A Appendix to "Efficient Interpolation of Computationally Expensive Posterior Densities with Variable Parameter Costs", by Nikolay Bliznyuk, David Ruppert and Christine A. Shoemaker, published in the *Journal of Computational and Graphical Statistics*

A.1 Proofs

The notation used in these proofs was introduced in Section 2.

Let l be a GP indexed by η, with mean 0 and covariance function C_η. Let $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ be any finite disjoint sets of values of η. Define $\Sigma_{ij} = C_\eta(\mathcal{E}_i, \mathcal{E}_j)$ for $1 \le i, j \le 3$. Since all finite-dimensional distributions of l are Gaussian, it is seen that

- $$E[l(\mathcal{E}_i)|l(\mathcal{E}_j)] = \Sigma_{ij}\Sigma_{jj}^{-1} l(\mathcal{E}_j), \tag{15}$$
- $$Var\{E[l(\mathcal{E}_i)|l(\mathcal{E}_j)]\} = \Sigma_{ij}\Sigma_{jj}^{-1}\Sigma_{ji}, \tag{16}$$
- $$l(\mathcal{E}_i)|l(\mathcal{E}_j) \sim MVN\left(E[l(\mathcal{E}_i)|l(\mathcal{E}_j)], \Sigma_{ii} - Var\{E[l(\mathcal{E}_i)|l(\mathcal{E}_j)]\}\right) \tag{17}$$

Notice that $Var[l(\mathcal{E}_i)|l(\mathcal{E}_j)]$ is the variance of the error from prediction of $l(\mathcal{E}_i)$ using the BLUP $E[l(\mathcal{E}_i)|l(\mathcal{E}_j)]$.

Proposition 1: *Under the above assumptions,*

$$Var[l(\mathcal{E}_1)] \ge Var[l(\mathcal{E}_1)|l(\mathcal{E}_3)] \ge Var[l(\mathcal{E}_1)|l(\mathcal{E}_2), l(\mathcal{E}_3)],$$

where $A \ge B$ iff $A - B$ is non-negative definite.

Proof: Follows from properties of a multivariate normal density.

Proposition 2: *Let $\mathcal{B}$ be a finite set of β-points and $\mathcal{Z}$ be a finite set of ζ-points. Define $\mathcal{Z}^* = \{\zeta^*\} \cup \mathcal{Z}$. If the covariance function for l is separable in a sense of Equation (4), then*

$$E[l([\beta^*, \zeta^*])|l(\mathcal{B} \oplus \mathcal{Z}^*)] = E[l([\beta^*, \zeta^*])|l(\mathcal{B} \oplus \zeta^*)]. \tag{18}$$

Proof: Without loss of generality, assume that $\sigma^2 = 1$ (because of Equation (15)), and that ζ^* is the first element of the list $\mathcal{Z}^*$.

Notice that, under the assumed separability,

- $Var[l(\mathcal{B} \oplus \mathcal{Z}^*)] = C_\eta(\mathcal{B} \oplus \mathcal{Z}^*, \mathcal{B} \oplus \mathcal{Z}^*) = C_\beta(\mathcal{B}, \mathcal{B}) \otimes C_\zeta(\mathcal{Z}^*, \mathcal{Z}^*)$, where $\otimes$ is the Kronecker product,
- $Cov[l(\mathcal{B} \oplus \mathcal{Z}^*), l(\beta^* \oplus \zeta^*)] = C_\eta(\mathcal{B} \oplus \mathcal{Z}^*, \beta^* \oplus \zeta^*) = C_\beta(\mathcal{B}, \beta^*) \otimes C_\zeta(\mathcal{Z}^*, \zeta^*)$.

•

$$[Var(l(\mathcal{B} \oplus \mathcal{Z}^*))]^{-1} Cov[l(\mathcal{B} \oplus \mathcal{Z}^*), l(\beta^* \oplus \zeta^*)] =$$
$$C_\beta(\mathcal{B}, \mathcal{B})^{-1} C_\beta(\mathcal{B}, \beta^*) \otimes C_\zeta(\mathcal{Z}^*, \mathcal{Z}^*)^{-1} C_\zeta(\mathcal{Z}^*, \zeta^*) =$$
$$C_\beta(\mathcal{B}, \mathcal{B})^{-1} C_\beta(\mathcal{B}, \beta^*) \otimes e_1,$$

where e_1 is the first standard basis vector for $(|\mathcal{Z}^*| + 1)$-dimensional vector space. [This is true since $C_\zeta(\mathcal{Z}^*, \zeta^*)$ is the first column of $C_\zeta(\mathcal{Z}^*, \mathcal{Z}^*)$, by definition of $\mathcal{Z}^*$.]

