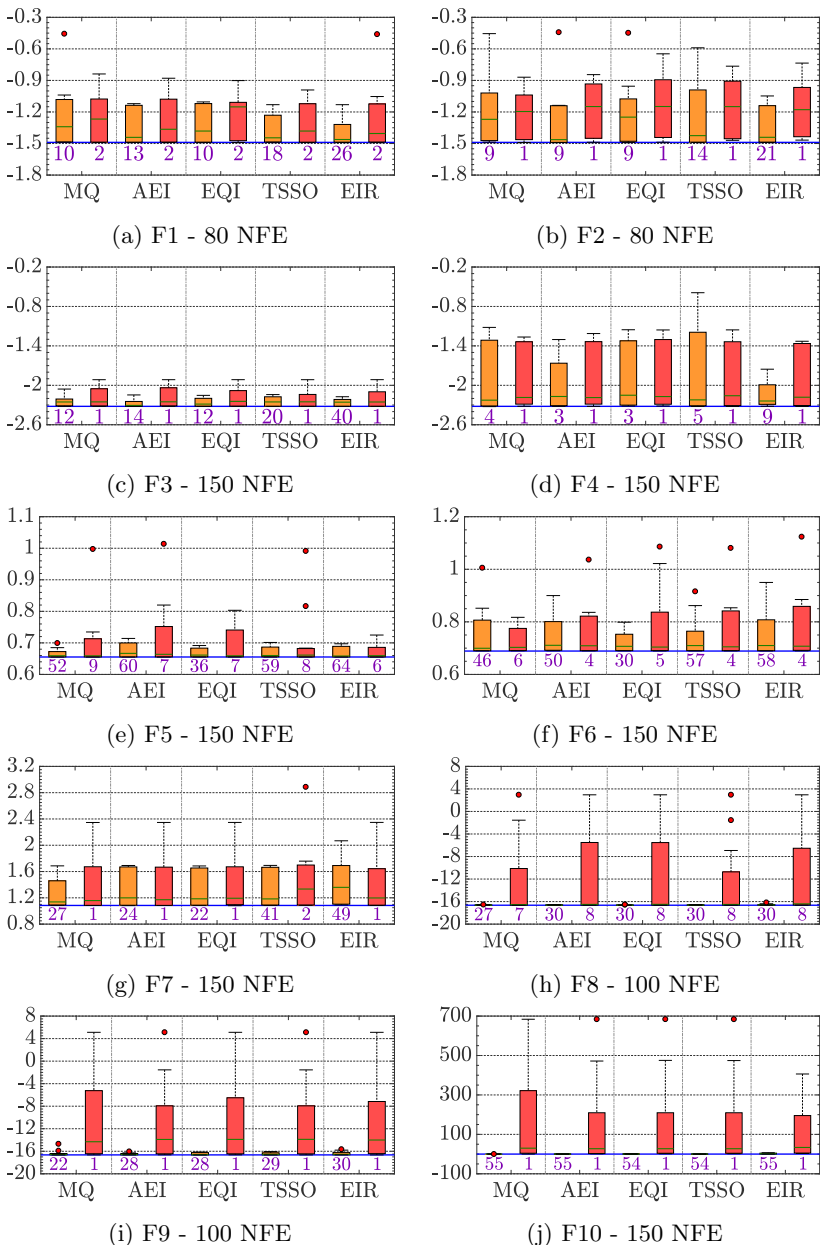


Figura 31 – Comparação entre o sEGO adaptativo com normalização e sem normalização. Os retângulos laranja e vermelho representam, respectivamente, os resultados com e sem normalização.

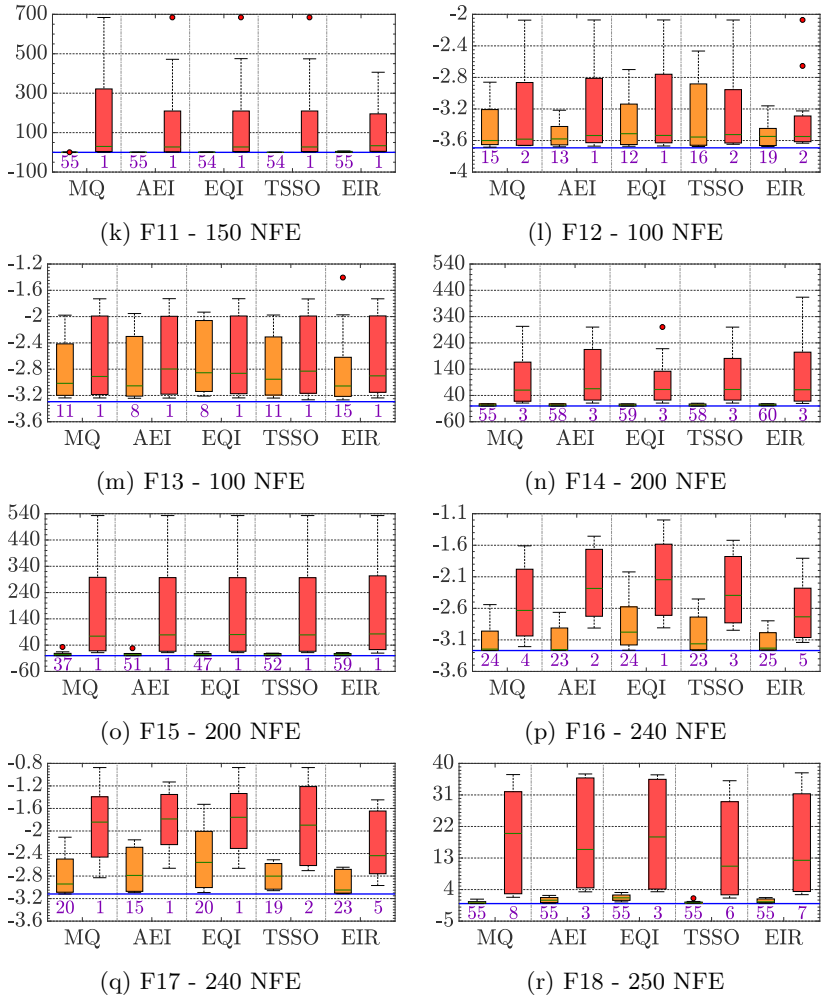


Figura 31 – (Continuação) Comparação entre o sEGO adaptativo com normalização e sem normalização. Os retângulos laranja e vermelho representam, respectivamente, os resultados com e sem normalização.

#### 5.4 INFLUÊNCIA DA VARIÂNCIA INICIAL PARA O sEGO ADAPTATIVO

Nesta seção fazemos uma análise de como o valor da variância inicial  $\bar{\sigma}_0^2$  pode influenciar na execução do sEGO adaptativo. Como visto

na seção anterior, é utilizada somente a abordagem com a normalização, já que esta traz muito mais robustez para o sEGO adaptativo.

Recordando, o valor da variância alvo inicial influencia diretamente toda a busca adaptativa do sEGO adaptativo. Se ajustarmos um valor alvo muito grande, corremos o risco de não conseguir avaliar a função com qualidade e ficarmos preso em uma região com alta variabilidade. Por outro lado, se ajustarmos a variância alvo muito baixa, trazemos uma qualidade superior para a avaliação da função, porém a um alto custo do orçamento computacional disponível.

Utilizamos para as análises três valores diferentes para a variância alvo inicial,  $\bar{\sigma}_0^2 = 0.1, 0.01, 0.0001$ . São analisadas as cinco métricas do sEGO adaptativo, utilizando os valores mínimos obtidos pelo algoritmo com os minimizadores do tipo P1 e P2 ao final da execução do sEGO adaptativo.

Os resultados são apresentados na Figura 32. Nessa figura temos que cada gráfico mostra os BP sobre as trinta simulações executadas. Foquemos em um bloco de uma determinada métrica. Podemos ver que há 6 BP nesse bloco, os três primeiros (amarelos) representam os valores mínimos obtidos com o minimizador P1, enquanto os três últimos (verdes) representam os valores mínimos obtidos pelo minimizador P2. Do subconjunto de três BP em um mesmo tipo de minimizador, o primeiro representa os resultados para  $\bar{\sigma}_0^2 = 0.1$ , o do meio  $\bar{\sigma}_0^2 = 0.01$  e o último  $\bar{\sigma}_0^2 = 0.0001$ . A reta contínua azul representa o objetivo do sEGO adaptativo, enquanto os valores numéricos, logo abaixo dessa reta e junto de um BP, representam a média de IPs adicionados naquele caso.

Analisando o conjunto completo de todos os problemas, vemos que a influência mais significativa está novamente relacionada à quantidade de IPs adicionados. Conforme a variância inicial diminui, aumenta-se o número de replicações em cada ponto, e naturalmente isso nos leva a um consumo mais rápido do orçamento computacional. Como vimos na seção anterior, as métricas que fazem a maior adição de

IPs para o sEGO adaptativo normalmente trazem maior robustez ao método. Assim, mostramos que realmente não deve-se optar por valores tão baixos para a variância inicial.

Dos resultados apresentados por Carraro (2017) e Carraro et al. (2019), concluímos que o ajuste de  $\bar{\sigma}_0^2$  é crucial para um bom resultado do sEGO adaptativo. Porém, com a normalização somos capazes de agregar ao sEGO adaptativo uma menor sensibilidade à variância inicial, onde é perceptível que em praticamente todos os casos e em todas as métricas, conseguimos alcançar o valor objetivo. A única exceção que cabe destaque é o da Figura 32(m), representando a função Hartman 3D, onde vemos uma maior dificuldade das métricas em aproximar o valor mínimo. Porém, os resultados obtidos ainda podem ser considerados razoáveis, dados a natureza do tipo de problema.

No Apêndice D apresentamos a Tabela 7 com os resultados obtidos para os problemas F9, F13, F15, F17 e F18.

#### 5.4.1 Uma análise via testes estatísticos para a influência da variância inicial

Para dar mais ênfase sobre como a normalização afeta a sensibilidade do sEGO adaptativo à variância alvo inicial, realizamos três testes estatísticos sobre as amostras de resultados advindos da métrica MQ, considerando o valor mínimo do problema de otimização com o minimizador do tipo P2. Todos os testes que serão realizados podem ser encontrados em mais detalhes no livro de (CONOVER, 1999).

O primeiro teste a ser realizado é o teste de Kruskal-Wallis, o qual serve para atestar se as soluções obtidas pelo sEGO adaptativo utilizando cada uma das três variâncias iniciais pertencem à mesma distribuição de probabilidade. Atribuímos ao teste as seguintes hipóteses:

$H_0$ : As três funções de distribuição são idênticas;

$H_1$ : Pelo menos uma das populações segue uma distribuição diferente

de pelo menos uma das outras duas populações.

Ajustando um nível de significância de 0.05, se falharmos em rejeitar a hipótese nula, então não haverá evidências estatísticas suficientes para mostrar que os resultados pertencem a distribuições diferentes. Esse teste serve para que possamos mostrar que o sEGO adaptativo utilizando a normalização, gera resultados próximos e que pertencem a uma mesma distribuição, seguindo um mesmo nível de variabilidade. Chamamos esse teste de T1.

Os próximos dois testes utilizados são testes não paramétricos e que possuem a mesma função entre si. Eles servirão para verificarmos se os resultados para as três variâncias não são idênticos. Utilizamos os testes de Friedman e de Quade para tal fim. Ambos os testes possuem as seguintes hipóteses:

$H_0$ : Os resultados para as três variâncias são idênticos;

$H_1$ : Os resultados não são idênticos.

Novamente, a um nível de significância de 0.05, se falharmos em rejeitar a hipótese nula, então não haverá evidências estatísticas suficientes para mostrar que os resultados são diferentes. Chamamos os testes de Friedman e de Quade por T2 e T3, respectivamente.

Podemos ver na Figura 33 o resultado obtido pelos testes. Nela temos representado pelo retângulo horizontal superior a região de rejeição dos testes. Dessa forma, quando o retângulo de um determinado teste atinge essa região temos que rejeitar a hipótese nula daquele determinado teste. Porém, podemos ver que em nenhum problema tivemos a rejeição de um dos testes. Assim, pode-se verificar que para a métrica MQ, a normalização traz resultados que podem ser classificados dentro de uma mesma distribuição de probabilidade, além de parecerem idênticos.

Para as outras métricas, fizemos uma análise análoga, porém descartamos sua apresentação já que o resultado é muito próximo do

obtido pelo MQ. Logo, podemos dizer que as variâncias iniciais  $\bar{\sigma}_0^2 = 0.1, 0.01, 0.0001$  resultaram em soluções que não podem ser diferenciadas entre si. Assim, podemos concluir que, de fato, a normalização traz uma menor sensibilidade à esse parâmetro, e com isso aumenta a robustez do sEGO adaptativo.

Com esse teste, o nosso intuito é dar enfoque que a normalização proporciona uma melhor robustez ao método sEGO adaptativo, melhorando significativamente os resultados obtidos por Carraro (2017). Porém, em hipótese nenhuma podemos dizer que o ajuste da variância inicial pode ser realizado de qualquer forma. O que podemos dizer é que esse resultado cria um certo intervalo para o valor da variância inicial. No entanto, a análise do problema pelo usuário ainda é a melhor saída para extrairmos a máxima performance do sEGO adaptativo.

#### 5.4.2 Análise do pior desempenho do sEGO adaptativo a partir das variâncias iniciais

Para finalizarmos esta seção de comparação entre as variâncias iniciais, faremos uma análise da pior performance do sEGO adaptativo para cada nível de variância inicial. Escolhemos como análise da performance seis estatísticas diferentes para as simulações: a mediana, o desvio padrão em torno da média (tem a função de computar a variabilidade dos dados), o melhor valor mínimo (considerado como o menor valor obtido), o pior valor mínimo (considerado como o maior valor obtido), o percentil de 10% e o percentil de 90%. Nosso intuito é realizar uma análise mais de tendência do sEGO adaptativo, podendo ser encarada como um estudo qualitativo da técnica.

A coleta dos resultados foi realizada da seguinte forma: para cada um dos dezoito problemas e para cada uma das três variâncias iniciais, fizemos o agrupamento dos valores mínimos do sEGO adaptativo, obtidos pelas cinco métricas, e com os dois tipos de minimizadores (P1 e P2). Após isso, aplicamos a estatística de análise a cada grupo de resultados. Dos três grupos iniciais, em um mesmo problema, obtemos

três subgrupos (um para cada variância) com 10 resultados estatísticos (5 métricas  $\times$  2 tipos de minimizadores). Desse novo subgrupo escolhemos qual possui a melhor estatística (por exemplo, menor mediana, menor melhor mínimo, etc.). Assim, restaremos com um único valor que representa a melhor estatística para determinada variância. Por fim encontramos qual variância apresenta o pior desempenho e computamos uma unidade a essa variância, por problema analisado. Como estamos trabalhando com problemas de minimização, o pior desempenho será representado pelo valor máximo entre as três variâncias.

Os resultados estão apresentados na Tabela 4. Logo, por exemplo, podemos observar que o sEGO adaptativo utilizando a variância inicial de 0.01, apresentou a pior mediana em apenas 1 dos dezoito problemas. Podemos verificar que a variância de 0.0001 só não foi pior em todos os problemas, unicamente na estatística pior mínimo, onde a variância de 0.1 gerou um sEGO adaptativo com o maior pior mínimo a mais. Portanto, vemos que a variância de 0.0001 foi aquela que apresentou as maiores medianas, a maior variabilidade de resultados, o pior melhor mínimo e os maiores percentis de 10% e 90%.

Podemos ver que, se somarmos os valores de uma mesma coluna dessa tabela, com exceção da última linha, teremos os valores 24, 22 e 62. Esse resultado por si só já poderia ser decisivo para se afirmar que a variância de 0.0001 é a que apresenta o pior desempenho, enquanto a variância de 0.01 apresenta o melhor desempenho. Porém, como estes valores numéricos ultrapassam o valor de dezoito, vamos criar um resultado alternativo para decidir qual variância tem o pior desempenho. Ponderamos cada resultado por meio de um nível de significância daquela estatística. Assim sendo, utilizamos a seguinte expressão:

$$w = 0.5 \cdot \text{Mediana} + 0.15 \cdot (\text{MelhorMínimo} + \text{Desvio}) + \\ 0.1 \cdot \text{PiorMínimo} + 0.05 \cdot (\text{Percentil10} + \text{Percentil90}), \quad (5.12)$$

para medir a qualidade dos resultados. Nessa expressão utilizamos que: 50% de importância da qualidade é dada pela quantidade de piores

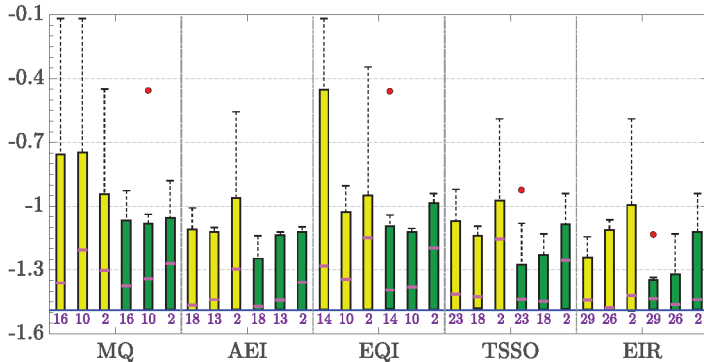


medianas; as estatísticas de melhor mínimo e desvio padrão representam, cada uma, 15% da qualidade final; a quantidade de pior mínimo agrega 10% à qualidade; e por fim cada um dos percentis influenciam 5% do resultado. Com isso, conseguimos extrair qual é a variância que influencia as estatísticas mais importantes para uma boa robustez do método, que são a mediana e a variabilidade dos dados.

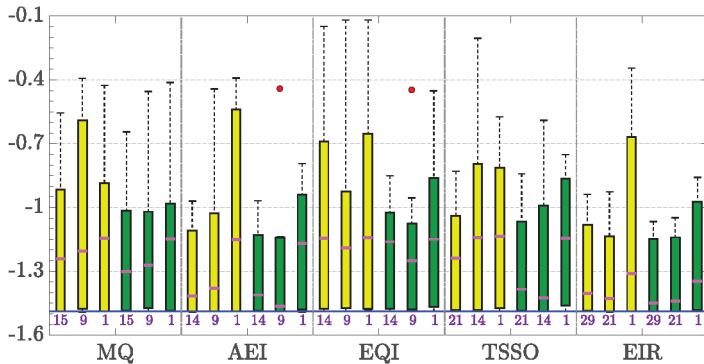
A última linha da Tabela 4 contém o resultado da métrica  $w$ . Podemos ver que, agora, os valores individuais de  $w$  somam 18. Portanto, podemos afirmar com mais significância que a variância de 0.0001 gerou os piores resultados, apresentando as maiores médias e variabilidade, enquanto a variância de 0.01 obteve os melhores resultados, com as menores médias e variabilidade. Essa foi a motivação de termos realizado a comparação do sEGO adaptativo com e sem normalização na seção anterior levando em consideração somente a variância inicial igual a 0.01.

