Adaptive Multiple Importance Sampling for Gaussian Processes: Finding Difference Makers in Personality Impressions

Xiaoyu Xiong[1], Maurizio Filippone[2], Alessandro Vinciarelli[1]

1-The University of Glasgow email: x.xiong.1@research.gla.ac.uk, Alessandro.Vinciarelli@glasgow.ac.uk

2- EURECOM, Sophia Antipolis, France email: maurizio.filippone@eurecom.fr

Motivation

- ✓ Traditional parametric approaches for Social Signal Processing (SSP):

  falls short in taking into account uncertainty due to **model misspecification** or **overfitting**
- ✓ Current solutions: Bayesian treatment with nonparametric models like Gaussian Processes (GPs)
- ✓ GPs requires accurately characterizing posterior over covariance parameters:

  this is normally done by means of Markov chain Monte Carlo (MCMC) methods
- ✓ However, MCMC for learning GP parameters are inefficient due to its rejection of expensive proposals
- ✓ In this work, we propose an alternative inference framework for GPs based on Adaptive Multiple Importance Sampling (AMIS).
- ✓ The experimental results suggests AMIS is competitive with MCMC for GP models and suitable for SSP

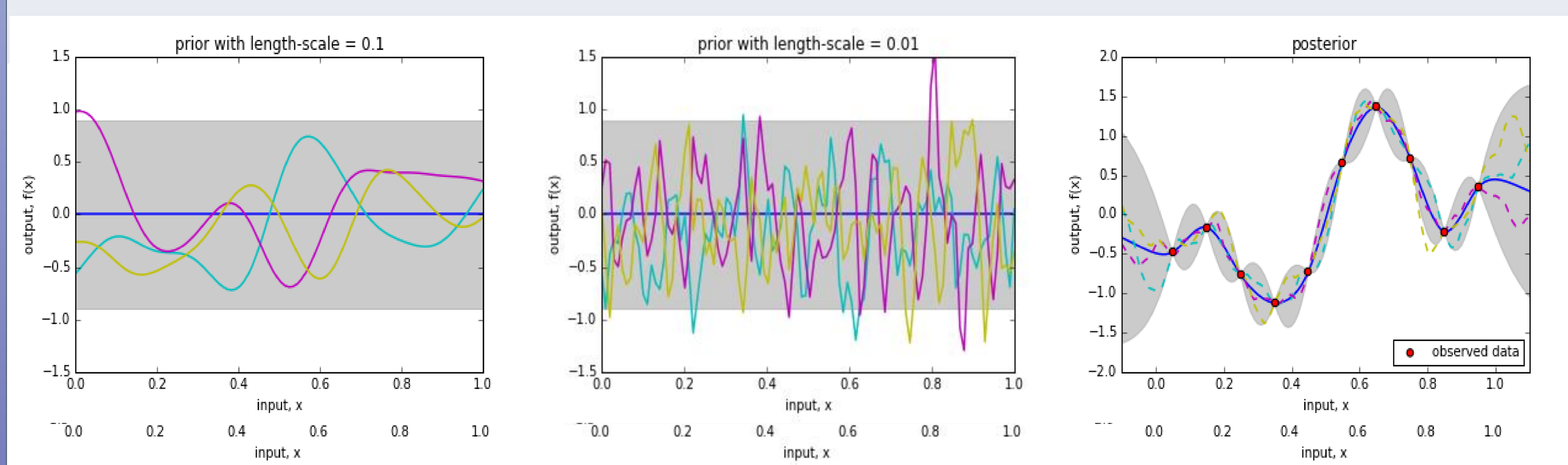
Bayesian Gaussian processes

Gaussian process assumption $p(\boldsymbol{f}(X)|\boldsymbol{\theta}) \sim \mathcal{N}(0, K(\boldsymbol{\theta}))$

- ✓ Gaussian likelihood $p(\boldsymbol{y}|\boldsymbol{\theta}) \sim p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{\theta}) \sim \mathcal{N}(0, K(\boldsymbol{\theta}) + \lambda \mathbf{I})$ where λ is the variance of the Gaussian noise on $\boldsymbol{y}$, $\mathbf{I}$ is the identity matrix