Therefore, $E[l(\eta^*)|l(\mathcal{B} \oplus \mathcal{Z}^*)]$

$$= l(\mathcal{B} \otimes \mathcal{Z}^*)^\mathsf{T} C_\beta(\mathcal{B}, \mathcal{B})^{-1} C_\beta(\mathcal{B}, \beta^*) \otimes e_1$$
$$= [C_\beta(\mathcal{B}, \beta^*)^\mathsf{T} C_\beta(\mathcal{B}, \mathcal{B})^{-1} \otimes e_1^\mathsf{T}] \cdot l(\mathcal{B} \otimes \mathcal{Z}^*)$$
$$= vec\{e_1^\mathsf{T} \cdot unvec[l(\mathcal{B} \oplus \mathcal{Z}^*)] \cdot C_\beta(\mathcal{B}, \mathcal{B})^{-1} C_\beta(\mathcal{B}, \beta^*)\}$$
$$= l(\mathcal{B} \oplus \zeta^*)^\mathsf{T} C_\beta(\mathcal{B}, \mathcal{B})^{-1} C_\beta(\mathcal{B}, \beta^*).$$

In this equation, $vec(\cdot)$ is the vectorization operator defined for a $m \times n$ matrix A as $vec(A) = [A_1^\mathsf{T}, \ldots, A_n^\mathsf{T}]^\mathsf{T}$, where A_i is the ith column of A. The jth column of $unvec[l(\mathcal{B} \oplus \mathcal{Z}^*)]$ is the column vector $l(\mathcal{B} \oplus \zeta^{(j)})$, where $\zeta^{(j)}$ is the jth element of $\mathcal{Z}^*$. We are using the identity $vec(ABC) = (C^\mathsf{T} \otimes A) \cdot vec(B)$ for any matrices A, B, C of such dimensions that the product ABC is defined (Harville 1997, chap. 16).

The proof follows by observing that $E[l(\eta^*)|l(\mathcal{B} \oplus \mathcal{Z}^*)]$ does not depend on $\mathcal{Z}$, and is equal to $E[l(\eta^*)|l(\mathcal{B} \oplus \zeta^*)]$, which can be verified by taking $\mathcal{Z}$ to be an empty set.

A.2 Details of Inference and Fitting

In the first subsection, we outline the exact steps that need to be taken to carry out Bayesian inference using the proposed interpolants. The later subsections provide details for fitting `DOSKA` and `INDA`.

A.2.1 Steps for Bayesian Inference with the Proposed Interpolants

This section is an adaptation of the procedure of Bliznyuk et al (2008) for Bayesian inference using RBF approximation to computationally expensive posterior density.

Step 1. Select the β-knots on a HPD region for β. For details, see Bliznyuk et al (2008) and Bliznyuk et al (2010, submitted).

Step 2. Fit `DOSKA` as discussed in Section 3 or `INDA` as discussed in Sections 2.2-2.4 or Appendix A.2.3.
Use the interpolants of the full log-posterior to define the approximate posterior density via Equation 2.

Step 3. Sample $\widetilde{\pi}$ using an MCMC algorithm.

A.2.2 Fitting of DOSKA

Unlike many popular RBF interpolants that involve no basis function parameters, successful application of kriging requires estimation of parameters θ of the correlation function C_β. In this section we review two methods of estimation, maximum likelihood and K-fold cross-validation. We assume that one has (i) knots $\mathcal{D} = \mathcal{B} \oplus \mathcal{Z}$ in a high-probability region of π and (ii) values of l at these points. For consistency with the assumption of zero mean Gaussian process made about l, we re-center l by subtracting from it the mean of $l(\mathcal{B} \oplus \mathcal{Z})$, as was done in Rasmussen (2003). This does not influence the interpretation of the log-posterior l since it is only known up to an additive constant.

The assumption that l is a realization of a Gaussian process allows one to write down the likelihood of $l(\mathcal{B} \oplus \mathcal{Z})$. This is a multivariate normal density with mean 0 and covariance matrix $\sigma^2 \cdot C_\beta(\mathcal{B}, \mathcal{B}) \otimes C_\zeta(\mathcal{Z}, \mathcal{Z})$, by separability of C_η and our choice of knots $\mathcal{D}$. Thanks to the Kronecker product representation, the log-likelihood can be evaluated efficiently.