Tabela 4 – Número dos piores resultados apresentados por cada variância inicial, entre todos os problemas.

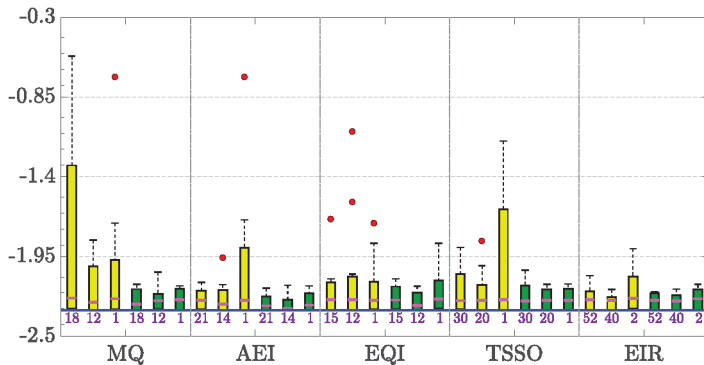
Estatísticas	Variâncias Iniciais		
	0.1	0.01	0.0001
Mediana	4	1	13
Desvio	2	4	12
Melhor Mínimo	5	5	8
Pior Mínimo	8	3	7
Percentil de 10%	4	6	8
Percentil de 90%	1	3	14
$w$	4.1	2.6	11.3



(a) F1 - 80 NFE

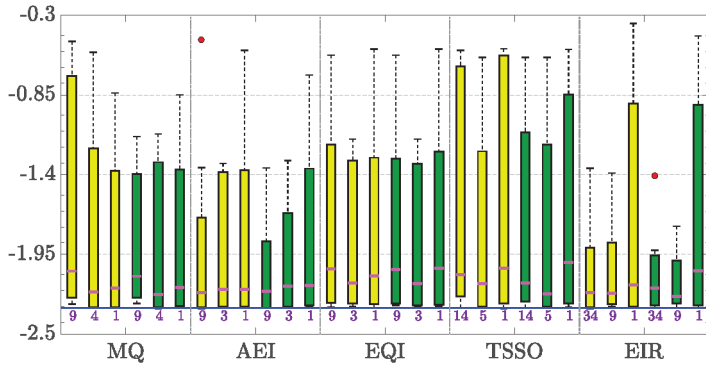


(b) F2 - 80 NFE

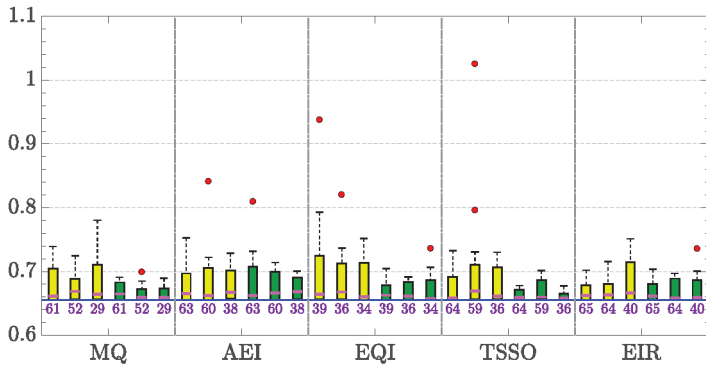


(c) F3 - 150 NFE

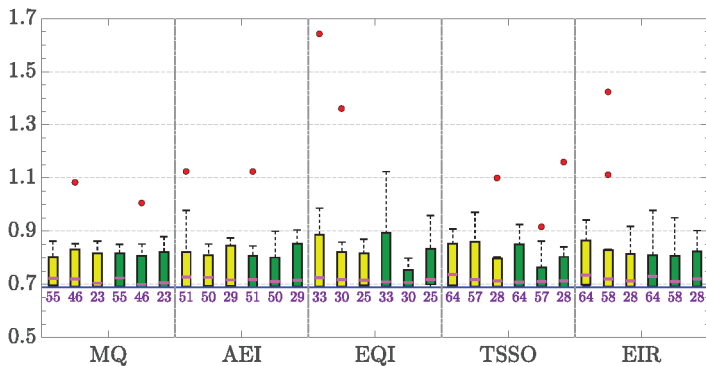
Figura 32 Influência de  $\sigma_0^2$  sobre o SEGO adaptativo com normalização.



(d) F4 - 150 NFE

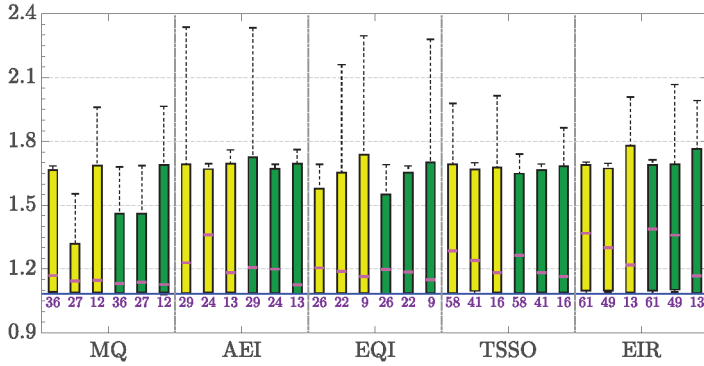


(e) F5 - 150 NFE

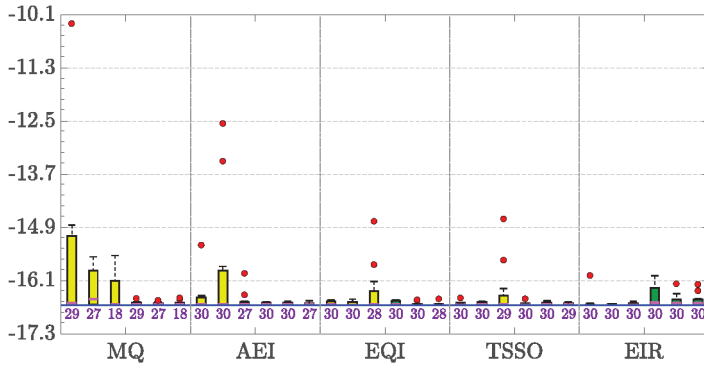


(f) F6 - 150 NFE

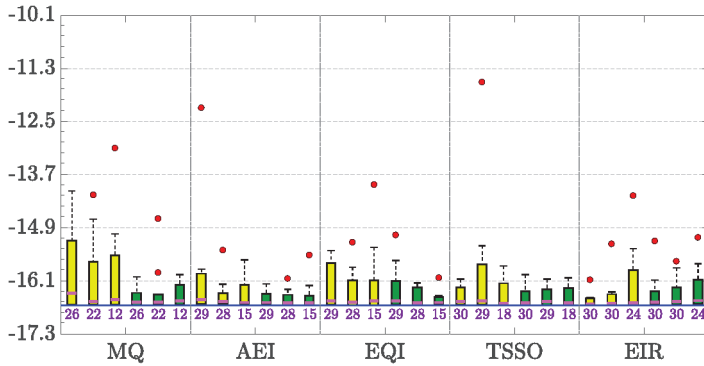
Figura 32 (Continuação) Influência de  $\sigma_0^2$  sobre o sEGO adaptativo com normalização.



(g) F7 - 150 NFE

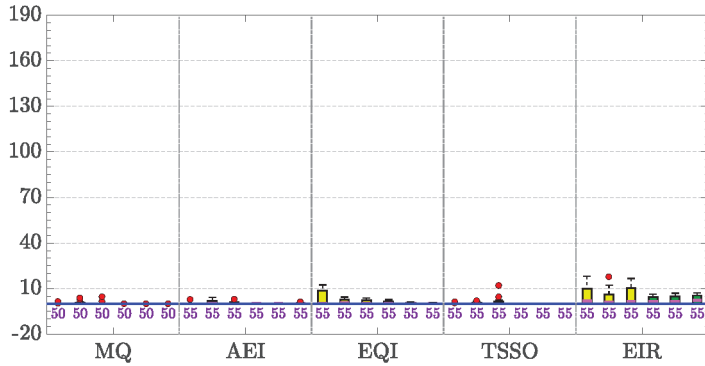


(h) F8 - 100 NFE

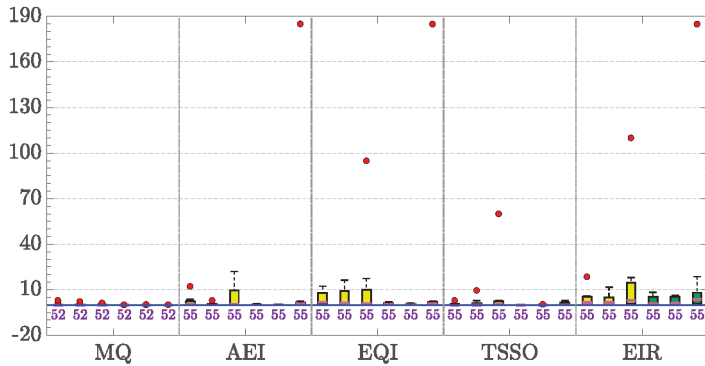


(i) F9 - 100 NFE

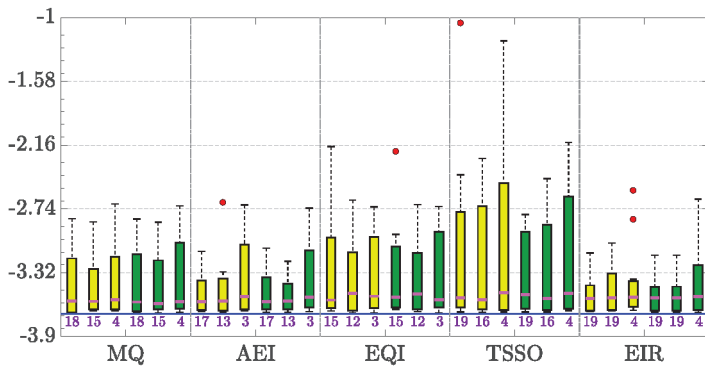
Figura 32 (Continuação) Influência de  $\bar{\sigma}_0^2$  sobre o sEGO adaptativo com normalização.



(j) F10 - 150 NFE

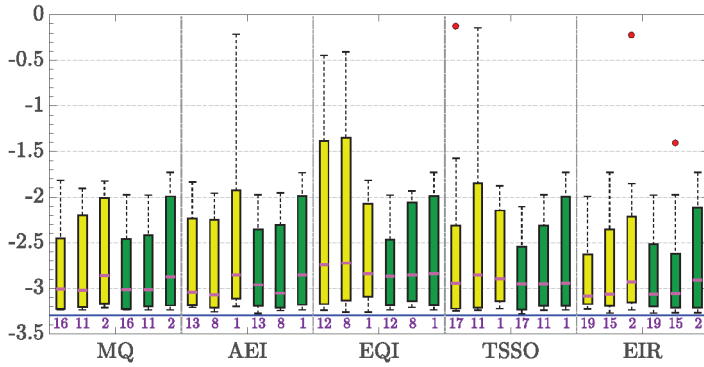


(k) F11 - 150 NFE

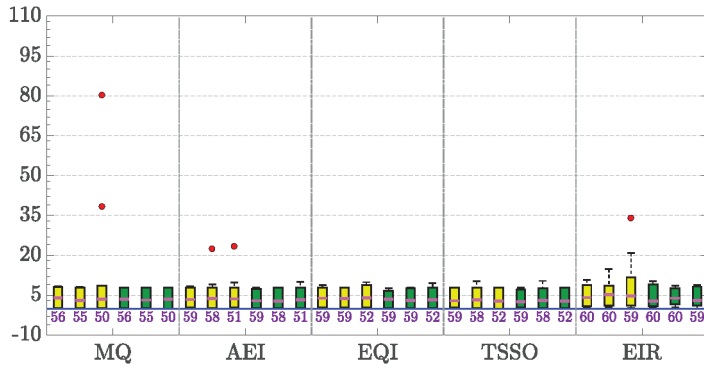


(l) F12 - 100 NFE

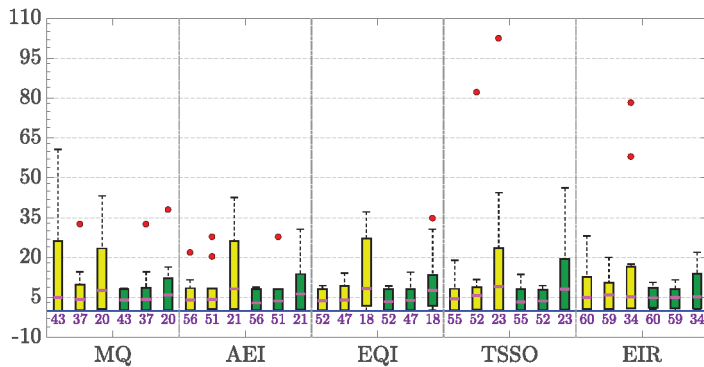
Figura 32 (Continuação) Influência de  $\sigma_0^2$  sobre o sEGO adaptativo com normalização.



(m) F13 - 150 NFE

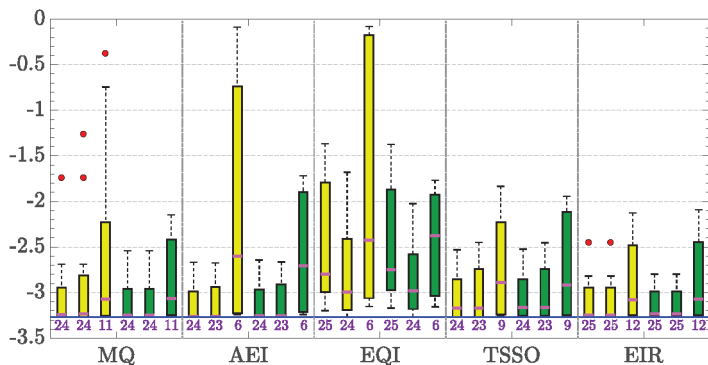


(n) F14 - 200 NFE

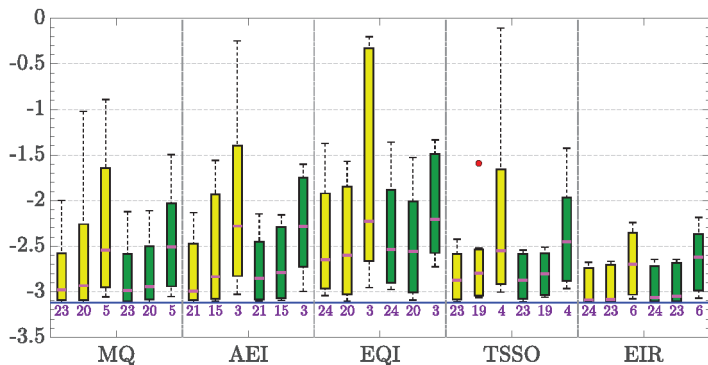


(o) F15 - 200 NFE

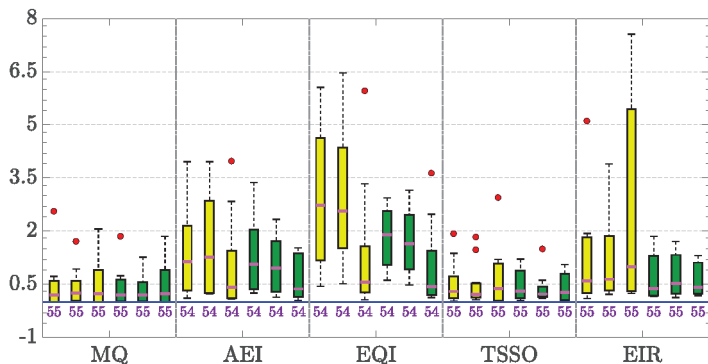
Figura 32 (Continuação) Influência de  $\sigma_0^2$  sobre o sEGO adaptativo com normalização.



(p) F16 - 240 NFE

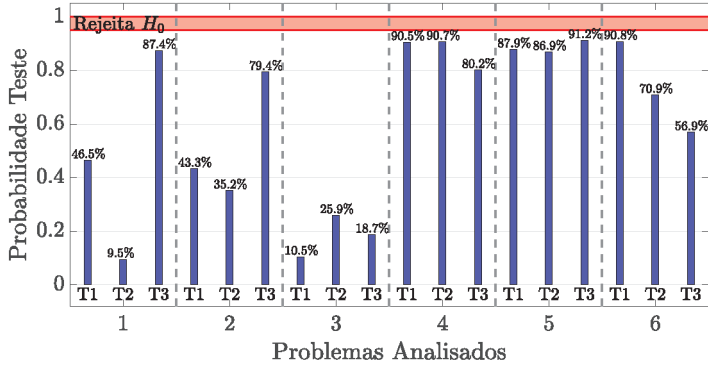


(q) F17 - 240 NFE

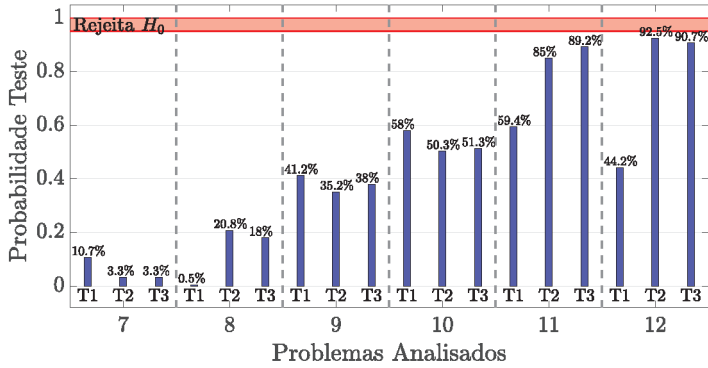


(r) F18 - 250 NFE

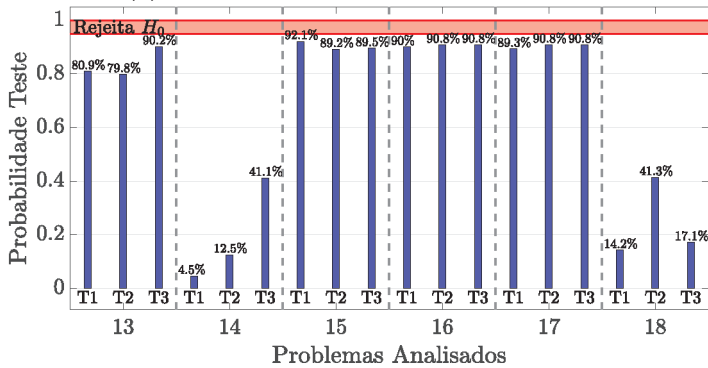
Figura 32 (Continuação) Influência de  $\sigma_0^2$  sobre o sEGO adaptativo com normalização.