  Posterior $\pi(\boldsymbol{\theta}) := p(\boldsymbol{\theta}|\boldsymbol{y}, X) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$
- ✓ Non-Gaussian likelihood $p(\boldsymbol{y}|\boldsymbol{\theta}) = \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{\theta})d\boldsymbol{f}$, $\tilde{p}(\boldsymbol{y}|\boldsymbol{\theta}) \simeq \frac{1}{N_{imp}}\sum_{i=1}^{N_{imp}} \frac{p(\boldsymbol{y}|\boldsymbol{f}_i)p(\boldsymbol{f}_i|\boldsymbol{\theta})}{q(\boldsymbol{f}_i|\boldsymbol{\theta},\boldsymbol{y})}$ (estimated)

  Posterior $\pi(\boldsymbol{\theta}) := p(\boldsymbol{\theta}|\boldsymbol{y}, X) \propto \tilde{p}(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$



The entries of covariance K is determined by a kernel function

RBF kernel: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma e^{-\frac{1}{\tau^2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}$, where

σ is the marginal variance of the function values at input locations $\boldsymbol{x}$, τ is the length-scale, controlling the smoothness of functions

ARD kernel: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma e^{-\sum_{r=1}^{d} \frac{1}{\tau_r^2}[\boldsymbol{x}_{i(r)} - \boldsymbol{x}_{j(r)}]^2}$, τ_r is the length-scale of each feature of $\boldsymbol{x}$

Inference in GPs is about learning the kernel parameters, e.g., $\boldsymbol{\theta} = (\sigma, \tau)$ (RBF) or $\boldsymbol{\theta} = (\sigma, \tau_r)$ (ARD)

✓ **Computing expectation using MCMC**

$$E_{\pi(\boldsymbol{\theta})}h(\boldsymbol{\theta}) = \int h(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} = \frac{1}{N}\sum_{i=0}^{N} h(\boldsymbol{\theta}^i)$$

✓ **Computing expectation using Importance Sampling (IS)**

$$E_{\pi(\boldsymbol{\theta})}h(\boldsymbol{\theta}) = \int h(\boldsymbol{\theta})\frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}q(\boldsymbol{\theta})d\boldsymbol{\theta} = \frac{1}{N}\sum_{i=0}^{N} h(\boldsymbol{\theta}^i)\boldsymbol{w}_i, \quad \boldsymbol{w}_i = \frac{\pi(\boldsymbol{\theta}^i)}{q(\boldsymbol{\theta}^i)}$$

(Importance weight)

MCMC for learning $\boldsymbol{\theta}$

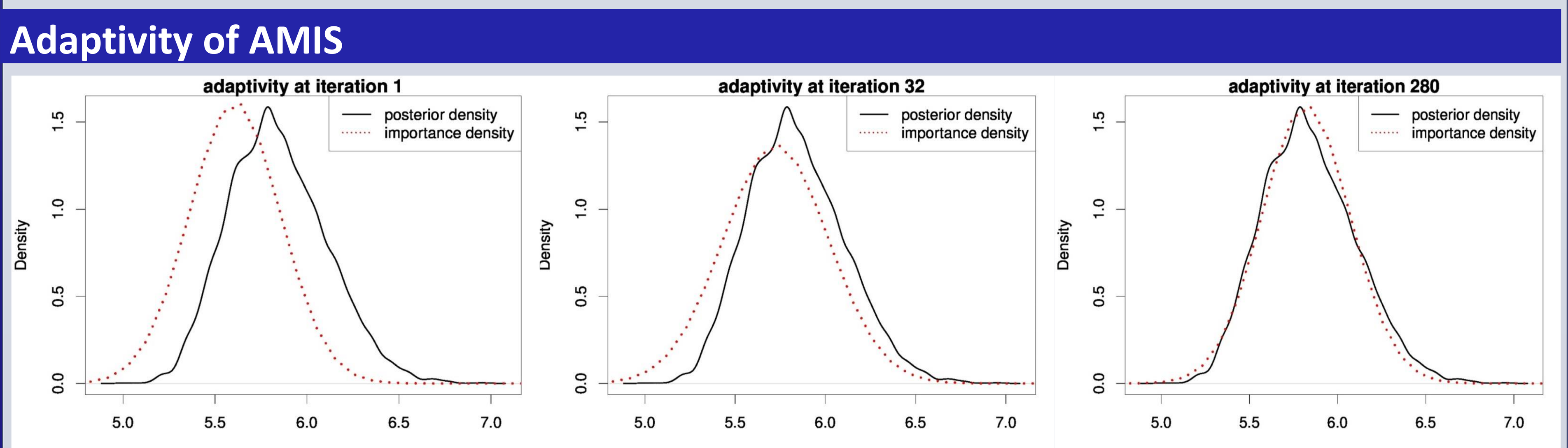
Hastings ratio:

$\frac{p(\boldsymbol{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}$ (MH for Gaussian likelihood) $\rightarrow$ $\frac{\tilde{p}(\boldsymbol{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{\tilde{p}(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}$ (Pseudo-Marginal MH for non-Gaussian likelihood)

Motivation for alternative sampling methods:

- ✓ In GPs with Gaussian likelihood, computing the marginal likelihood and its gradient with respect to $\boldsymbol{\theta}$ is expensive and standard MCMC algorithms reject proposals leading to a waste of computations
- ✓ In GPs with non-Gaussian likelihood, PM-MH may further cause inefficiencies due to large overestimation of the marginal likelihood

Adaptivity of AMIS



✓ AMIS adaptively constructs an approximate posterior over $\boldsymbol{\theta}$ used to build an increasingly more accurate importance estimator :

$$\frac{1}{\sum_{t=0}^{T-1}\sum_{i=1}^{N_t} w_i{}^t}\sum_{t=0}^{T-1}\sum_{i=1}^{N_t} w_i{}^t\, h(\boldsymbol{\theta}_i{}^t) \qquad \text{where}$$

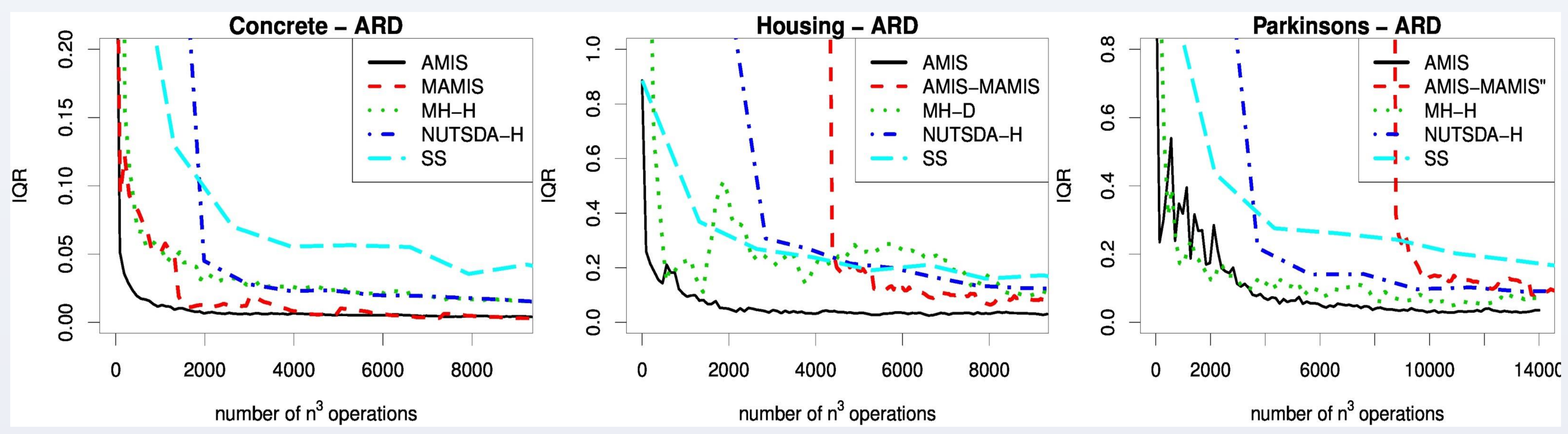
$w_i{}^t = g(\boldsymbol{\theta}_i^t) \Big/ \frac{1}{\sum_{t=0}^{T-1} N_t}\sum_{t=0}^{T-1} N_t q_t(\boldsymbol{\theta}_i{}^t;\gamma_t)$, $g(\boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ weight of **AMIS** for Gaussian likelihood