(a) Resultado para os problemas F1 a F6



(b) Resultado para os problemas F7 a F12



(c) Resultado para os problemas F13 a F18

Figura 33 Testes estatísticos ao nível de 0.05 como confiança, para os resultados do sEGO adaptativo normalizado via MQ com P2.



## 5.5 DETERMINAÇÃO DA MÉTRICA COM OS MELHORES RESULTADOS OBTIDOS PELO sEGO ADAPTATIVO

Esta seção tem por intenção encontrar a métrica de adição de IPs que obtém os melhores resultados, entre todos os problemas analisados. Para extração dos resultados, fazemos a análise somente do sEGO adaptativo, utilizando a variância inicial  $\bar{\sigma}_0^2 = 0.01$ .

Nosso intuito é classificar qualitativamente as métricas, de acordo com as técnicas descritas nessa seção, levando em consideração a robustez dos resultados (menor variabilidade associada aos menores valores mínimos). Iremos realizar uma classificação das métricas, analisando o cenário formado pelos dezoito problemas propostos. Dessa forma, consideramos em cada uma das cinco métricas os valores mínimos obtidos pelo sEGO adaptativo com os minimizadores do tipo P1 e P2. Para melhor distinção entre as métricas, fazemos a junção da sigla da métrica com o número da sigla do tipo de minimizador. Logo, por exemplo, o resultado do sEGO adaptativo com a métrica AEI, cujo valor mínimo é obtido pelo minimizador P2, é apresentado com a notação AEI2.

Novamente, reiteramos que tal análise tem carácter qualitativo sobre os resultados obtidos, de modo que a métrica criada e utilizada seja capaz de ilustrar a tendência que determinada métrica possui em ser a melhor, ou a pior, escolha para o sEGO adaptativo.

### 5.5.1 Primeira análise: um olhar sobre a variabilidade dos resultados na perspectiva do percentil de 90%

A primeira técnica que utilizamos é a construção de um escore da variabilidade dos valores mínimos obtidos pelo sEGO adaptativo, dentro das trinta simulações realizadas. A variabilidade, nesse contexto, é medida pelo valor do percentil de 90% das soluções obtidas pelo sEGO adaptativo.

A construção dos escores se dará da seguinte forma: primeira-

mente associamos o escore 0 ao valor mínimo que é o objetivo do sEGO adaptativo, ou seja, os valores mínimos  $y_{\min}$  de cada uma das funções apresentadas na Seção 5.1. Em segundo lugar, associamos o escore 1 ao maior percentil de 90%, obtido por uma das cinco métricas, e com algum dos dois tipos de minimizadores. Chamaremos esse maior percentil de  $p90_{\max}$ . Por fim, cada percentil de 90% das demais métricas terá um escore resultante de uma interpolação linear entre  $y_{\min}$  e  $p90_{\max}$ . A relação linear para o escore será dada por

$$\text{escore}(p90) = \frac{p90 - y_{\min}}{p90_{\max} - y_{\min}}, \quad (5.13)$$

onde  $p90$  é o valor do percentil de 90% no qual queremos obter o escore.

Somente a título de ilustração, no problema F1 temos que  $y_{\min} = -1.489072$  assumirá o escore 0. A métrica MQ1 é a que resulta no sEGO adaptativo com o maior percentil de 90%, logo  $p90_{\max} = -0.747818$ . Assim, por exemplo, a métrica EIR2 com o percentil  $p90 = -1.318625$  possui escore de 0.2299, enquanto a métrica EQI1 com o percentil  $p90 = -1.028611$  tem escore de 0.6212.

Cada uma das métricas tem seus escores computados em cada um dos dezoito problemas. O escore final de uma determinada métrica, é a soma de todos os seus escores. A métrica que detiver o menor escore acumulado será considerado aquela mais robusta para o sEGO adaptativo.

A Figura 34 apresenta a evolução do escore de cada métrica ao longo dos dezoito problemas analisados. Na legenda, as métricas aparecem na ordem da pior para a melhor métrica. Podemos ver que o sEGO adaptativo, utilizando métrica MQ, e com os valores mínimos obtidos pelo minimizador do tipo P2 apresenta a melhor robustez perante as demais. O pior comportamento é apresentado pelo sEGO adaptativo com a métrica EQI1.

Pode-se notar o domínio da qualidade dos resultados do sEGO adaptativo obtidos com o minimizador do tipo P2 perante o minimizador do tipo P1, onde somente a métrica EIR1 possui um escore menor do que

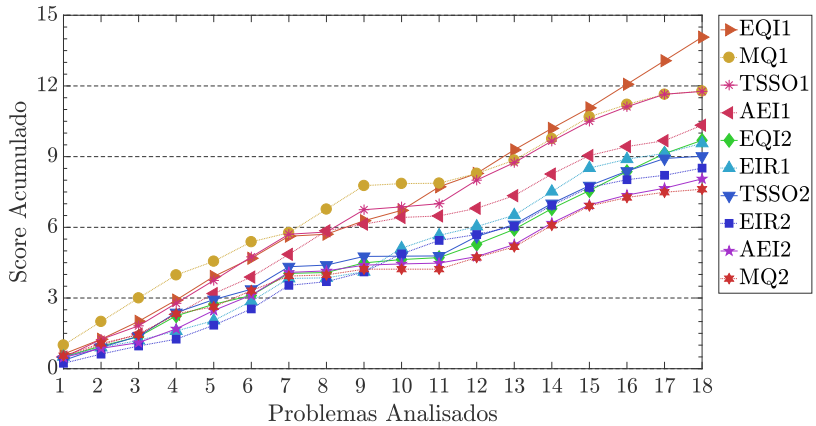


Figura 34 – Escores acumulados por cada métrica ao longo da análise dos dezoito problemas propostos.

o EQI2. Na verdade, o que se pode concluir é que o valor mínimo obtido pelo minimizador do tipo P1 é muito sensível ao modelo SK construído ao final da execução do sEGO adaptativo. Portanto, se o modelo SK não conseguir reproduzir com sucesso a região do mínimo global, o SGA não conseguirá encontrar o valor mínimo com boa qualidade. Por isso vemos a superioridade dos resultados considerando os valores obtidos com o minimizador do tipo P2, que depende somente de um percentil do valor do SK naquela região. Entretanto, o fato de somente o EIR1 ter sido superior a uma métrica com o minimizador do tipo P2, nos comprova efetivamente a superioridade do EIR em pesquisar o domínio de busca do problema. O que faz o EIR2 não ser superior ao AEI2 e nem ao MQ2, é o poder de refinamento dessas duas últimas métricas. Como estamos construindo um escore onde analisamos o maior valor de um percentil, então quanto menor for esse percentil, melhor será a métrica, e isso é o condicionante para a posição do EIR perante as demais. Porém, podemos verificar que o EIR1 é superior ao MQ1 e ao AEI1.

Cabe ressaltar aqui que a métrica MQ2 não é superior a todas

as demais em todos os problemas. Na verdade, podemos observar que a superioridade do MQ só começou a ser significativa a partir do problema F10. Antes desse problema, podemos ver que a métrica EIR2 dominou, mostrando a menor variabilidade perante as demais métricas. Esse fato é perceptível pois para problemas em grandes dimensões, o EIR gasta muitos pontos percorrendo todo o espaço viável, deixando a pesquisa local para o futuro, enquanto o MQ parte diretamente para o refino de uma região que contenha o menor percentil do SK.

A Tabela 8 contida no Apêndice D mostra todos os escores acumulados obtidos pelas cinco métricas, cada uma com os dois minimizadores e para todos os dezoito problemas.

### 5.5.2 Segunda análise: um olhar sobre a significância de seis estatísticas dos resultados

Para uma segunda análise das métricas, utilizamos uma abordagem semelhante à realizada na Subseção 5.4.2. Na referida subseção, fizemos uma análise das estatísticas mediana, desvio padrão, melhor valor mínimo, pior valor mínimo, percentil de 10% e percentil de 90%, no contexto do pior desempenho do sEGO adaptativo sobre uma determinada variância inicial. Aqui nessa subseção, fazemos praticamente a mesma análise, porém sobre a ótica de encontrar a melhor métrica de adição de IPs para o sEGO adaptativo.

A coleta dos valores é realizada de forma análoga a realizada na Subseção 5.4.2. Para cada problema, procuramos a métrica que possui a melhor estatística entre as métricas, considerando os dois tipos de minimizadores. Aqui, nossa condição de melhor valor é dada por aquela métrica que resultar no menor valor da estatística em questão. À métrica que possuir a melhor estatística será atribuído o valor 1, e faremos isso para cada um dos problemas. Ao fim, o somatório dos valores de cada métrica, em cada estatística, deve somar 18.

A Tabela 5 apresenta os valores assumidos por cada uma das métricas analisadas. Dessa forma, podemos ver como cada métrica se

sai, de acordo com uma das estatísticas. Como destaques positivos temos que a métrica MQ2 possui o maior número de melhores médias enquanto o EIR2 possui o maior número de menores desvios padrões. Isso comprova, respectivamente, o poder refinador do MQ e o poder exploratório do EIR. Diferentemente da subseção anterior, a métrica que apresenta os piores resultados estatísticos é o TSSO2, onde somente em três problemas ele foi destaque em alguma métrica.

Novamente, podemos somar todos os valores recebidos por uma determinada métrica. O resultado se encontra na penúltima linha da Tabela 5. Olhando somente para essa soma, poderíamos inferir que a métrica MQ2 é a melhor entre todas, porém as demais métricas não poderiam ser ordenadas, já que existem somas iguais para métricas diferentes. Além disso, o fato da soma não ser igual a 18 atrapalha a perspectiva geral de cada métrica do sEGO adaptativo. Dessa forma recorreremos ao valor  $w$  sugerido em (5.12) para ponderar as métricas de acordo com a melhor mediana associada a menor variabilidade dos dados. O resultado de  $w$  encontra-se na última linha da Tabela 5. Seguindo essa ponderação e ordenando da melhor para a pior métrica, temos

MQ2, AEI2, EIR2, EIR1, MQ1, AEI1, TSSO1, EQI1, EQI2, TSSO2.

Podemos ver que essa ordenação não é idêntica a apresentada pela Figura 34. Como  $w$  tem 50% do seu valor assumido pela mediana

Tabela 5 – Número de melhores resultados para cada métrica entre todos os problemas.

Estatísticas	Métricas									
	MQ1	MQ2	AEI1	AEI2	EQI1	EQI2	TSSO1	TSSO2	EIR1	EIR2
Mediana	1	5	3	4	0	0	0	0	3	2
Desvio	1	4	0	1	0	1	0	1	2	8
Melhor Mínimo	6	4	1	2	1	0	3	0	1	0
Pior Mínimo	4	0	2	0	5	0	5	0	2	0
Percentil 10%	4	5	1	2	2	1	0	0	2	1
Percentil 90%	1	3	0	3	0	2	0	2	1	6
<b>Soma</b>	<b>17</b>	<b>21</b>	<b>7</b>	<b>12</b>	<b>8</b>	<b>4</b>	<b>8</b>	<b>3</b>	<b>11</b>	<b>17</b>
$w$	2.2	4.1	1.9	2.7	0.75	0.3	0.95	0.25	2.3	2.55

e 15% retido pelo desvio e melhor mínimo, nessa nova disposição, as métricas que possuem os melhores resultados de valor mínimo, tendem a aparecer como melhores. Por isso vemos a dominância das métricas MQ, AEI e EIR na solução dos problemas.

Novamente temos as claras diferenças entre as métricas exploratórias (EIR) e as de refinamento (MQ e AEI). Como  $w$  é altamente sensível à mediana, métricas que refinam a região ótima e encontram um valor mínimo mais próximo do objetivo, tendem a ser superiores às métricas que possuem mais o viés de explorador. Porém, mesmo com esse comportamento, a métrica EIR ainda tende a gerar excelentes resultados, associando valor médio e variabilidade.

## 6 CONCLUSÃO E TRABALHOS FUTUROS

### 6.1 CONCLUSÃO

Nesta dissertação abordamos a solução de problemas de minimização, cujas funções objetivo são definidas em formas integrais. Essas formas integrais normalmente não possuem tratamento analítico, por isso devemos recorrer a solução por técnicas numéricas. Escolhendo a Integração por Monte Carlo (*Monte Carlo Integration* - MCI) como técnica numérica, aproximamos a função por meio de uma amostra aleatória de valores da função. Portanto, tratamos problemas de minimização cujas funções possuem incertezas ou ruídos advindos da aproximação do MCI, e que geralmente são problemas de alto custo computacional.

Dentro desse contexto, utilizamos um método eficaz de substituição de funções, conhecido como, Kriging Estocástico (*Stochastic Kriging* - SK). Esse metamodelo é construído a partir da suposição de duas parcelas de erro associadas à aproximação da função. A parcela Extrínseca do erro é obtida da suposição de que existe uma correlação espacial entre todos os valores da função em um espaço amostral. Já a parcela Intrínseca do erro é advinda da própria natureza estocástica da função com incertezas. Mostramos com sucesso, que a parcela intrínseca para o nosso problema de minimização, pode ser definida por meio de uma matriz de covariância, onde a diagonal principal é formada a partir da variância da integração via Monte Carlo.

Esse modelo foi então acoplado ao já conhecido EGO, e dessa forma ficou instaurado o método sEGO para solução de problemas não convexos de otimização, com funções objetivo sujeitas a incertezas, e caras computacionalmente. Dentro das métricas de adição dos chamados Pontos de Preenchimento (*Infill Point* - IP), escolhemos cinco técnicas para análise: Critério do Percentil Mínimo (*Minimal Quantile Criteria* - MQ), Melhora Esperada Aumentada (*Augmented Expected Improvement* - AEI), Percentil de Melhora Esperada (*Expected Quantile Improvement* - EQI), Otimização Sequencial de Dois Estágios (*Two-Stage Sequential*

*Optimization* - TSSO) e Melhora Esperada com Reinterpolação (*Expected Improvement with Reinterpolation* - EIR). Todas as cinco técnicas passaram por um ajuste, de modo a serem aplicadas no contexto de ruídos heterogêneos. Uma contribuição deste trabalho é a extensão do EIR para o caso heterogêneo. Outra contribuição se encontra na intensa análise geométrica e analítica sobre as diferentes métricas, de forma a extrairmos o comportamento do sEGO nas abordagens adaptativa e normalizada. Fomos capazes de observar que as métricas MQ, AEI e EQI possuem mais o viés de refinador, permitindo que pontos de suporte sejam considerados como IP, ou mesmo adicionando vários pontos em uma mesma região, para diminuição da incerteza nestas localidades. A métrica AEI é capaz de fugir da zona de refinamento graças a parcela penalizadora de sua métrica. A métrica TSSO mostrou um excelente balanço entre exploração local e global, não permitindo que pontos de suporte fossem considerados como novos IPs. Por fim, a métrica EIR se mostra como a mais robusta no quesito exploração global, não permitindo que pontos de suporte sejam novos IPs, conseguindo assim, varrer todo o domínio de busca da função, conseqüentemente diminuindo a incerteza da parcela extrínseca do metamodelo SK.

Para auxiliar o sEGO na alocação do recurso computacional, foi utilizada a abordagem adaptativa da variância alvo para o MCI. Nessa técnica, propomos um valor para a variância do MCI e fazemos replicações do valor da função até que atinjamos tal variância alvo. A abordagem adaptativa é baseada no mesmo conceito da Melhora Esperada (*Expected Improvement* - EI), prezando por um balanceamento entre a pesquisa global e a pesquisa local dos pontos do domínio de busca. Dessa forma, o valor da variância adaptativa é construída como uma expressão exponencial, parametrizada pelo número de dimensões do problema, bem como pelo número de pontos próximos ao ponto em que desejamos aproximar o valor da função objetivo. No início da execução do sEGO adaptativo, devemos ajustar um valor para a variância alvo inicial.



Vimos que um fator importante para o sucesso do sEGO adaptativo é a amplitude do contradomínio, representando pelo valor CD como a diferença entre o maior e menor valor da função objetivo. Funções com CD alto fazem com que o MCI demore para convergir para a variância alvo proposta, fazendo todo o orçamento computacional ser gasto em poucas ou em uma única adição de IP. Por isso, uma das grandes contribuições deste trabalho é a aplicação de uma normalização na função objetivo, antes de construirmos o modelo em SK. Essa normalização foi proposta por uma transformação não linear conhecida como Tunelamento Estocástico. Vimos que o tunelamento estocástico não altera as tendências de superfície da função objetivo, e também não altera a variabilidade estocástica da função. Logo, a função da normalização está em apenas mapear o contradomínio da função original para o domínio  $[0,1]$ . Outro fator associado à normalização, e estudado neste trabalho, é a sensibilidade do sEGO adaptativo ao valor da variância alvo inicial. Assim, foram analisadas três tipos de variâncias iniciais com os valores de 0.1, 0.01 e 0.0001 para os casos onde o sEGO foi executado com a normalização. Vimos como estas variâncias influenciam o resultado obtido pelo sEGO analisando, além dos valores mínimos obtidos, a quantidade de IPs adicionados com o orçamento computacional disponível.

Para podermos conferir a viabilidade de todas as modificações e adaptações propostas neste trabalho para o sEGO adaptativo, utilizamos nove funções conhecidas na literatura de otimização. Cada uma dessas funções possui parâmetros estocásticos que pertencem a uma determinada distribuição de probabilidade. Logo, para diferentes níveis de variabilidade, analisamos um conjunto de dezoito problemas de minimização. A forma escolhida para se mostrar os resultados foi realizando o total de 30 simulações do sEGO para cada problema, e em cada métrica. Foram definidos duas formas de extração do minimizador: o tipo P1 é obtido da utilização do SGA na otimização do último modelo SK ajustado pelo sEGO. O minimizador do tipo P2, é o ponto pertencente ao último espaço amostral formado durante a execução do sEGO, e que minimiza um percentil de 70% do SK criado a partir desse plano

amostral.

Após todas as execuções, fomos capazes de inferir as seguintes conclusões:

- O sEGO utilizando a normalização possui um comportamento muito superior ao sEGO sem a normalização. A principal diferença está na quantidade de IPs que todas as cinco métricas são capazes de adicionar com a normalização. Essa superioridade do valor de IPs é traduzida ou em um melhor refinamento do valor mínimo ou em uma exploração maior do domínio de busca. Em alguns problemas mais multimodais, somente a normalização foi capaz de aproximar com qualidade o valor objetivo, além de mostrar as menores variabilidade dos resultados;
- A normalização também agregou ao sEGO uma resistência maior ao valor da variância inicial. Foi possível mostrar por testes estatísticos que os três níveis de variância utilizados neste trabalho geraram resultados que podem ser considerados idênticos e que possuem a mesma distribuição de probabilidade. Dessa forma, vemos que não teremos tanto impacto no resultado do sEGO a partir da variância inicial. Porém, como visto no texto, não deve-se ajustar qualquer valor à variância inicial; logo, fizemos uma análise em seis estatísticas que traduzem a qualidade das aproximações, e dessa análise concluímos que a variância inicial de 0.01 foi superior às demais. Portanto, esse é um excelente ponto de partida para a utilização da pesquisa adaptativa;
- Também conseguimos averiguar que dos dois tipos de minimizadores propostos, o minimizador do tipo P2 se saiu melhor do que o minimizador do tipo P1. O minimizador P1 necessita de uma excelente substituição por parte do metamodelo SK na região do mínimo; logo, esse ponto pode falhar caso essa região não seja bem modelada. De uma outra perspectiva, o minimizador P2 não

precisa dessa condição, bastando apenas possuir na região um valor médio baixo do SK, associado a um pequeno erro RMSE;

- Utilizamos de duas técnicas diferentes para encontrarmos qual é a melhor métrica de adição de IPs para o sEGO. A primeira técnica foi uma análise da variabilidade dos dados, a partir do valor do percentil de 90% dos trinta resultados obtidos. Vemos que, em ordem, os três melhores algoritmos são o MQ, AEI e o EIR, utilizando o minimizador do tipo P2. Os dois primeiros são considerados como extremamente locais, enquanto o EIR é o mais explorador. O pior algoritmo nesta técnica foi o EQI, que obteve os piores desempenhos, considerando os dois tipos de minimizadores. A segunda técnica utilizada foi uma ponderação da quantidade de melhores estatísticas obtidas por cada métrica, entre os dezoito problemas. Entre os valores estatísticos estão a mediana e o desvio padrão. Dessa forma, é encontrado o algoritmo que possui a melhor mediana e apresenta a menor variabilidade dos dados. Novamente, as três melhores métricas foram MQ, AEI e o EIR, utilizando o minimizador P2. A pior métrica nessa abordagem foi o TSSO2;
- Da conclusão anterior podemos elucidar que, quando trabalhamos com problemas convexos, ou com superfícies não tão multimodais, a métrica MQ deverá ser utilizada, pois trabalha muito melhor a condição de refinador e não se preocupa em obter o melhor modelo SK. Entretanto, para problemas mais multimodais, dos quais precisamos ter pontos em diversas localidades do domínio para que possamos ter um SK com menos incerteza, então deverá ser utilizada a métrica EIR.

Finalmente todos os resultados mostraram que o sEGO é uma excelente opção de algoritmo global de otimização, sendo extremamente valioso na substituição de funções de difícil tratamento computacional. Todas as métricas foram capazes de obter valores mínimos e minimi-

zadores com extrema qualidade, quando associadas à normalização proposta.

## 6.2 TRABALHOS FUTUROS

Tendo em vista todas as abordagens realizadas neste estudo, propõem-se para trabalhos futuros os seguintes pontos:

- Criar uma nova métrica de adição de IPs que seja uma combinação das métricas MQ e EIR;
- Definir uma alternativa para se obter a matriz de covariância para o erro intrínseco do SK;
- Construir uma maneira alternativa de se agregar a parcela intrínseca à parcela extrínseca do SK, prezando pela simplificação da função de verossimilhança e de sua otimização;
- Ampliar o estudo dos impactos gerados no SK com a utilização da normalização, principalmente no comportamento do RMSE;
- Identificar novas alternativas à pesquisa adaptativa, prezando por realizar mais replicações do valor da função, sem aumentar o orçamento computacional;
- Verificar a eficiência de outras funções de correlação para a criação da parcela extrínseca do sEGO;
- Estender as técnicas de construção do SK, bem como do sEGO, a problemas com restrições.

## Referências

- ANDERSON, T. W. *An introduction to multivariate statistical analysis*. [S.l.]: Wiley-Interscience, 2003. ISBN 0-471-36091-0. Citado na página [201](#).
- ANKENMAN, B.; NELSON, B. L.; STAUM, J. Stochastic kriging for simulation metamodeling. *Operations Research, Informatics*, v. 58, n. 02, p. 371–382, mar. 2010. ISSN 0030-364X. Citado 8 vezes nas páginas [66](#), [87](#), [88](#), [93](#), [96](#), [97](#), [98](#) e [99](#).
- ARORA, J. S. *Introduction to optimum design*. 4. ed. San Diego, USA: Elsevier Academic Press, 2017. ISBN 9780128008065. Citado 5 vezes nas páginas [41](#), [42](#), [44](#), [47](#) e [49](#).
- BAZARAA, M. S.; SHERALI, H. D.; SHETTY, C. M. *Nonlinear Programming: Theory and Algorithms*. 3. ed. New Jersey, USA: John Wiley & Sons, Inc, 2006. ISBN 0471486000. Citado 2 vezes nas páginas [43](#) e [47](#).
- BECK, A. T.; KOUGIOUMTZOGLOU, I. A.; SANTOS, K. R. M. dos. Optimal performance-based design of non-linear stochastic dynamical structures subject to stationary wind excitation. *Engineering Structures*, v. 78, p. 145–153, 2014. Citado na página [31](#).
- BECK, J.; DIA, B. M.; ESPATH, L. F. R.; LONG, Q.; TEMPONE, R. Fast bayesian experimental design: Laplace-based importance sampling for the expected information gain. *Computer Methods in Applied Mechanics and Engineering*, v. 334, p. 523–553, jun. 2018. Citado na página [32](#).
- BEERS, W. C. M. van; KLEIJNEN, J. P. C. Kriging for interpolation in random simulation. *Journal of the Operational Research Society*, v. 54, p. 255–262, 2003. Citado 3 vezes nas páginas [66](#), [88](#) e [92](#).
- BHATTACHARJYA, S.; CHAKRABORTY, S. An improved robust multi-objective optimization of structure with random parameters. *Advances in Structural Engineering*, v. 21, n. 11, p. 1597–1607, 2018. Citado na página [31](#).
- BHATTI, M. A. *Practical optimization methods: with mathematica® applications*. [S.l.]: Springer Science & Business Media, 2012. Citado 2 vezes nas páginas [41](#) e [47](#).

- BIERMANN, D.; WEINERT, K.; WAGNER, T. Model-based optimization revisited: Towards real-world processes. In: *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*. [S.l.: s.n.], 2008. p. 2975–2982. ISSN 1089-778X. Citado na página 87.
- BOBBY, S.; SPENCE, S. M. J.; KAREEM, A. Data-driven performance-based topology optimization of uncertain wind-excited tall buildings. *Structural and Multidisciplinary Optimization*, v. 54, n. 6, p. 1379–1402, dez. 2016. Citado 2 vezes nas páginas 31 e 32.
- BROOMHEAD, D.; LOWE, D. Multivariable functional interpolation and adaptive networks. *Complex Systems*, v. 2, p. 321–355, 1988. Citado na página 52.
- CAFLISCH, R. E. Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica*, Cambridge University Press, v. 7, p. 1–49, 1998. Citado na página 88.
- CAPIEZ-LERNOUT, E.; SOIZE, C. Robust design optimization in computational mechanics. *Journal of Applied Mechanics-Transactions of the Asme*, v. 75, n. 2, p. Article Number: 021001, 2008. Disponível em: <https://hal-upec-upem.archives-ouvertes.fr/hal-00686134>. Citado na página 30.
- CARLON, A. G.; DIA, B. M.; ESPATH, L. F. R.; LOPEZ, R. H.; TEMPONE, R. Nesterov-Aided Stochastic Gradient Methods Using Laplace Approximation for Bayesian Design Optimization. Preprint submitted to Elsevier. 2019. Disponível em: <https://arxiv.org/pdf/1807.00653.pdf>. Citado na página 32.
- CARRARO, F. *A stochastic Kriging approach for the minimization of integrals*. Dissertação (Mestrado) — Departamento de Engenharia Civil, Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina, Brasil, 2017. Citado 6 vezes nas páginas 65, 75, 107, 116, 163 e 165.
- CARRARO, F.; LOPEZ, R. H.; MIGUEL, L. F. F.; TORII, A. J. Monte carlo integration with adaptive variance selection for improved stochastic efficient global optimization. *Structural and Multidisciplinary Optimization*, Feb 2019. ISSN 1615-1488. Disponível em: <https://doi.org/10.1007/s00158-019-02212-y>. Citado 9 vezes nas páginas 34, 35, 37, 109, 114, 115, 140, 153 e 163.

CHAUDHURI, A.; HAFTKA, R. Efficient global optimization with adaptive target setting. *AIAA Journal, American Institute of Aeronautics and Astronautics*, v. 52, p. 1573–1578, 07 2014. Citado na página 66.

CHEN, X.; KIM, K.-K. Stochastic kriging with biased sample estimates. *ACM Transactions on Modeling and Computer Simulation*, ACM, New York, NY, USA, v. 24, n. 2, p. 8:1–8:23, fev. 2014. ISSN 1049-3301. Disponível em: <http://doi.acm.org/10.1145/2567893>. Citado na página 88.

CONOVER, W. J. *Practical Nonparametric Statistics*. Third. [S.l.]: John Wiley & Sons, 1999. ISBN 0471160687. Citado na página 163.

COX, D. D.; JOHN, S. Sdo: A statistical method for global optimization. In: *in Multidisciplinary Design Optimization: State-of-the-Art*. [S.l.: s.n.], 1997. p. 315–329. Citado na página 104.

CRISTIANINI, N.; SHAWE-TAYLOR, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. [S.l.]: Cambridge University Press, 2000. ISBN 978-0521780193. Citado na página 52.

DORIGO, M. *Optimization, learning and natural algorithms*. Tese (Doutorado) — Politecnico di Milano, Italy, 1992. Citado na página 49.

EIDORS. *Electrical Impedance Tomography and Diffuse Optical Tomography Reconstruction Software*. 2019. Acessado em 04/04/2019. Disponível em: [http://eidors3d.sourceforge.net/tutorial/EIDORS\\_basics/contrasts.shtml](http://eidors3d.sourceforge.net/tutorial/EIDORS_basics/contrasts.shtml). Citado na página 33.

EOM, Y.-S.; YOO, K.-S.; PARK, J.-Y.; HAN, S.-Y. Reliability-based topology optimization using a standard response surface method for three-dimensional structures. *Structural and Multidisciplinary Optimization*, v. 43, n. 2, p. 287–295, 2011. Citado na página 52.

FLETCHER, R. *Practical Methods of Optimization*. Second. [S.l.]: Wiley, 1987. ISBN 0471494631. Citado 2 vezes nas páginas 41 e 47.

FORRESTER, A.; SOBESTER, A. S.; KEANE, A. *Engineering design via surrogate modelling: a practical guide*. Chichester, West Sussex, United Kingdom: John Wiley & Sons, 2008. ISBN 978-0-470-06068-1. Citado 16 vezes nas páginas 52, 57, 58, 60, 61, 63, 65, 66, 72, 74, 77, 84, 87, 91, 93 e 120.

FORRESTER, A. I.; KEANE, A. J. Recent advances in surrogate-based optimization. *Progress in Aerospace Sciences*, v. 45, n. 1, p. 50–79, 2009. ISSN 0376-0421. Citado 2 vezes nas páginas 61 e 66.

FORRESTER, A. I. J.; KEANE, A. J.; BRESSLOFF, N. W. Design and analysis of "noisy" computer experiments. *AIAA Journal*, v. 44, n. 10, p. 2331–2339, jan. 2006. Disponível em: (<https://doi.org/10.2514/1.20068>). Citado 9 vezes nas páginas 66, 88, 91, 93, 101, 102, 103, 111 e 129.

FORRESTER, A. I. J.; SÓBESTER, A.; KEANE, A. J. Multi-fidelity optimization via surrogate modelling. *Proceedings of the Royal Society of London A*, v. 463, n. 2088, p. 3251–3269, 2007. Disponível em: (<https://eprints.soton.ac.uk/64698/>). Citado na página 77.

GILES, M. B. Multilevel Monte Carlo Path Simulation. *OPERATIONS RESEARCH*, v. 56, n. 3, p. 607–617, jun. 2008. ISSN 0030-364X. Citado na página 88.

GOLDBERG, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*. 1st. ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989. ISBN 0201157675. Citado na página 49.

GOMES, W. J.; BECK, A. T.; LOPEZ, R. H.; MIGUEL, L. F. A probabilistic metric for comparing metaheuristic optimization algorithms. *Structural Safety*, v. 70, p. 59 – 70, 2018. ISSN 0167-4730. Citado na página 154.

GONÇALVES, M. S.; LOPEZ, R. H.; MIGUEL, L. F. F. Search group algorithm: A new metaheuristic method for the optimization of truss structures. *Computers and Structures*, v. 153, p. 165–184, 2015. Citado 2 vezes nas páginas 49 e 155.

Haji-Ali, A.-L.; NOBILE, F.; TEMPONE, R. Multi-index monte carlo: when sparsity meets sampling. *Numerische Mathematik*, v. 132, n. 4, p. 767–806, 2016. Citado na página 88.

HAMMERSLEY, J. M.; HANDSCOMB, D. C. Monte carlo methods. *Springer*, 1964. Citado na página 89.

HOERL, A. E.; KENNARD, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, v. 12, n. 1, p. 55–67, 1970. Citado 2 vezes nas páginas 93 e 94.



- HOYLE, N. *Automated multi-stage geometry parameterization of internal fluid flow applications*. Tese (phdthesis) — University of Southampton, 2006. Disponível em: (<https://eprints.soton.ac.uk/72028/>). Citado 3 vezes nas páginas 51, 60 e 71.
- HUAN, X.; MARZOUK, Y. M. Simulation-based optimal bayesian experimental design for nonlinear systems. *Journal of Computational Physics*, v. 232, n. 1, p. 288–317, 2013. Citado na página 32.
- HUANG, D.; ALLEN, T. T.; NOTZ, W. I.; ZENG, N. Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of Global Optimization*, v. 34, p. 441–466, 2006. Citado 6 vezes nas páginas 88, 93, 94, 102, 103 e 129.
- HUSSAIN, M. F.; BARTON, R. R.; JOSHI, S. B. Metamodeling: Radial basis functions, versus polynomials. *European Journal of Operational Research*, v. 138, n. 1, p. 142–154, 2002. Citado na página 52.
- IOOSS, B.; LHUILLIER, C.; JEANNEAU, H. Numerical simulation of transit-time ultrasonic flowmeters: uncertainties due to flow profile and fluid turbulence. *Ultrasonics*, v. 40, n. 9, p. 1009–1015, 2002. ISSN 0041-624X. Disponível em: (<http://www.sciencedirect.com/science/article/pii/S0041624X02003876>). Citado na página 87.
- IZMAILOV, A.; SOLODOV, M. *Otimização: Métodos Computacionais*. Segunda. [S.l.]: Associação Instituto Nacional de Matemática Pura e Aplicada - IMPA, 2012. v. 2. ISBN 9788524402685. Citado na página 47.
- JALALI, H.; NIEUWENHUYSE, I. V.; PICHENY, V. Comparison of kriging-based algorithms for simulation optimization with heterogeneous noise. *European Journal of Operational Research*, v. 261, n. 1, p. 279–301, 2017. ISSN 0377-2217. Citado 9 vezes nas páginas 66, 84, 102, 103, 104, 114, 122, 129 e 155.
- JOHNSON, M. E.; MOORE, L. M.; YLVISAKER, D. Minimax and maximin distance designs. *Journal of statistical planning and inference, Elsevier*, v. 26, n. 2, p. 131–148, 1990. Citado na página 75.
- JONES, D. R. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, v. 21, n. 4, p. 345–383, 2001. Citado 8 vezes nas páginas 60, 61, 66, 71, 74, 75, 84 e 105.

JONES, D. R.; SCHONLAU, M.; WELCH, W. J. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, v. 13, n. 4, p. 455–492, dez. 1998. Citado 12 vezes nas páginas 34, 58, 61, 66, 74, 76, 77, 78, 84, 101, 102 e 152.

JR, C. W. C. Response surface methods for optimizing and improving reproducibility of crystal growth. In: *Macromolecular Crystallography Part A*. 1. ed. San Diego, USA: Academic Press, 1997. v. 276, p. 74–90. ISBN 978-0-12-182177-7. Citado na página 52.

KAGAN, N.; SCHMIDT, H. P.; OLIVEIRA, C. C. B. de; KAGAN, H. *Métodos de otimização aplicados a sistemas elétricos de potência*. 1. ed. [S.l.]: Editora Blucher, 2009. ISBN 978-85-212-0472-5. Citado na página 45.

KALOS, M. H.; WHITLOCK, P. A. *Monte Carlo Methods. Vol. 1: Basics*. New York, NY, USA: Wiley-Interscience, 1986. ISBN 0-471-89839-2. Citado na página 89.

KEANE, A. J.; NAIR, P. B. *Computational approaches for aerospace design: the pursuit of excellence*. Chichester, West Sussex, England: John Wiley & Sons, 2005. ISBN 9780470855485. Citado na página 93.

KENNEDY, J.; EBERHART, R. C. Particle swarm optimization. *IEEE International Conference on Neural Network*, IV, p. 1942–1948, 1995. Citado na página 49.

KIRKPATRICK, S.; GELATT, C. D.; VECCHI, M. P. Optimization by simulated annealing. *Science*, v. 220, n. 4598, p. 671–680, 1983. Citado na página 49.

KLEIJNEN, J. P. C.; MEHDAD, E. Estimating the variance of the predictor in stochastic kriging. *Center Discussion Paper*, n. 2015-041, 2015. Disponível em: <https://ssrn.com/abstract=2646459>. Citado na página 88.

KRIGE, D. G. A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Chemical, Metallurgical and Mining Engineering Society of South Africa*, v. 52, n. 6, p. 119–139, 1951. Citado na página 58.

LEPAGE, G. P. A new algorithm for adaptive multidimensional integration. *Journal of Computational Physics*, v. 27, n. 2, p. 192–203, 1978. ISSN 0021-9991. Disponível em: [http:](http://)

[//www.sciencedirect.com/science/article/pii/S0021999178900049](http://www.sciencedirect.com/science/article/pii/S0021999178900049)).

Citado na página 88.

LI, W.; HUYSE, L.; PADULA, S. Robust airfoil optimization to achieve drag reduction over a range of mach numbers. *Structural and Multidisciplinary Optimization*, v. 24, n. 1, p. 38–50, ago. 2002. ISSN 1615-1488. Disponível em: (<https://doi.org/10.1007/s00158-002-0212-4>). Citado na página 87.

LIU, B.; HAFTKA, R.; AKGÜN, M. Two-level composite wing structural optimization using response surfaces. *Structural and Multidisciplinary Optimization*, v. 20, n. 2, p. 87–96, 2000. ISSN 1615-1488. Citado na página 52.

LOCATELLI, M. Bayesian algorithms for one-dimensional global optimization. *Journal of Global Optimization*, Springer Nature, v. 10, n. 1, p. 57–76, 1997. Citado na página 77.