$w_i{}^t = \tilde{g}(\boldsymbol{\theta}_i^t) \Big/ \frac{1}{\sum_{t=0}^{T-1} N_t}\sum_{t=0}^{T-1} N_t q_t(\boldsymbol{\theta}_i{}^t;\gamma_t)$, $\tilde{g}(\boldsymbol{\theta}) = \tilde{p}(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ weight of **PM-AMIS** for non-Gaussian likelihood

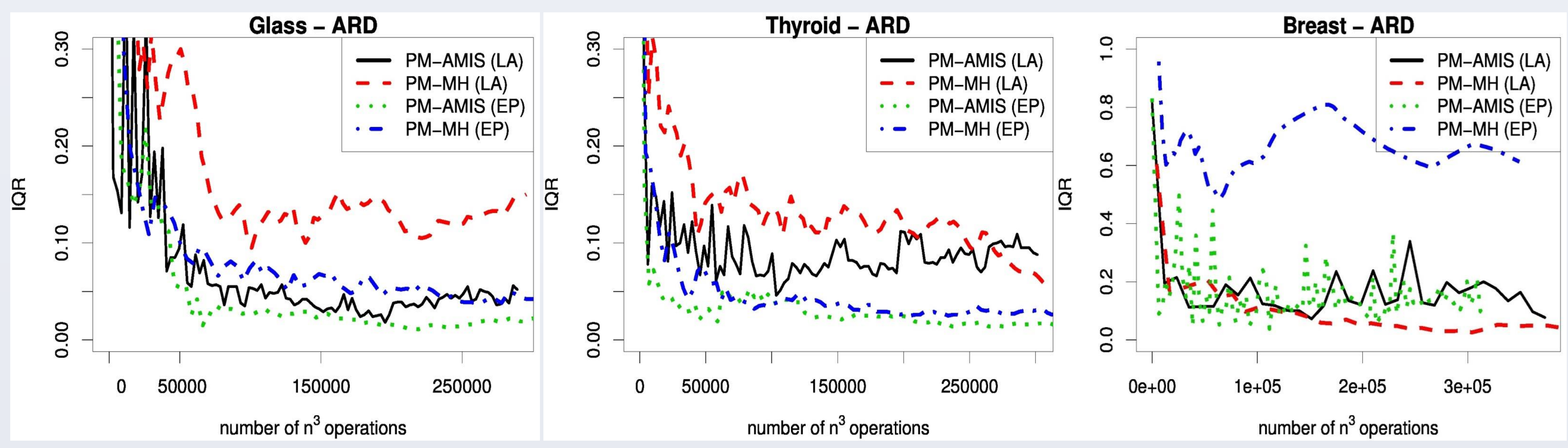
Experiments

- ✓ IQR: Interquartile Range of the expectation of the norm of the parameters
- ✓ Convergence analysis: IQR against computational costs (number of n^3 operations)

GP regression - AMIS vs MH, Best of HMC family, slice sampling (SS)



GP classification - PM-AMIS vs PM-MH



Application of PM-AMIS for GPs: Personality inference from Flickr images

- ✓ **Data:** "fave", i.e., pictures that Flickr users have tagged as favourite, with personality traits attributed by Asian and UK assessors
- ✓ **Task:** predicting whether a Flickr user is perceived to be above median with respect to the Big-Five traits – Openness (Ope), Conscientiousness (Con), Extraversion (Ext), Agreeableness (Agr), Neuroticism (Neu)
- ✓ **Approaches:** Support Vector Machine (SVM) and PM-AMIS for GPs with G-ARD kernel:
  - $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma e^{-\sum_{r=1}^{N_g}\frac{1}{N_r\tau_r^2}\{\sum_{s\in\mathcal{G}_r}[\boldsymbol{x}_{i(s)} - \boldsymbol{x}_{j(s)}]^2\}}$
  - τ_r is the length-scale parameter for group r, N_r is the number of features in group r, N_g is the number of groups, $\mathcal{G}_r$ is the set of indexes of the features that belong to group r