LOPEZ, R. H.; RITTO, T. G.; SAMPAIO, R.; CURSI, J. E. S. de. A new algorithm for the robust optimization of rotor-bearing systems. *Engineering Optimization*, v. 46, n. 8, p. 1123–1138, 2014. Citado na página 30.

LUENBERGER, D. G. *Optimization by Vector Space Methods*. New York, USA: Wiley, 1969. ISBN 9780471181170. Citado 2 vezes nas páginas 44 e 47.

MATHERON, G. Principles of geostatistics. *Economic Geology*, v. 58, n. 8, p. 1246–1266, 1963. Citado 2 vezes nas páginas 52 e 58.

MATHWORKS, I. T. *MatLab, version R2017a (9.2.0.538062)*. 2017. Citado 2 vezes nas páginas 130 e 139.

MIGUEL, L. F. F.; LOPEZ, R. H.; TORII, A. J.; MIGUEL, L. F. F.; BECK, A. T. Robust design optimization of tmds in vehicle–bridge coupled vibration problems. *Engineering Structures*, v. 126, p. 703–711, nov. 2016. ISSN 0141-0296. Citado na página 30.

MIGUEL, L. F. F.; MIGUEL, L. F. F.; LOPEZ, R. H. Failure probability minimization of buildings through passive friction dampers. *The Structural Design of Tall and Special Buildings*, v. 25, n. 17, p. 869–885, 2016. Citado na página 30.

MISS, J.; JACQUET, O.; HEULERS, L. The moret 4.b monte carlo code: New features to treat complex criticality systems. In: *Integrating*

*criticality safety into the resurgence of nuclear power*. Knoxville, TN: American Nuclear Society, 2005. p. 61. Citado na página 87.

MITCHELL, T. J.; MORRIS, M. D. The spatial correlation function approach to response surface estimation. In: *Proceedings of the 24th Conference on Winter Simulation*. New York, NY, USA: ACM, 1992. (WSC '92), p. 565–571. ISBN 0-7803-0798-4. Disponível em: <http://doi.acm.org/10.1145/167293.167638>. Citado na página 75.

MOCKUS, J. Application of bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization*, v. 4, n. 4, p. 347–365, Jun 1994. ISSN 1573-2916. Disponível em: <https://doi.org/10.1007/BF01099263>. Citado na página 81.

MOCKUS, J.; TIESIS, V.; ZILINSKAS, A. The application of bayesian methods for seeking the extremum. *Towards Global Optimization*, North Holland, Amsterdam, v. 2, p. 117–129, 09 1978. Citado na página 77.

NASCENTES, F. F. S.; LOPEZ, R. H.; CURSI, J. E. S.; SAMPAIO, R.; MIGUEL, L. F. F. An efficient global optimization approach for reliability maximization of friction-tuned mass damper-controlled structures. *Shock and Vibration*, v. 2018, p. 8, 2018. Citado na página 77.

NOCEDAL, J.; WRIGHT, S. J. *Numerical Optimization*. second. New York, NY, USA: Springer, 2006. Citado 2 vezes nas páginas 45 e 46.

PENROSE, R. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, Cambridge University Press, v. 51, n. 3, p. 406–413, 1955. Citado na página 55.

PICHENY, V.; GINSBOURGER, D. Noisy kriging-based optimization methods: A unified implementation within the diceoptim package. *Computational Statistics & Data Analysis*, v. 71, p. 1035–1053, 2014. ISSN 0167-9473. Citado 2 vezes nas páginas 102 e 155.

PICHENY, V.; GINSBOURGER, D.; RICHEL, Y.; CAPLIN, G. Quantile-based optimization of noisy computer experiments with tunable precision. *Technometrics*, v. 55, p. 2–13, 2013. Citado 6 vezes nas páginas 84, 94, 102, 108, 129 e 155.

PICHENY, V.; WAGNER, T.; GINSBOURGER, D. A benchmark of kriging-based infill criteria for noisy optimization. Working paper or reprint. 2012. Disponível em: <https://>

- [//hal.archives-ouvertes.fr/hal-00658212](https://hal.archives-ouvertes.fr/hal-00658212)). Citado 3 vezes nas páginas 88, 102 e 129.
- PICHENY, V.; WAGNER, T.; GINSBOURGER, D. A benchmark of kriging-based infill criteria for noisy optimization. *Structural and Multidisciplinary Optimization*, v. 48, n. 3, p. 607–626, set. 2013. ISSN 1615-1488. Citado 4 vezes nas páginas 66, 102, 103 e 105.
- PLUMLEE, M.; TUO, R. Building accurate emulators for stochastic simulations via quantile kriging. *Technometrics*, Taylor & Francis, v. 56, n. 4, p. 466–473, 2014. Citado na página 88.
- POWELL, M. J. D. Radial basis functions for multivariable interpolation: A review. In: MASON, J. C.; COX, M. G. (Ed.). *Algorithms for Approximation*. New York, NY, USA: Clarendon Press, 1987. p. 143–167. ISBN 0-19-853612-7. Citado na página 52.
- PRONZATO, L.; MÜLLER, W. Design of computer experiments: space filling and beyond. *Statistics and Computing*, Springer, v. 22, n. 3, p. 681–701, 2012. Citado na página 75.
- QUAN, N.; YIN, J.; NG, S. H.; LEE, L. H. Simulation optimization via kriging: a sequential search using expected improvement with computing budget constraints. *IIE Transactions*, Taylor & Francis, v. 45, n. 7, p. 763–780, abr. 2013. Citado 3 vezes nas páginas 103, 109 e 129.
- RITTO, T. G.; LOPEZ, R. H.; SAMPAIO, R.; CURSI, J. E. S. de. Robust optimization of a flexible rotor-bearing system using the campbell diagram. *Engineering Optimization*, Taylor & Francis, v. 43, n. 1, p. 77–96, 2011. Citado na página 30.
- ROUSTANT, O.; GINSBOURGER, D.; DEVILLE, Y. DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. Working paper or preprint. 2010. Disponível em: (<https://hal.archives-ouvertes.fr/hal-00495766>). Citado na página 139.
- RUBINSTEIN, R. Y.; KROESE, D. P. *Simulation and the Monte Carlo Method*. 3st. ed. New York, NY, USA: John Wiley & Sons, Inc., 2017. ISBN 9781118632208. Citado na página 88.
- SACKS, J.; WELCH, W. J.; MITCHELL, T. J.; WYNN, H. P. Design and analysis of computer experiments. *Statistical Science*, v. 4, n. 4, p. 409–423, 1989. Citado 3 vezes nas páginas 58, 66 e 71.

SAKATA, S.; ASHIDA., F. Ns-kriging based microstructural optimization applied to minimizing stochastic variation of homogenized elasticity of fiber reinforced composites. *Structural and Multidisciplinary Optimization*, v. 38, p. 443–453, 2009. Citado na página 87.

SAKATA, S.; ASHIDA, F.; ZAKO, M. Microstructural design of composite materials using fixed-grid modeling and noise-resistant smoothed kriging-based approximate optimization. *Structural and Multidisciplinary Optimization*, v. 36, p. 273–287, 2008. Citado na página 87.

SASENA, M. J. *Flexibility and efficiency enhancements for constrained global design optimization with kriging approximations*. Tese (Doutorado) — University of Michigan, 2002. Citado 6 vezes nas páginas 55, 58, 59, 66, 71 e 84.

SCHONLAU, M. *Computer experiments and global optimization*. Tese (phdthesis) — University of Waterloo, Waterloo, Ontario, Canada, 1997. Citado 6 vezes nas páginas 77, 80, 81, 102, 113 e 137.

SCHONLAU, M.; WELCH, W. J.; JONES, D. R. Global versus local search in constrained optimization of computer models. In: \_\_\_\_\_. *New developments and applications in experimental design*. Hayward, CA: Institute of Mathematical Statistics, 1998. (Lecture Notes–Monograph Series, Volume 34), p. 11–25. Disponível em: <https://doi.org/10.1214/lnms/1215456182>. Citado 2 vezes nas páginas 66 e 77.

SIMPSON, T. W.; MAUERY, T. M.; KORTE, J. J.; MISTREE, F. Kriging models for global approximation in simulation-based multidisciplinary design optimization. *AIAA Journal*, v. 39, n. 12, p. 2233–2241, 2001. Citado 2 vezes nas páginas 59 e 60.

SOIZE, C.; CAPIEZ-LERNOUT, E.; OHAYON, R. Robust updating of uncertain computational models using experimental modal analysis. *AIAA Journal*, v. 46, n. 11, p. 2955–2965, nov. 2008. Citado na página 30.

SPENCE, S. M. J.; KAREEM, A. Performance-based design and optimization of uncertain wind-excited dynamic building systems. *Engineering Structures*, v. 78, p. 133–144, 2014. Citado na página 31.

STAUM, J. Better simulation metamodelling: the why, what and how of stochastic kriging. In: ROSSETTI, M. D.; HILL, R. R.; JOHANSSON,

- B.; DUNKIN, A.; INGALLS, R. G. (Ed.). *Proceedings of the 2009 Winter Simulation Conference*. [S.l.: s.n.], 2009. p. 119–133. Citado na página 88.
- STEPHENS, R. I.; FUCHS, H. O. *Metal Fatigue in Engineering*. [S.l.]: Wiley, 2001. Citado na página 87.
- TIKHONOV, A. N.; ARSENIN, V. Y. *Solutions of ill-posed problems*. [S.l.]: V. H. Winston & Sons, 1977. Citado 2 vezes nas páginas 93 e 94.
- TORII, A. J.; LOPEZ, R. H. Reliability analysis of water distribution networks using the adaptive response surface approach. *Journal of Hydraulic Engineering*, v. 138, n. 3, p. 227–236, 2011. Citado na página 52.
- TORII, A. J.; LOPEZ, R. H.; BIONDINI, F. An approach to reliability-based shape and topology optimization of truss structures. *Engineering Optimization*, Taylor & Francis, v. 44, n. 1, p. 37–53, 2012. Citado na página 52.
- WENZEL, W.; HAMACHER, K. Stochastic tunneling approach for global minimization of complex potential energy landscapes. *Physical Review Letters*, American Physical Society, v. 82, p. 3003–3007, Apr 1999. Citado na página 122.

Imprime uma página indicando o início dos apêndices





## Apêndices



## APÊNDICE A – A DISTRIBUIÇÃO NORMAL MULTIVARIADA

Esse apêndice tem a função de demonstrar a distribuição de probabilidade utilizada na definição da função de verossimilhança na Seção 3.3.3. Todo o desenvolvimento é uma adaptação da demonstração que consta em [Anderson \(2003\)](#).

Seja  $X$  uma variável aleatória que segue uma distribuição normal com média  $\mu$  e desvio padrão  $\sigma$ . Então temos que a função densidade de probabilidade para  $X$  será dada por

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x - \mu)^2\right]. \quad (\text{A.1})$$

De uma maneira geral, a função de densidade normal em uma dimensão pode ser escrita por

$$f(x) = q \exp\left[-\frac{1}{2}(x - b)a(x - b)\right], \quad (\text{A.2})$$

onde  $a$  é uma constante positiva e  $q$  é determinado de modo a satisfazer a condição  $P[-\infty < X < \infty] = 1$  dada por

$$\int_{-\infty}^{\infty} f(x)dx = 1. \quad (\text{A.3})$$

Por comparação entre as equações (A.1) e (A.2) temos que  $q = \frac{1}{\sigma\sqrt{2\pi}}$ ,  $a = \frac{1}{\sigma^2}$  e  $b = \mu$ .

Duas características essenciais da distribuição univariacional são a média e o desvio padrão. A média é uma medida de posição, e o desvio padrão é uma medida de variabilidade da variável aleatória. Para a análise multivariacional, a média e o desvio padrão possuem sua correspondente relevância. Porém, outro aspecto de extrema importância surge como sendo a medida de dependência entre as diferentes variáveis aleatórias. A dependência entre duas variáveis irá envolver a *covariância* entre elas, ou seja, a média do produto dos desvios em

relação a média. Da definição de covariância, se normalizada pelos correspondentes desvios padrões, temos definido o coeficiente de *correlação*, que serve como uma medida do grau de dependência linear entre as variáveis. Dessa forma, um conjunto básico de estatísticas para uma distribuição multivariada será formado pelo vetor de médias (consiste do vetor de médias univariacionais) e pela matriz de covariância (consistindo das variâncias univariacionais e das covariâncias bivariacionais). Pode-se usar, no lugar da matriz de covariância, o conjunto de desvios padrões e a matriz de correlação (como utilizado para definir o modelo do Kriging determinístico).

Tomando então, um vetor de variáveis aleatórias que será denotado por  $\mathbf{X} = \{X_1, X_2, \dots, X_n\}^T$ , teremos que uma distribuição normal multivariada de probabilidade para esse vetor será completamente determinada ou definida por um vetor de médias  $\boldsymbol{\mu}$  e por uma matriz de covariância  $\boldsymbol{\Sigma}$ . Podemos escrever que  $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  onde

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}] = \{\mathbb{E}[X_1], \mathbb{E}[X_2], \dots, \mathbb{E}[X_n]\}^T \quad (\text{A.4})$$

e

$$\boldsymbol{\Sigma} = \text{Cov}[\mathbf{X}, \mathbf{X}] = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T], \quad (\text{A.5})$$

com

$$[\boldsymbol{\Sigma}]_{ij} = \text{Cov}[X_i, X_j]. \quad (\text{A.6})$$

A função de densidade de probabilidade normal do vetor  $\mathbf{X}$  possui uma forma análoga a (A.2), onde basta trocarmos o escalar  $x$  pelo vetor  $\mathbf{x} = \{x_1, x_2, \dots, x_n\}^T$ , o escalar  $b$  pelo vetor  $\mathbf{b} = \{b_1, b_2, \dots, b_n\}^T$  e a constante positiva  $a$  pela matriz

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}.$$

A matriz  $\mathbf{A}$  será considerada como real, simétrica e positiva definida. O termo quadrático  $(x - b)a(x - b)$  de (A.2) poderá ser substituído pela

seguinte forma quadrática

$$(\mathbf{x} - \mathbf{b})^T \mathbf{A}(\mathbf{x} - \mathbf{b}) = (\mathbf{x} - \mathbf{b}) \cdot \mathbf{A}(\mathbf{x} - \mathbf{b}) = \sum_{i,j=1}^n (x_i - b_i) a_{ij} (x_j - b_j).$$

Com essas novas relações podemos definir a função de densidade de probabilidade normal multivariada como

$$f(\mathbf{x}) = Q \exp \left[ -\frac{1}{2}(\mathbf{x} - \mathbf{b})^T \mathbf{A}(\mathbf{x} - \mathbf{b}) \right], \quad (\text{A.7})$$

onde  $Q > 0$  será escolhido de modo a se ter  $P[-\infty < \mathbf{X} < \infty] = 1$ .

Para se determinar o valor de  $Q$  temos que

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} Q \exp \left[ -\frac{1}{2}(\mathbf{x} - \mathbf{b})^T \mathbf{A}(\mathbf{x} - \mathbf{b}) \right] dx_n \cdots dx_1 = 1. \quad (\text{A.8})$$

Como fizemos a suposição de que  $\mathbf{A}$  é positiva definida, então existe uma matriz não singular real  $\mathbf{C}$  tal que

$$\mathbf{C}^T \mathbf{A} \mathbf{C} = \mathbf{I}, \quad (\text{A.9})$$

onde  $\mathbf{I}$  é a matriz identidade de ordem  $n$ . Para simplificar a integração em (A.8) façamos a seguinte mudança de variáveis

$$\mathbf{x} - \mathbf{b} = \mathbf{C} \mathbf{y} \Rightarrow x_i - b_i = \sum_{j=1}^n C_{ij} y_j, \quad (\text{A.10})$$

de onde temos que

$$[\mathbf{J}]_{ij} = \frac{\partial x_i}{\partial y_j} = C_{ij},$$

onde  $\mathbf{J}$  representa a Jacobiana da transformação entre as variáveis. Nesse contexto, podemos escrever que

$$\begin{aligned} dx_n \cdots dx_1 &= \text{abs}(|\mathbf{J}|) dy_n \cdots dy_1 \\ &= \text{abs}(|\mathbf{C}|) dy_n \cdots dy_1. \end{aligned}$$

O operador  $\text{abs}(\cdot)$  representa o valor absoluto e será considerado para garantirmos que a mudança de variáveis mantenha a integração positiva, e conseqüentemente que  $Q > 0$ . Portanto, temos que (A.8) se torna

$$\begin{aligned} Q \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp \left[ -\frac{1}{2} (\mathbf{C}\mathbf{y})^T \mathbf{A} (\mathbf{C}\mathbf{y}) \right] \text{abs}(|\mathbf{C}|) dy_n \cdots dy_1 &= 1 \\ Q \text{abs}(|\mathbf{C}|) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp \left[ -\frac{1}{2} (\mathbf{y}^T \mathbf{C}^T \mathbf{A} \mathbf{C} \mathbf{y}) \right] dy_n \cdots dy_1 &= 1 \\ Q \text{abs}(|\mathbf{C}|) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp \left[ -\frac{1}{2} (\mathbf{y}^T \mathbf{y}) \right] dy_n \cdots dy_1 &= 1. \end{aligned} \quad (\text{A.11})$$

Para essa nova integração vemos que

$$\begin{aligned} \exp \left[ -\frac{1}{2} (\mathbf{y}^T \mathbf{y}) \right] &= \exp \left[ -\frac{1}{2} (\mathbf{y} \cdot \mathbf{y}) \right] \\ &= \exp \left( -\frac{1}{2} \sum_{i=1}^n y_i^2 \right) \\ &= \prod_{i=1}^n \exp \left( -\frac{1}{2} y_i^2 \right). \end{aligned} \quad (\text{A.12})$$

Se aplicarmos o determinante em ambos os lados da equação (A.9) temos que

$$\begin{aligned} |\mathbf{C}^T \mathbf{A} \mathbf{C}| &= |\mathbf{I}| \\ |\mathbf{C}^T| |\mathbf{A}| |\mathbf{C}| &= 1, \end{aligned}$$

e como  $|\mathbf{C}^T| = |\mathbf{C}|$  temos que

$$|\mathbf{C}| = \frac{1}{\sqrt{|\mathbf{A}|}} \Rightarrow \text{abs}(|\mathbf{C}|) = \frac{1}{\sqrt{|\mathbf{A}|}}.$$

Assim, vemos que (A.11) é reduzida a

$$\begin{aligned} Q \frac{1}{\sqrt{|\mathbf{A}|}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^n \exp \left( -\frac{1}{2} y_i^2 \right) dy_n \cdots dy_1 &= 1 \\ Q \frac{1}{\sqrt{|\mathbf{A}|}} \prod_{i=1}^n \left[ \int_{-\infty}^{\infty} \exp \left( -\frac{1}{2} y_i^2 \right) dy_i \right] &= 1. \end{aligned} \quad (\text{A.13})$$

A integral remanescente em (A.13) é conhecida como Integral Gaussiana e seu valor,  $\sqrt{2\pi}$ , é determinado na Seção A.1. Logo,

$$Q \frac{1}{\sqrt{|\mathbf{A}|}} \prod_{i=1}^n \left[ \sqrt{2\pi} \right] = 1$$

$$Q \frac{1}{\sqrt{|\mathbf{A}|}} \left( \sqrt{2\pi} \right)^n = 1,$$

e finalmente encontramos que

$$Q = \frac{\sqrt{|\mathbf{A}|}}{(2\pi)^{\frac{n}{2}}}. \quad (\text{A.14})$$

Portanto a função densidade de probabilidade normal multivariada será dada por

$$f(\mathbf{x}) = \frac{\sqrt{|\mathbf{A}|}}{(2\pi)^{\frac{n}{2}}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \mathbf{b})^T \mathbf{A} (\mathbf{x} - \mathbf{b}) \right]. \quad (\text{A.15})$$

Para identificarmos o significado dos termos  $\mathbf{b}$  e  $\mathbf{A}$  na equação (A.15), no contexto de um vetor de variáveis aleatórias, é necessário que façamos o cálculo dos primeiro e segundo momentos do vetor de variáveis aleatórias  $\mathbf{X}$ , ou seja, os momentos das variáveis  $X_1, X_2, \dots, X_n$ . Para tanto, voltemos a mudança de coordenadas (A.10) para podermos escrever que

$$\mathbf{X} = \mathbf{C}\mathbf{Y} + \mathbf{b}, \quad (\text{A.16})$$

com  $\mathbf{C}$  e  $\mathbf{b}$  sendo, respectivamente, uma matriz e vetor de valores reais. Dessa forma temos que o primeiro momento do vetor  $\mathbf{X}$  (que representa a média ou esperança matemática) será dado por

$$\begin{aligned} \mathbb{E}[\mathbf{X}] &= \mathbb{E}[\mathbf{C}\mathbf{Y} + \mathbf{b}] \\ &= \mathbf{C}\mathbb{E}[\mathbf{Y}] + \mathbf{b}. \end{aligned} \quad (\text{A.17})$$

Da transformação (A.16) espera-se, pelos resultados apresentados anteriormente, que a função densidade de probabilidade de  $\mathbf{Y}$  seja proporcional a (A.12), dessa forma, vemos que

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp \left[ -\frac{1}{2} (\mathbf{y}^T \mathbf{y}) \right] = \prod_{i=1}^n \left[ \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} y_i^2 \right) \right], \quad (\text{A.18})$$

e conseqüentemente a  $i$ -ésima componente do vetor de esperanças de  $\mathbf{Y}$  será dado por

$$\begin{aligned} [\mathbb{E}[\mathbf{Y}]]_i &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} y_i \prod_{j=1}^n \left[ \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y_j^2\right) \right] dy_n \cdots dy_1 \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y_i \exp\left(-\frac{1}{2}y_i^2\right) dy_i \prod_{j=1, j \neq i}^n \left[ \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y_j^2\right) dy_j \right] \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y_i \exp\left(-\frac{1}{2}y_i^2\right) dy_i \\ &= 0. \end{aligned}$$

Este último resultado é consequência direta da função  $y_i \exp\left(-\frac{1}{2}y_i^2\right)$  ser ímpar em  $y_i$ . Conseqüentemente, temos que  $\mathbb{E}[\mathbf{Y}] = \mathbf{0}$ . Logo, (A.17) será dado por

$$\boldsymbol{\mu} = \mathbf{C}\mathbb{E}[\mathbf{Y}] + \mathbf{b} = \mathbf{C}\mathbf{0} + \mathbf{b} = \mathbf{b}. \quad (\text{A.19})$$

Dessa forma, podemos afirmar que se um vetor de variáveis aleatórias  $\mathbf{X}$  possui média  $\boldsymbol{\mu}$ , então seu equivalente na função de distribuição (A.15) será o vetor  $\mathbf{b}$ .

Para o segundo momento da variável aleatória  $\mathbf{X}$ , que será representado pela covariância entre as variáveis aleatórias, começaremos analisando a expressão (A.5), onde

$$\begin{aligned} \text{Cov}[\mathbf{X}, \mathbf{X}] &= \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] \\ &= \mathbb{E}[(\mathbf{C}\mathbf{Y} + \mathbf{b} - (\mathbf{C}\mathbb{E}[\mathbf{Y}] + \mathbf{b}))(\mathbf{C}\mathbf{Y} + \mathbf{b} - (\mathbf{C}\mathbb{E}[\mathbf{Y}] + \mathbf{b}))^T] \\ &= \mathbb{E}[(\mathbf{C}\mathbf{Y})(\mathbf{C}\mathbf{Y})^T] \\ &= \mathbb{E}[\mathbf{C}\mathbf{Y}\mathbf{Y}^T\mathbf{C}^T] \\ &= \mathbf{C}\mathbb{E}[\mathbf{Y}\mathbf{Y}^T]\mathbf{C}^T. \end{aligned} \quad (\text{A.20})$$

O elemento  $ij$  da matriz  $\mathbb{E}[\mathbf{Y}\mathbf{Y}^T]$  poderá ser encontrado a partir de (A.18) utilizando

$$[\mathbb{E}[\mathbf{Y}\mathbf{Y}^T]]_{ij} = \mathbb{E}[Y_i Y_j]$$



$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} y_i y_j \prod_{l=1}^n \left[ \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y_l^2\right) \right] dy_n \cdots dy_1. \quad (\text{A.21})$$

Vamos supor dois casos:  $i = j$  e  $i \neq j$ . O primeiro nos dará os elementos da diagonal principal de  $\mathbb{E}[\mathbf{Y}\mathbf{Y}^T]$ , enquanto o segundo nos dará os demais elementos. Fazendo  $i = j$  temos

$$\begin{aligned} \mathbb{E}[Y_i^2] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y_i^2 \exp\left(-\frac{1}{2}y_i^2\right) dy_i \cdot \prod_{l=1, l \neq i}^n \left[ \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y_l^2\right) dy_l \right] \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y_i^2 \exp\left(-\frac{1}{2}y_i^2\right) dy_i \\ &= \frac{1}{\sqrt{2\pi}} \sqrt{2\pi} \\ &= 1. \end{aligned}$$

A última integral que aparece na expressão acima é deduzida na Seção A.2. Por fim, tomando  $i \neq j$  temos

$$\begin{aligned} \mathbb{E}[Y_i Y_j] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y_i \exp\left(-\frac{1}{2}y_i^2\right) dy_i \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y_j \exp\left(-\frac{1}{2}y_j^2\right) dy_j \cdot \\ &\quad \prod_{l=1, l \neq i, j}^n \left[ \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y_l^2\right) dy_l \right] \\ &= 0 \cdot 0 \cdot 1 \\ &= 0. \end{aligned}$$

Portanto, encontramos que  $\mathbb{E}[\mathbf{Y}\mathbf{Y}^T] = \mathbf{I}$  e (A.20) se torna

$$\text{Cov}[\mathbf{X}, \mathbf{X}] = \mathbf{C}\mathbf{I}\mathbf{C}^T = \mathbf{C}\mathbf{C}^T. \quad (\text{A.22})$$

Da equação (A.9), se tomarmos a inversa em ambos os lados temos

$$\begin{aligned} (\mathbf{C}^T \mathbf{A} \mathbf{C})^{-1} &= \mathbf{I}^{-1} \\ \mathbf{C}^{-1} \mathbf{A}^{-1} \mathbf{C}^{-T} &= \mathbf{I} \\ \mathbf{C} \mathbf{C}^{-1} \mathbf{A}^{-1} \mathbf{C}^{-T} \mathbf{C}^T &= \mathbf{C} \mathbf{I} \mathbf{C}^T \\ \mathbf{A}^{-1} &= \mathbf{C} \mathbf{C}^T, \end{aligned}$$

e portanto encontramos que a matriz de covariância será uma matriz positiva definida dada por

$$\Sigma = \text{Cov}[\mathbf{X}, \mathbf{X}] = \mathbf{A}^{-1}. \quad (\text{A.23})$$

Vemos assim que, se um vetor de variáveis aleatórias  $\mathbf{X}$  possuir uma matriz de covariância  $\Sigma$  positiva definida, então sua correspondente na função de distribuição (A.15) será  $\mathbf{A}^{-1}$ .

Portanto, temos que se  $\mathbf{X}$  é tal que  $\boldsymbol{\mu}$  é seu vetor de médias e  $\Sigma$  sua matriz de covariância, então a função densidade de probabilidade normal multivariada fica definida e dada por

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]. \quad (\text{A.24})$$

## A.1 INTEGRAL GAUSSIANA

Vamos determinar o valor da seguinte integral

$$I = \int_{-\infty}^{\infty} \exp \left( -\frac{1}{2} x^2 \right) dx, \quad (\text{A.25})$$

que também poderá ser escrita como

$$I = \int_{-\infty}^{\infty} \exp \left( -\frac{1}{2} y^2 \right) dy. \quad (\text{A.26})$$

Multiplicando as equações (A.25) e (A.26) temos que

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} \exp \left( -\frac{1}{2} x^2 \right) dx \int_{-\infty}^{\infty} \exp \left( -\frac{1}{2} y^2 \right) dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp \left[ -\frac{1}{2} (x^2 + y^2) \right] dx dy. \end{aligned}$$

Realizando uma troca das coordenadas cartesianas  $(x, y)$  para as coordenadas polares  $(r, \theta)$  encontramos que

$$I^2 = \int_0^{2\pi} \int_0^{\infty} \exp \left[ -\frac{1}{2} r^2 \right] r dr d\theta.$$

Fazendo  $t = -\frac{1}{2}r^2$  temos que  $dt = -rdr$  e que se  $r = 0, \infty$  então  $t = 0, -\infty$ , respectivamente. Dessa forma

$$\begin{aligned} I^2 &= \int_0^{2\pi} \int_{-\infty}^0 \exp(t) dt d\theta \\ &= \int_0^{2\pi} \left[ \int_{-\infty}^0 \exp(t) dt \right] d\theta \\ &= \int_0^{2\pi} \left[ \exp(t) \Big|_{-\infty}^0 \right] d\theta \\ &= \int_0^{2\pi} 1 d\theta \\ &= 2\pi. \end{aligned}$$

Portanto temos que

$$I^2 = 2\pi \Rightarrow I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}x^2\right) dx = \sqrt{2\pi}. \quad (\text{A.27})$$

## A.2 INTEGRAL AUXILIAR

Seja a seguinte função primitiva

$$I(x) = \int x^2 \exp\left(-\frac{1}{2}x^2\right) dx.$$

Tomemos agora as seguintes funções  $a(x) = x$  e  $b(x) = x \exp\left(-\frac{1}{2}x^2\right)$ , de onde o integrando de  $I$  se torna  $a(x) \cdot b(x)$ . Façamos

$$u = a(x) \Rightarrow du = dx$$

e

$$dv = b(x)dx \Rightarrow v = -\exp\left(-\frac{1}{2}x^2\right),$$

dessa forma temos que

$$\begin{aligned} I(x) &= \int u dv = uv - \int v du \\ &= -x \exp\left(-\frac{1}{2}x^2\right) + \int \exp\left(-\frac{1}{2}x^2\right) dx. \end{aligned}$$

Portanto, temos que

$$\begin{aligned} \int_{-\infty}^{\infty} x^2 \exp\left(-\frac{1}{2}x^2\right) dx &= -x \exp\left(-\frac{1}{2}x^2\right) \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}x^2\right) dx \\ &= -x \exp\left(-\frac{1}{2}x^2\right) \Big|_{-\infty}^{\infty} + \sqrt{2\pi} \\ &= \sqrt{2\pi} - \left[ \lim_{x \rightarrow \infty} x \exp\left(-\frac{1}{2}x^2\right) - \right. \\ &\quad \left. \lim_{x \rightarrow -\infty} x \exp\left(-\frac{1}{2}x^2\right) \right]. \end{aligned}$$

Para calcular os limites que se apresentam entre colchetes, podemos utilizar a seguinte expansão em série de Taylor

$$\exp\left(\frac{1}{2}x^2\right) = 1 + \frac{x^2}{2} + \frac{x^4}{2^2 2!} + \frac{x^6}{2^3 3!} + \cdots = 1 + \sum_{i=1}^{\infty} \frac{x^{2i}}{2^i i!},$$

assim os limites poderão ser calculados por

$$\begin{aligned} \lim_{x \rightarrow \pm\infty} x \exp\left(-\frac{1}{2}x^2\right) &= \lim_{x \rightarrow \pm\infty} \frac{x}{\exp\left(\frac{1}{2}x^2\right)} \\ &= \lim_{x \rightarrow \pm\infty} \frac{x}{1 + \sum_{i=1}^{\infty} \frac{x^{2i}}{2^i i!}} \\ &= \lim_{x \rightarrow \pm\infty} \frac{x}{x \left( \frac{1}{x} + \sum_{i=1}^{\infty} \frac{x^{2i-1}}{2^i i!} \right)} \\ &= \lim_{x \rightarrow \pm\infty} \frac{1}{\frac{1}{x} + \sum_{i=1}^{\infty} \frac{x^{2i-1}}{2^i i!}}. \end{aligned}$$

Pode-se ver que  $\frac{1}{x} \rightarrow 0$  conforme  $x \rightarrow \pm\infty$  e que  $\sum_{i=1}^{\infty} \frac{x^{2i-1}}{2^i i!} \rightarrow \pm\infty$  quando  $x \rightarrow \pm\infty$ , logo

$$\lim_{x \rightarrow \pm\infty} x \exp\left(-\frac{1}{2}x^2\right) = \frac{1}{0 \pm \infty} = 0.$$

Assim concluímos que

$$\int_{-\infty}^{\infty} x^2 \exp\left(-\frac{1}{2}x^2\right) dx = \sqrt{2\pi} - [0 - 0]$$

$$= \sqrt{2\pi}.$$



## APÊNDICE B – DERIVADAS DA FUNÇÃO DE VEROSSIMILHANÇA PARA O SK

Seja a função de verossimilhança dada por

$$L_{\ln}(\mu, \sigma_z^2, \boldsymbol{\theta}) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} (\bar{\mathbf{y}} - \mu \mathbf{1})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{y}} - \mu \mathbf{1}), \quad (\text{B.1})$$

onde  $\boldsymbol{\Sigma} = \sigma_z^2 \boldsymbol{\Psi} + \boldsymbol{\Sigma}_\varepsilon$  e  $\boldsymbol{\Psi}$  é uma função do parâmetro  $\boldsymbol{\theta}$ .

Para encontrarmos os parâmetros que maximizam (B.1), devemos resolver simultaneamente

$$\frac{\partial L_{\ln}(\mu, \sigma_z^2, \boldsymbol{\theta})}{\partial \mu} = 0 \quad \frac{\partial L_{\ln}(\mu, \sigma_z^2, \boldsymbol{\theta})}{\partial \sigma_z^2} = 0 \quad \frac{\partial L_{\ln}(\mu, \sigma_z^2, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}. \quad (\text{B.2})$$

Antes de partirmos para as derivadas, precisaremos de um resultado importante do cálculo diferencial para matrizes. Tomemos uma matriz quadrada qualquer  $\mathbf{A}$ , que possua elementos que são funções de algum parâmetro  $\beta$ . Então temos que:

$$\frac{\partial |\mathbf{A}|}{\partial \beta} = |\mathbf{A}| \operatorname{tr} \left( \mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial \beta} \right) \quad (\text{B.3})$$

$$\frac{\partial \mathbf{A}^{-1}}{\partial \beta} = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial \beta} \mathbf{A}^{-1}, \quad (\text{B.4})$$

onde o operador  $\operatorname{tr}(\cdot)$  representa o traço da matriz.

Tomando a derivada de (B.1) em função do parâmetro  $\mu$  temos

$$\begin{aligned} \frac{\partial L_{\ln}(\mu, \sigma_z^2, \boldsymbol{\theta})}{\partial \mu} &= -\frac{1}{2} \left[ -\mathbf{1}^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{y}} - \mu \mathbf{1}) - (\bar{\mathbf{y}} - \mu \mathbf{1})^T \boldsymbol{\Sigma}^{-1} \mathbf{1} \right] \\ &= \mathbf{1}^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{y}} - \mu \mathbf{1}) = 0 \end{aligned} \quad (\text{B.5})$$

e dessa forma

$$\mu = \frac{\mathbf{1}^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{y}}}{\mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{1}}. \quad (\text{B.6})$$

Podemos ver que a equação (B.5) depende dos três parâmetros  $\mu, \sigma_z^2$  e  $\boldsymbol{\theta}$ , porém podemos colocar  $\mu$  em função dos outros dois parâmetros, vide equação (B.6).

Para a derivada em função de  $\sigma_z^2$ , primeiro vemos que

$$\begin{aligned} \frac{\partial \Sigma}{\partial \sigma_z^2} &= \frac{\partial(\sigma_z^2 \Psi + \Sigma_\varepsilon)}{\partial \sigma_z^2} \\ &= \Psi. \end{aligned}$$

Logo, temos que

$$\begin{aligned} \frac{\partial L_{\ln}(\mu, \sigma_z^2, \theta)}{\partial \sigma_z^2} &= -\frac{1}{2} \left\{ \frac{1}{|\Sigma|} |\Sigma| \operatorname{tr} [\Sigma^{-1} \Psi] \right. \\ &\quad \left. + (\bar{\mathbf{y}} - \mu \mathbf{1})^T [-\Sigma^{-1} \Psi \Sigma^{-1}] (\bar{\mathbf{y}} - \mu \mathbf{1}) \right\}, \end{aligned}$$

de onde conseguimos a equação

$$\operatorname{tr} [\Sigma^{-1} \Psi] - (\bar{\mathbf{y}} - \mu \mathbf{1})^T [\Sigma^{-1} \Psi \Sigma^{-1}] (\bar{\mathbf{y}} - \mu \mathbf{1}) = 0. \quad (\text{B.7})$$

Essa última equação é uma função dos três parâmetros  $\mu, \sigma_z^2$  e  $\theta$ .