Table 1: Groups of features (number in brackets denotes the number of features for each group)

G1	G2	G3	G4	G5	G6	G7	G8	G9
Faces (1)	Colour Properties(3)	Colour Distribution(1)	Homogeneous Regions (4)	Composition (2)	Texture Wavelets (12)	GIST filters (24)	GLCM (24)	Texture Statistics (4)

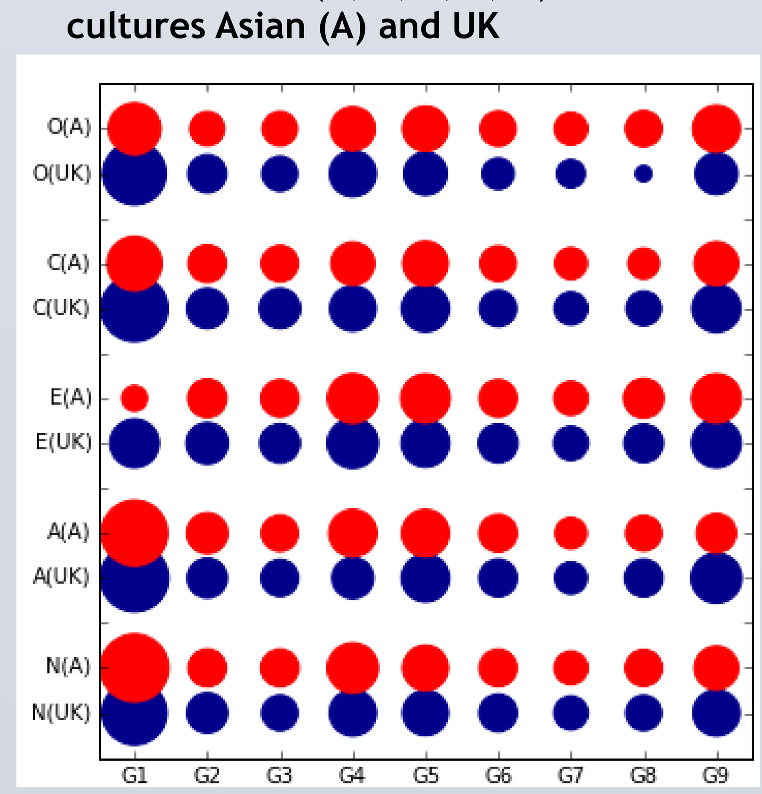
✓ **Results:**

- PM-AMIS for GPs achieves comparable accuracies with SVM (see Table 2)
- G-ARD is able to identify the groups of features (G1, G5, G9, G4) that mostly influence personality impression
- Weights difference (Figure 1) across Asian and UK personality assessors suggests cultural difference on personality perception

Table 2: Prediction Accuracy

	Ope	Con	Ext	Agr	Neu
PM-AMIS for GPs (UK)	65%	58%	71%	73%	79%
SVM(UK)	59%	62%	71%	74%	77%
PM-AMIS for GPs (Asia)	68%	52%	74%	68%	69%
SVM(Asia)	68%	47%	68%	69%	70%

Figure 1: Coefficients of the G-ARD for the five traits (O, C, E, A, N) and two cultures Asian (A) and UK



Conclusions

- ✓ AMIS achieves faster convergence than MCMC for GPs while being easy to tune and implement and facilitating massive parallelization
- ✓ AMIS for GPs offers an efficient probabilistic framework for SSP

References

[1] Xiong, X., Šmídl, V. and Filippone, M. Adaptive Multiple Importance Sampling for Gaussian Processes, eprint arXiv:1508.01050v2, 2016.

[2] Segalin, C., Perina, A., Cristani, M. and Vinciarelli, A. The pictures we like are our image: continuous mapping of favourite pictures into self-assessed and attributed personality traits. IEEE Transactions on Affective Computing (to appear), 2016.

[3] Filippone, M., Girolami, M. Pseudo-marginal Bayesian inference for Gaussian processes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014.

[4] Cornuet, J.-M., Marin, J.-M., Mira, A., Robert, C. P. Adaptive multiple importance sampling. Scandinavian Journal of Statistics 39, 798-812, 2012.