Por fim, para a derivada em função de  $\theta$ , temos primeiro que para a  $i$ -ésima dimensão do vetor  $\theta$

$$\begin{aligned} \frac{\partial \Sigma}{\partial \theta_i} &= \frac{\partial(\sigma_z^2 \Psi + \Sigma_\varepsilon)}{\partial \theta_i} \\ &= \sigma_z^2 \frac{\partial \Psi}{\partial \theta_i}, \end{aligned}$$

e dessa forma

$$\begin{aligned} \frac{\partial L_{\ln}(\mu, \sigma_z^2, \theta)}{\partial \theta_i} &= -\frac{1}{2} \left\{ \frac{1}{|\Sigma|} |\Sigma| \operatorname{tr} \left[ \Sigma^{-1} \sigma_z^2 \frac{\partial \Psi}{\partial \theta_i} \right] \right. \\ &\quad \left. + (\bar{\mathbf{y}} - \mu \mathbf{1})^T \left[ -\Sigma^{-1} \sigma_z^2 \frac{\partial \Psi}{\partial \theta_i} \Sigma^{-1} \right] (\bar{\mathbf{y}} - \mu \mathbf{1}) \right\}, \end{aligned}$$

e consequentemente

$$\sigma_z^2 \left\{ \operatorname{tr} \left[ \Sigma^{-1} \frac{\partial \Psi}{\partial \theta_i} \right] - (\bar{\mathbf{y}} - \mu \mathbf{1})^T \left[ \Sigma^{-1} \frac{\partial \Psi}{\partial \theta_i} \Sigma^{-1} \right] (\bar{\mathbf{y}} - \mu \mathbf{1}) \right\} = 0.$$

Como devemos ter  $\sigma_z^2 > 0$  então

$$\operatorname{tr} \left[ \Sigma^{-1} \frac{\partial \Psi}{\partial \theta_i} \right] - (\bar{\mathbf{y}} - \mu \mathbf{1})^T \left[ \Sigma^{-1} \frac{\partial \Psi}{\partial \theta_i} \Sigma^{-1} \right] (\bar{\mathbf{y}} - \mu \mathbf{1}) = 0. \quad (\text{B.8})$$



Podemos ver que (B.8) é novamente dependente dos parâmetros  $\mu, \sigma_z^2, \boldsymbol{\theta}$ .

Foram obtidas duas equações com as derivadas em relação a  $\mu$  e  $\sigma_z^2$ , mais  $k$  equações com as derivadas em relação a  $\theta_i, i = 1, 2, \dots, k$ ; portanto, temos formado um sistema com  $k + 2$  equações não lineares para  $k + 2$  incógnitas. Dessa forma, a solução desse sistema é capaz de nos fornecer os estimadores  $\hat{\mu}, \hat{\sigma}_z^2$  e  $\hat{\boldsymbol{\theta}}$ .

Para encerrar, se definirmos a priori que a função de correlação entre os pontos amostrais  $\mathbf{d}$  será dada por  $h(\mathbf{d}^{(i)}, \mathbf{d}^{(j)}) = \exp\left(-\sum_{r=1}^k \theta_r |d_r^{(i)} - d_r^{(j)}|^2\right)$ , então teremos que o elemento  $ij$  da matriz  $\partial\Psi/\partial\theta_q$  para a  $q$ -ésima dimensão de  $\boldsymbol{\theta}$  será

$$\left[\frac{\partial\Psi}{\partial\theta_q}\right]_{ij} = -|d_q^{(i)} - d_q^{(j)}|^2 \exp\left(-\sum_{r=1}^k \theta_r |d_r^{(i)} - d_r^{(j)}|^2\right).$$



## APÊNDICE C – SIMPLIFICAÇÕES NUMÉRICAS PARA O ALGORITMO

### C.1 TÉCNICAS MATRICIAIS PARA A ANÁLISE NUMÉRICA

Pode ser visto nas equações (3.17), (3.18), (3.31) e (3.33) que a maior carga computacional associada à construção do metamodelo com o Kriging (determinístico ou estocástico) está na inversão da matriz de correlação  $\Psi$ , para o Kriging determinístico, ou na matriz de covariância  $\Sigma = \sigma_z^2 \Psi + \Sigma_\varepsilon$ . Entretanto, uma vez que os parâmetros  $\theta$  estejam definidos, a inversa  $\Psi^{-1}$  precisará ser calculada uma única vez.

Este apêndice tem por objetivo fornecer a técnica utilizada na implementação do algoritmo numérico que gera o metamodelo em Kriging determinístico e realiza a otimização via EGO. Uma análise análoga poderá ser feita para o caso do SK e do sEGO.

Vamos começar com o benefício da matriz de correlação  $\Psi$  ser simétrica e positiva definida, assim temos que

$$\Psi = \mathbf{C}^T \mathbf{C} \tag{C.1}$$

onde  $\mathbf{C}$  é a matriz obtida pela decomposição Cholesky de  $\Psi$ . Logo temos que a inversa de  $\Psi$  é dada por

$$\begin{aligned} \Psi^{-1} &= (\mathbf{C}^T \mathbf{C})^{-1} \\ &= \mathbf{C}^{-1} \mathbf{C}^{-T}. \end{aligned} \tag{C.2}$$

Iremos partir para algumas simplificações extremamente úteis. Primeiro tomemos o estimador da média do metamodelo,  $\mu$  (3.17). Assim,

$$\begin{aligned} \hat{\mu} &= \frac{\mathbf{1}^T \Psi^{-1} \mathbf{y}}{\mathbf{1}^T \Psi^{-1} \mathbf{1}} \\ &= \frac{\mathbf{1} \cdot \mathbf{C}^{-1} \mathbf{C}^{-T} \mathbf{y}}{\mathbf{1} \cdot \mathbf{C}^{-1} \mathbf{C}^{-T} \mathbf{1}} \\ &= \frac{\mathbf{C}^{-T} \mathbf{y} \cdot \mathbf{C}^{-T} \mathbf{1}}{\mathbf{C}^{-T} \mathbf{1} \cdot \mathbf{C}^{-T} \mathbf{1}}. \end{aligned}$$

Sejam os vetores  $\mathbf{a} = \mathbf{C}^{-T} \mathbf{1}$  e  $\mathbf{b} = \mathbf{C}^{-T} \mathbf{y}$ , então temos que

$$\hat{\mu} = \frac{\mathbf{b} \cdot \mathbf{a}}{\mathbf{a} \cdot \mathbf{a}}, \quad (\text{C.3})$$

onde teremos que  $\mathbf{a} \cdot \mathbf{a} = \sum_{i=1}^n a_i^2$ .

Partindo para o estimador da variância da parcela  $Z(\mathbf{d})$  (3.18) temos que

$$\begin{aligned} \hat{\sigma}^2 &= \frac{(\mathbf{y} - \mathbf{1}\hat{\mu})^T \boldsymbol{\Psi}^{-1} (\mathbf{y} - \mathbf{1}\hat{\mu})}{n} \\ &= \frac{(\mathbf{y} - \mathbf{1}\hat{\mu}) \cdot \mathbf{C}^{-1} \mathbf{C}^{-T} (\mathbf{y} - \mathbf{1}\hat{\mu})}{n} \\ &= \frac{\mathbf{C}^{-T} (\mathbf{y} - \mathbf{1}\hat{\mu}) \cdot \mathbf{C}^{-T} (\mathbf{y} - \mathbf{1}\hat{\mu})}{n}. \end{aligned}$$

Seja o vetor  $\boldsymbol{\rho} = \mathbf{C}^{-T} (\mathbf{y} - \mathbf{1}\hat{\mu})$  teremos que

$$\hat{\sigma}^2 = \frac{\boldsymbol{\rho} \cdot \boldsymbol{\rho}}{n} = \frac{\sum_{i=1}^n \rho_i^2}{n}, \quad (\text{C.4})$$

onde podemos aplicar que  $\boldsymbol{\rho} = \mathbf{b} - \mu \mathbf{a}$ .

Podemos ver que, até o momento, para o cálculo dos estimadores da média e da variância do metamodelo, só necessitamos da inversa da matriz decomposta  $\mathbf{C}^T$ . O cálculo dessa inversa é menos caro computacionalmente do que o cálculo da inversa de  $\boldsymbol{\Psi}$ , pois  $\mathbf{C}^T$  é uma matriz triangular inferior. Outro detalhe importante é que  $\boldsymbol{\Psi}$ ,  $\mathbf{C}^T$ ,  $\mathbf{C}^{-T}$ ,  $\mathbf{a}$ ,  $\mathbf{b}$ ,  $\boldsymbol{\rho}$ ,  $\hat{\mu}$  e  $\hat{\sigma}^2$  só serão calculadas dentro do algoritmo de otimização da verossimilhança e para a avaliação da função (3.14). Após o ajuste dos parâmetros, esses valores serão fixos, e poderemos carregá-los como pilares para construção do metamodelo.

Para o preditor (3.31) teremos que

$$\begin{aligned} \hat{Y}(\mathbf{d}) &= \hat{\mu} + \mathbf{h}^T \boldsymbol{\Psi}^{-1} (\mathbf{y} - \mathbf{1}\hat{\mu}) \\ &= \hat{\mu} + \mathbf{h} \cdot \mathbf{C}^{-1} \mathbf{C}^{-T} (\mathbf{y} - \mathbf{1}\hat{\mu}) \\ &= \hat{\mu} + \mathbf{C}^{-T} (\mathbf{y} - \mathbf{1}\hat{\mu}) \cdot \mathbf{C}^{-T} \mathbf{h}, \end{aligned}$$

onde podemos fazer  $\mathbf{v} = \mathbf{C}^{-T} \mathbf{h}$  e obter

$$\hat{Y}(\mathbf{d}) = \hat{\mu} + \boldsymbol{\rho} \cdot \mathbf{v}. \quad (\text{C.5})$$

Para o cálculo do MSE (3.33) teremos que analisar duas parcelas diferentes. Primeiro vamos começar com

$$\begin{aligned} \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h} &= \mathbf{h} \cdot \mathbf{C}^{-1} \mathbf{C}^{-T} \mathbf{h} \\ &= \mathbf{C}^{-T} \mathbf{h} \cdot \mathbf{C}^{-T} \mathbf{h} \\ &= \mathbf{v} \cdot \mathbf{v} \\ &= \sum_{i=1}^n v_i^2 \end{aligned}$$

A segunda parcela a ser analisada é

$$\begin{aligned} \frac{(1 - \mathbf{1}^T \boldsymbol{\Psi}^{-1} \mathbf{h})^2}{\mathbf{1}^T \boldsymbol{\Psi}^{-1} \mathbf{1}} &= \frac{(1 - \mathbf{1} \cdot \mathbf{C}^{-1} \mathbf{C}^{-T} \mathbf{h})^2}{\mathbf{1} \cdot \mathbf{C}^{-1} \mathbf{C}^{-T} \mathbf{1}} \\ &= \frac{(1 - \mathbf{C}^{-T} \mathbf{h} \cdot \mathbf{C}^{-T} \mathbf{1})^2}{\mathbf{C}^{-T} \mathbf{1} \cdot \mathbf{C}^{-T} \mathbf{1}} \\ &= \frac{(1 - \mathbf{v} \cdot \mathbf{a})^2}{\mathbf{a} \cdot \mathbf{a}} \\ &= \left( \frac{1 - \mathbf{v} \cdot \mathbf{a}}{\|\mathbf{a}\|} \right)^2. \end{aligned}$$

Portanto, temos que o MSE será dado por

$$\begin{aligned} s^2(\mathbf{d}) &= \hat{\sigma}^2 \left[ 1 - \mathbf{h}^T \boldsymbol{\Psi}^{-1} \mathbf{h} + \frac{(1 - \mathbf{1}^T \boldsymbol{\Psi}^{-1} \mathbf{h})^2}{\mathbf{1}^T \boldsymbol{\Psi}^{-1} \mathbf{1}} \right] \\ &= \hat{\sigma}^2 \left[ 1 - \sum_{i=1}^n v_i^2 + \left( \frac{1 - \mathbf{v} \cdot \mathbf{a}}{\|\mathbf{a}\|} \right)^2 \right]. \end{aligned} \quad (\text{C.6})$$

Podemos ver que, nas equações (C.5) e (C.6), o que irá mudar de ponto para ponto a ser predito é o valor de  $\mathbf{v}$ . Logo, como o valor de  $\mathbf{C}^{-T}$  já foi previamente calculado no fim da otimização dos parâmetros, não necessitamos do cálculo de uma nova inversa.

Dessa forma, somos capazes de mostrar que necessitamos do cálculo de uma única inversão, a da matriz  $\mathbf{C}^T$ . Porém, mesmo com essa simplificação, ainda devemos tomar um cuidado com a matriz  $\mathbf{C}$ . Conforme o número de pontos amostrais cresce, o custo computacional

para obtermos a fatoração Cholesky de  $\Psi$  e a respectiva inversa  $\mathbf{C}^{-T}$  também aumenta. Então, mesmo tendo reduzido o processo de predição para o cálculo de apenas uma inversa no preditor e no MSE, durante a otimização da verossimilhança, precisaremos calcular essa inversa várias vezes. Por isso ainda manteremos o custo computacional elevado para espaços amostrais grandes. Isso inviabiliza a utilização do Kriging para problemas com dimensões elevadas (normalmente maiores que 20), onde necessitamos de uma grande quantidade de pontos para realizar uma boa aproximação.

Outro complicador no algoritmo de otimização da verossimilhança é que, conforme ajustamos o parâmetro  $\theta$ , a matriz  $\Psi$  pode se aproximar da matriz singular (isso se deve aos erros numéricos de aproximação). Quando isso acontece, o algoritmo Cholesky irá falhar, mesmo com a matriz sendo positiva definida. Dessa forma, precisamos penalizar a função de verossimilhança para esse  $\theta$ . Uma alternativa básica é setar o valor de  $\mathcal{L}_{\ln}(\theta) = 10^{20}$  (um valor alto) para aqueles parâmetros  $\theta$  onde o Cholesky falha. Espera-se com isso que o algoritmo fuja desses pontos durante a otimização dos parâmetros.

Uma abordagem semelhante deverá ser utilizada, caso a fatoração Cholesky funcione, porém o determinante da matriz  $\Psi$  seja muito próximo de zero. Nesses casos, o valor de  $\ln(|\Psi|)$  tenderá a menos infinito na equação (3.19) e conseqüentemente fará  $\mathcal{L}_{\ln}(\theta) \rightarrow \infty$ . Assim, um valor infinito em um algoritmo de maximização fará com que esse  $\theta$ , que gera a matriz quase singular, seja o maximizador. É nítido que tais comportamentos só ocorrerão devido aos erros numéricos das operações intrínsecas da máquina que os executa. Portanto, achamos viável ajustar  $\mathcal{L}_{\ln}(\theta) = 10^{20}$  para os parâmetros  $\theta$  que gerassem  $|\Psi| < 10^{-300}$ . O valor  $10^{-300}$  foi idealizado por meio de debuggs realizados no código e a respectiva análise do comportamento da função de verossimilhança. Foram realizados testes com várias limitações, entre elas  $10^{-10}$ ,  $10^{-20}$ ,  $10^{-30}$ ,  $10^{-100}$ ,  $10^{-300}$  e  $10^{-1000}$ . Os quatro primeiros limites eram ativados por muitos  $\theta$  e observou-se que, para eles, o valor

da verossimilhança não tendia ao infinito. Logo, vários parâmetros eram descartados sem sua prévia análise. Isso compromete todo o processo, pois é possível que deixemos de lado pontos que trariam um melhor modelo ajustado. Verificou-se também que esses quatro valores causavam o aumento do MSE, conforme os IPs iam sendo adicionados. Porém, esse não é o comportamento esperado, já que vimos que o MSE deve diminuir conforme diminuimos a incerteza no modelo. Para o caso  $10^{-1000}$ , vimos que poucos parâmetros ativavam essa restrição, e alguns dos que não ativavam, faziam a verossimilhança tender ao infinito. Para alguns valores analisados entre  $10^{-300}$  e  $10^{-1000}$ , não encontrou-se valor que gerasse um comportamento fora do esperado.

Todas as simplificações e considerações numéricas realizadas neste apêndice são estendidas para o caso de se determinar o modelo em SK.

## C.2 CÁLCULO SIMPLIFICADO DA MÉTRICA EI

Como apresentado na Seção 3.4.1, a métrica EI é uma das formas mais tradicionais de se escolher IPs. Ela é utilizada tanto no EGO quanto em quatro das métricas para o sEGO. Portanto, devemos de forma adequada realizar seu cálculo. Relembrando sua definição, temos que seu valor é dado por

$$\mathbb{E}(I(\mathbf{d})) = \left( f_{\min} - \hat{Y}(\mathbf{d}) \right) \Phi \left( \frac{f_{\min} - \hat{Y}(\mathbf{d})}{s(\mathbf{d})} \right) + s(\mathbf{d}) \phi \left( \frac{f_{\min} - \hat{Y}(\mathbf{d})}{s(\mathbf{d})} \right).$$

Omitindo a dependência de  $\mathbf{d}$  em  $\hat{Y}(\mathbf{d})$  e  $s(\mathbf{d})$  e considerando a variável  $u = \frac{f_{\min} - \hat{Y}}{s}$  temos que

$$\mathbb{E}(I(\mathbf{d})) = su\Phi(u) + s\phi(u). \quad (\text{C.7})$$

As funções auxiliares  $\Phi(u)$  e  $\phi(u)$  são, respectivamente, a função distribuição cumulativa de probabilidade e a função densidade de probabilidade, ambas, para uma variável normal padrão  $\mathcal{N}(0, 1)$ . Logo, podemos

avaliar a segunda parcela de (C.7), de acordo com a função densidade de probabilidade normal, e obter

$$\begin{aligned} s\phi(u) &= s \left( \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{u^2}{2} \right] \right) \\ &= s \left( \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{(f_{\min} - \hat{Y})^2}{2s^2} \right] \right). \end{aligned} \quad (\text{C.8})$$

Portanto, deve-se escolher a forma mais adequada (ótima relação custo computacional e aproximação numérica) para o cálculo da função exponencial, que será a responsável pelo maior tempo computacional do cálculo dessa parcela.

Para a primeira parcela de (C.7), suponha que a variável aleatória normal padrão em questão seja representada por  $T$ , então temos que

$$su\Phi(u) = su \int_{-\infty}^u \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{T^2}{2} \right] dT. \quad (\text{C.9})$$

A análise numérica dessa parcela pode ser realizada de diversas formas. A mais básica seria utilizar de alguma técnica de integração e tentar aproximar a integral imprópria que aparece na expressão. Porém, essa é uma escolha ruim, devido ao alto tempo computacional que gastaríamos para obter uma aproximação razoável para seu valor. Para os usuários do MatLab<sup>®</sup> ou outra ferramenta estatística, uma saída poderia ser definir uma distribuição aleatória normal padrão e automaticamente calcular o valor da função distribuição cumulativa de probabilidade. No MatLab<sup>®</sup> essa tarefa é realizada utilizando a função `cdf`.

Porém, durante a implementação numérica foi utilizada uma outra forma de cálculo, que foi capaz de economizar alguns segundos computacionais. Tomemos somente a integral imprópria em (C.9) e vamos utilizar a propriedade da soma de integrais e do valor da Integral Gaussiana, definida em (A.27) para encontrar que

$$\int_{-\infty}^u \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{T^2}{2} \right] dT = \int_{-\infty}^0 \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{T^2}{2} \right] dT$$



$$\begin{aligned}
& + \int_0^u \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{T^2}{2}\right] dT \\
& = \frac{1}{2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{T^2}{2}\right] dT \\
& \quad + \int_0^u \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{T^2}{2}\right] dT \\
& = \frac{1}{2} \frac{1}{\sqrt{2\pi}} \sqrt{2\pi} + \int_0^u \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{T^2}{2}\right] dT \\
& = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^u \exp\left[-\left(\frac{T}{\sqrt{2}}\right)^2\right] dT.
\end{aligned}$$

Fazendo  $q = \frac{T}{\sqrt{2}}$  temos  $\sqrt{2}dq = dT$ . Para os limites de integração, podemos ver que, quando  $T = 0$  então  $q = 0$ , e quando  $T = u$  então  $q = \frac{u}{\sqrt{2}}$ . Dessa forma, teremos que

$$\begin{aligned}
\int_{-\infty}^u \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{T^2}{2}\right] dT & = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^u \exp\left[-\left(\frac{T}{\sqrt{2}}\right)^2\right] dT. \\
& = \frac{1}{2} + \frac{1}{\sqrt{2}} \frac{1}{\sqrt{\pi}} \int_0^{\frac{u}{\sqrt{2}}} \exp(-q^2) \sqrt{2}dq. \\
& = \frac{1}{2} + \frac{1}{2} \frac{2}{\sqrt{\pi}} \int_0^{\frac{u}{\sqrt{2}}} \exp(-q^2) dq. \\
& = \frac{1}{2} \left(1 + \frac{2}{\sqrt{\pi}} \int_0^{\frac{u}{\sqrt{2}}} \exp(-q^2) dq\right).
\end{aligned}$$

Temos que a segunda parcela da soma dentro dos parêntesis na expressão anterior é a definição da Função Erro, também denotada por  $\text{erf}$ , ou seja, temos que

$$\text{erf}\left(\frac{u}{\sqrt{2}}\right) = \frac{2}{\sqrt{\pi}} \int_0^{\frac{u}{\sqrt{2}}} \exp(-q^2) dq. \quad (\text{C.10})$$

Portanto temos que a primeira parcela de (C.7) será dada por

$$\begin{aligned}
su\Phi(u) & = \frac{1}{2} su \left(1 + \text{erf}\left(\frac{u}{\sqrt{2}}\right)\right) \\
& = \frac{1}{2} (f_{\min} - \hat{Y}) \left(1 + \text{erf}\left[\frac{f_{\min} - \hat{Y}}{s\sqrt{2}}\right]\right). \quad (\text{C.11})
\end{aligned}$$

Substituindo (C.8) e (C.11) em (C.7) encontramos que

$$\mathbb{E}(I(\mathbf{d})) = \frac{(f_{\min} - \hat{Y})}{2} \left( 1 + \operatorname{erf} \left[ \frac{f_{\min} - \hat{Y}}{s\sqrt{2}} \right] \right) + \frac{s}{\sqrt{2\pi}} \exp \left[ -\frac{(f_{\min} - \hat{Y})^2}{2s^2} \right]. \quad (\text{C.12})$$

A equação (C.12) poderá ser calculada utilizando aproximações para as funções erro e exponencial. Ambas podem ser definidas em forma de séries de potências, o que facilita a implementação numérica e o tempo computacional de cálculo. Neste trabalho foram utilizadas as funções básicas `erf` e `exp`, já implementadas no MatLab<sup>®</sup>.

## APÊNDICE D – RESULTADOS ESTATÍSTICOS

Tabela 6 – Resultados obtidos na comparação entre o sEGO com e sem normalização. Minimizador P2 e  $\bar{\sigma}_0^2 = 0.01$ .

Função	Métrica	Com Normalização			Sem Normalização		
		Mediana	$\bar{J}_{\min}^{\text{melhor}}$	$\bar{J}_{\min}^{\text{pior}}$	Mediana	$\bar{J}_{\min}^{\text{melhor}}$	$\bar{J}_{\min}^{\text{pior}}$
<b>F7</b>	MQ	1.1378	1.0856	1.6865	1.1583	1.0841	2.3464
	AEI	1.1995	1.0820	1.6921	1.1713	1.0832	2.3469
	EQI	1.1853	1.0846	1.6851	1.1936	1.0807	2.3471
	TSSO	1.1837	1.0842	1.6933	1.3336	1.0856	2.8876
	EIR	1.3594	1.0920	2.0680	1.1980	1.0841	2.3477
<b>F9</b>	MQ	-16.5742	-16.6412	-14.6876	-14.3075	-16.6426	5.1165
	AEI	-16.5903	-16.6344	-16.0415	-13.9031	-16.6264	5.1158
	EQI	-16.5889	-16.6464	-16.1370	-13.8932	-16.6273	5.1166
	TSSO	-16.5591	-16.6442	-16.0543	-13.8857	-16.6415	5.1164
	EIR	-16.5533	-16.6373	-15.6506	-14.0029	-16.6414	5.1157
<b>F11</b>	MQ	0.0000	0.0000	0.2442	29.2303	0.0475	208.7176
	AEI	0.0955	0.0056	0.5473	23.3116	0.2189	1165.5331
	EQI	0.2524	0.0224	1.4297	20.3952	0.2266	317.0679
	TSSO	0.0176	0.0035	0.2526	26.0059	0.2268	406.7843
	EIR	1.1420	0.0442	6.3264	45.7770	4.6239	773.6387
<b>F13</b>	MQ	-3.0157	-3.2376	-1.9783	-2.9111	-3.2378	-1.7294
	AEI	-3.0537	-3.2427	-1.9535	-2.7976	-3.2396	-1.7273
	EQI	-2.8544	-3.2097	-1.9322	-2.8632	-3.2370	-1.7279
	TSSO	-2.9532	-3.2394	-1.9760	-2.8302	-3.2660	-1.7312
	EIR	-3.0559	-3.2674	-1.4072	-2.9016	-3.2374	-1.7294
<b>F15</b>	MQ	4.2464	0.0002	32.5217	74.2089	11.3434	532.9242
	AEI	3.6918	0.0573	27.7859	79.0368	11.4609	532.8973
	EQI	3.9640	0.1456	14.5498	80.2349	11.4779	532.9093
	TSSO	3.7163	0.0136	9.5383	79.0242	11.4770	532.9130
	EIR	5.0166	0.0421	11.6708	82.6871	9.6647	532.9137
<b>F17</b>	MQ	-2.9412	-3.1182	-2.1104	-2.6332	-3.2067	-1.6094
	AEI	-2.7890	-3.0942	-2.1569	-2.2847	-2.9122	-1.4566
	EQI	-2.5579	-3.0926	-1.5266	-2.1458	-2.9091	-1.1992
	TSSO	-2.8013	-3.0596	-2.5110	-2.3948	-2.9474	-1.5215
	EIR	-3.0478	-3.1133	-2.6434	-2.7342	-3.1402	-1.8061
<b>F18</b>	MQ	0.2066	0.0014	1.2689	20.0046	1.7965	36.8064
	AEI	0.9633	0.1321	2.3332	15.4422	3.3576	36.9859
	EQI	1.6439	0.4807	3.1596	19.0144	3.3803	36.7367
	TSSO	0.2259	0.0989	1.4945	10.6718	1.5788	35.0547
	EIR	0.5194	0.1251	1.7128	12.3466	2.5488	37.3076

Tabela 7 – Resultados obtidos em alguns problemas utilizando as cinco métricas e para três variâncias iniciais diferentes.

Função	Métrica	$\sigma_0^2$	Minimizador P1			Minimizador P2		
			Mediana	$\bar{J}_{\text{melhor}}^{\text{min}}$	$\bar{J}_{\text{pior}}^{\text{min}}$	Mediana	$\bar{J}_{\text{melhor}}^{\text{min}}$	$\bar{J}_{\text{pior}}^{\text{min}}$
F9	MQ	$10^{-1}$	-16.3652	-16.6361	-14.0643	-16.5764	-16.6418	-15.9998
		$10^{-2}$	-16.5629	-16.6429	-14.1526	-16.5742	-16.6412	-14.6876
		$10^{-4}$	-16.5127	-16.6375	-13.0977	-16.5475	-16.6444	-15.9499
	AEI	$10^{-1}$	-16.5128	-16.6326	-12.1877	-16.5800	-16.6422	-16.1594
		$10^{-2}$	-16.5615	-16.6413	-15.3997	-16.5903	-16.6344	-16.0415
		$10^{-4}$	-16.5780	-16.6436	-15.6171	-16.5846	-16.6427	-15.5110
	EQI	$10^{-1}$	-16.5417	-16.6392	-15.4060	-16.5393	-16.6375	-15.0579
		$10^{-2}$	-16.5750	-16.6431	-15.2221	-16.5889	-16.6464	-16.1370
		$10^{-4}$	-16.5480	-16.6436	-13.9216	-16.5786	-16.6432	-16.0226
	TSSO	$10^{-1}$	-16.5561	-16.6443	-16.0517	-16.5822	-16.6431	-15.9505
		$10^{-2}$	-16.5390	-16.6466	-11.6090	-16.5591	-16.6442	-16.0543
		$10^{-4}$	-16.5941	-16.6375	-15.7526	-16.5896	-16.6358	-16.0231
	EIR	$10^{-1}$	-16.6188	-16.6410	-16.0697	-16.5676	-16.6372	-15.1940
		$10^{-2}$	-16.5970	-16.6433	-15.2590	-16.5533	-16.6373	-15.6506
		$10^{-4}$	-16.5916	-16.6453	-14.1718	-16.5345	-16.6259	-15.1137
F13	MQ	$10^{-1}$	-3.0066	-3.2329	-1.8151	-3.0169	-3.2343	-1.9734
		$10^{-2}$	-3.0206	-3.2370	-1.9044	-3.0157	-3.2376	-1.9783
		$10^{-4}$	-2.8605	-3.2133	-1.8222	-2.8755	-3.2375	-1.7292
	AEI	$10^{-1}$	-3.0426	-3.2078	-1.8323	-2.9620	-3.2772	-1.9742
		$10^{-2}$	-3.0730	-3.2581	-1.9579	-3.0537	-3.2427	-1.9535
		$10^{-4}$	-2.8530	-3.1968	-0.2120	-2.8532	-3.2369	-1.7306
	EQI	$10^{-1}$	-2.7437	-3.2385	-0.4466	-2.8683	-3.2380	-1.9772
		$10^{-2}$	-2.7231	-3.2616	-0.4069	-2.8544	-3.2097	-1.9322
		$10^{-4}$	-2.8373	-3.2622	-1.8175	-2.8387	-3.2379	-1.7301
	TSSO	$10^{-1}$	-2.9465	-3.2483	-0.1269	-2.9514	-3.2784	-2.1031
		$10^{-2}$	-2.8522	-3.2389	-0.1421	-2.9532	-3.2394	-1.9760
		$10^{-4}$	-2.8953	-3.2231	-1.8749	-2.9451	-3.2363	-1.7290
	EIR	$10^{-1}$	-3.0870	-3.2271	-1.9932	-3.0647	-3.2712	-1.9766
		$10^{-2}$	-3.0630	-3.2714	-1.7274	-3.0559	-3.2674	-1.4072
		$10^{-4}$	-2.9326	-3.2366	-0.2245	-2.9087	-3.2680	-1.7290
F15	MQ	$10^{-1}$	5.0985	0.0002	60.6876	4.1437	0.0005	8.3448
		$10^{-2}$	4.3646	0.0001	32.5742	4.2464	0.0002	32.5217
		$10^{-4}$	7.6734	0.0145	43.2824	6.0824	0.0235	38.0134
	AEI	$10^{-1}$	4.1092	0.1078	21.9157	2.9960	0.1627	8.8761
		$10^{-2}$	4.2570	0.0134	27.8121	3.6918	0.0573	27.7859
		$10^{-4}$	8.2839	0.1534	42.6207	6.3525	0.1458	30.7469
	EQI	$10^{-1}$	3.8433	0.0072	9.4688	3.4443	0.0406	9.3594
		$10^{-2}$	3.9836	0.1726	14.2241	3.9640	0.1456	14.5498

		$10^{-4}$	8.3304	0.1025	37.2112	7.5988	0.1935	34.8216	
		<b>TSSO</b>	$10^{-1}$	4.5948	0.0113	18.9760	3.3328	0.0234	13.6895
			$10^{-2}$	5.7635	0.0299	82.1644	3.7163	0.0136	9.5383
	<b>EIR</b>	$10^{-4}$	9.1437	0.0600	102.4360	8.1100	0.0402	46.2685	
		$10^{-1}$	5.0728	0.2805	28.2271	4.8900	0.4575	10.7356	
		$10^{-2}$	5.9708	0.4229	20.1442	5.0166	0.0421	11.6708	
	<b>F17</b>	<b>MQ</b>	$10^{-4}$	5.1987	0.7721	78.2459	5.2085	0.6280	21.9807
			$10^{-1}$	-2.9762	-3.1198	-1.9994	-2.9831	-3.1211	-2.1202
			$10^{-2}$	-2.9329	-3.1194	-1.0216	-2.9412	-3.1182	-2.1104
<b>AEI</b>		$10^{-4}$	-2.5416	-3.0561	-0.8901	-2.5072	-3.0523	-1.4954	
		$10^{-1}$	-2.9907	-3.1178	-2.1312	-2.8503	-3.1049	-2.1447	
		$10^{-2}$	-2.8349	-3.1040	-1.5594	-2.7890	-3.0942	-2.1569	
<b>EQI</b>		$10^{-4}$	-2.2789	-3.0270	-0.2458	-2.2855	-2.9959	-1.6003	
		$10^{-1}$	-2.6464	-3.0419	-1.3709	-2.5360	-2.9758	-1.3580	
		$10^{-2}$	-2.5970	-3.1044	-1.5690	-2.5579	-3.0926	-1.5266	
<b>TSSO</b>		$10^{-4}$	-2.2237	-2.9547	-0.1999	-2.2075	-2.7258	-1.3332	
		$10^{-1}$	-2.8758	-3.1120	-2.4243	-2.8717	-3.1118	-2.5442	
		$10^{-2}$	-2.7978	-3.0616	-1.5929	-2.8013	-3.0596	-2.5110	
<b>EIR</b>		$10^{-4}$	-2.5480	-3.0019	-0.1063	-2.4539	-2.9632	-1.4254	
		$10^{-1}$	-3.0893	-3.1139	-2.6758	-3.0607	-3.1133	-2.6434	
		$10^{-2}$	-3.0851	-3.1187	-2.6654	-3.0478	-3.1133	-2.6434	
<b>F18</b>		<b>MQ</b>	$10^{-4}$	-2.6998	-3.0788	-2.2393	-2.6210	-3.0714	-2.1837
			$10^{-1}$	0.2044	0.0002	2.5571	0.2062	0.0004	1.8514
			$10^{-2}$	0.2469	0.0013	1.7091	0.2066	0.0014	1.2689
	<b>AEI</b>	$10^{-4}$	0.2398	0.0001	2.0613	0.2345	0.0002	1.8541	
		$10^{-1}$	1.1333	0.1117	3.9568	1.0694	0.2567	3.3766	
		$10^{-2}$	1.2646	0.2261	3.9631	0.9633	0.1321	2.3332	
	<b>EQI</b>	$10^{-4}$	0.4127	0.0830	3.9721	0.3703	0.0437	1.5321	
		$10^{-1}$	2.7299	0.4403	6.0595	1.9028	0.6170	2.9428	
		$10^{-2}$	2.5745	0.5194	6.4648	1.6439	0.4807	3.1596	
	<b>TSSO</b>	$10^{-4}$	0.5617	0.0618	5.9555	0.4283	0.1234	3.6315	
		$10^{-1}$	0.2988	0.0290	1.9261	0.3008	0.0421	1.2124	
		$10^{-2}$	0.2231	0.0618	1.8303	0.2259	0.0989	1.4945	
	<b>EIR</b>	$10^{-4}$	0.3745	0.0236	2.9439	0.2726	0.0330	1.0547	
		$10^{-1}$	0.5912	0.1042	5.1058	0.3874	0.1638	1.8541	
		$10^{-2}$	0.6413	0.2172	3.8969	0.5194	0.1251	1.7128	
			$10^{-4}$	0.9880	0.2432	7.5624	0.4129	0.1780	1.3079

Tabela 8 – Escores acumulados de cada uma das cinco métricas com os dois minimizadores e para os dezoito problemas analisados.

Funções	Métricas									
	MQ1	MQ2	AEI1	AEI2	EQI1	EQI2	TSSO1	TSSO2	EIR1	EIR2
<b>F1</b>	1.00	0.55	0.50	0.48	0.62	0.50	0.47	0.35	0.51	0.23
<b>F2</b>	2.00	1.07	1.01	0.86	1.25	0.96	1.24	0.90	0.90	0.62
<b>F3</b>	3.00	1.44	1.47	1.11	2.01	1.36	1.82	1.38	1.20	0.96
<b>F4</b>	3.98	2.33	2.31	1.69	2.91	2.24	2.78	2.38	1.60	1.25
<b>F5</b>	4.56	2.63	3.19	2.47	3.91	2.74	3.75	2.93	2.04	1.85
<b>F6</b>	5.38	3.31	3.89	3.12	4.69	3.11	4.75	3.37	2.86	2.54
<b>F7</b>	5.77	3.93	4.85	4.09	5.62	4.05	5.71	4.33	3.83	3.54
<b>F8</b>	6.77	3.98	5.85	4.15	5.71	4.08	5.80	4.40	3.86	3.69
<b>F9</b>	7.77	4.22	6.13	4.39	6.28	4.49	6.75	4.77	4.11	4.11
<b>F10</b>	7.86	4.22	6.42	4.45	6.72	4.63	6.87	4.78	5.11	4.87
<b>F11</b>	7.87	4.22	6.48	4.48	7.72	4.72	7.00	4.79	5.66	5.44
<b>F12</b>	8.29	4.72	6.81	4.76	8.29	5.28	8.00	5.61	6.03	5.69
<b>F13</b>	8.85	5.17	7.34	5.27	9.29	5.92	8.75	6.12	6.51	6.04
<b>F14</b>	9.77	6.09	8.26	6.18	10.20	6.79	9.66	7.01	7.51	6.93
<b>F15</b>	10.69	6.91	9.05	6.95	11.07	7.56	10.50	7.77	8.51	7.69
<b>F16</b>	11.22	7.27	9.43	7.37	12.07	8.37	11.12	8.38	8.89	8.02
<b>F17</b>	11.64	7.49	9.68	7.66	13.07	9.13	11.65	8.92	9.14	8.21
<b>F18</b>	11.78	7.62	10.33	8.06	14.07	9.69	11.77	9.02	9.56	8.51