An alternative K-fold cross-validation criterion (KfCV) reuses subsets of the "data" $l(\mathcal{D})$ for validation, thereby guarding against overfitting. In our setting, its form is

$$F(\theta) := \sum_{i=1}^{K} \|\widetilde{l}_{i,\theta}(\mathcal{B}_i \oplus \mathcal{Z}) - l(\mathcal{B}_i \oplus \mathcal{Z})\|_F^2, \tag{19}$$

where

$$\widetilde{l}_{i,\theta}([\beta^*, \zeta^*]) := C_\beta(\beta^*, \mathcal{B}_{-i}; \theta) \cdot C_\beta(\mathcal{B}_{-i}, \mathcal{B}_{-i}; \theta)^{-1} \cdot l(\mathcal{B}_{-i} \oplus \zeta^*), \tag{20}$$

$\{\mathcal{B}_1, \ldots, \mathcal{B}_K\}$ is a partition of $\mathcal{B}$, $\mathcal{B}_{-i} := \mathcal{B} \backslash \mathcal{B}_i$ is the set difference and, for a matrix A, $\|A\|_F^2 = \sum_{i,j} A_{i,j}^2$ (squared Frobenius norm of A). To compute $\widetilde{l}_{i,\theta}(\mathcal{B}_i \oplus \mathcal{Z})$ for a given value of θ, it is necessary to obtain a factorization of $C_\beta(\mathcal{B}_{-i}, \mathcal{B}_{-i}; \theta)$ and to evaluate $l(\mathcal{B}_{-i} \oplus \zeta)$ for all $\zeta \in \mathcal{Z}$. The overall cost of factorizing $C_\beta(\mathcal{B}_{-i}, \mathcal{B}_{-i}; \theta)$ for all i can be made equal to a small multiple of $|\mathcal{B}|^3$, as opposed to $\mathcal{O}(K \cdot |\mathcal{B}|^3)$ in a naïve implementation, if one computes QR or Cholesky factorizations of $C_\beta(\mathcal{B}_{-i}, \mathcal{B}_{-i}; \theta)$ by downdating a single factorization of $C_\beta(\mathcal{B}, \mathcal{B}; \theta)$ for each i (Golub and Van Loan, 1996, sec. 12.5). For example, for the choice $K = |\mathcal{B}|/4$ that we use, computational savings can be enormous if $|\mathcal{B}|$ is large.

Many of the popular correlation functions are differentiable in θ, and so both F and the negative of the log-likelihood function can be minimized efficiently by numerical optimization. In our experiments, both of these criteria often had multiple minimizers, so multiple starting points for optimization are necessary.

In preliminary experiments to determine which method requires the fewest knots for a given level of accuracy, we had more success with KfCV. In particular, on Rasmussen's test problem 2 discussed below in Section 4, for higher values of $\dim(\beta)$, the MLE required roughly twice as many β-knots as KfCV. For this reason, we use the KfCV criterion in the experiments of this paper.

A.2.3 Fitting INDA using RBF Interpolation

We now describe the procedure for fitting INDA using RBF interpolation with the cubic basis function and a linear polynomial tail $q(\boldsymbol{\beta}) = (1, \boldsymbol{\beta}^\mathsf{T}) \cdot \boldsymbol{c}$. Discussion of fitting for

other choices of basis functions is in Powell (1996).

Define the matrix $\boldsymbol{\Phi} \in \mathbb{R}^{N\times N}$ by: $\boldsymbol{\Phi}_{i,j} = \phi(\|\beta^{(i)} - \beta^{(j)}\|_2)$, for $i, j = 1, \ldots, N$. Let $\boldsymbol{P} \in \mathbb{R}^{N\times(p+1)}$ be the matrix with $(1, \{\beta^{(i)}\}^{\mathsf{T}})$ as the ith row for $i = 1, \ldots, N$. The coefficients for the RBF surface that interpolates $G_{E,k}$, the kth component of G_E, at the points $\beta^{(1)}, \ldots, \beta^{(N)}$ are obtained by solving the system

$$\begin{pmatrix} \boldsymbol{\Phi} & \boldsymbol{P} \\ \boldsymbol{P}^{\mathsf{T}} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{c} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mathcal{G}}_k \\ \mathbf{0} \end{pmatrix}, \tag{21}$$

where $\boldsymbol{\mathcal{G}}_k = \left[G_{E,k}(\beta^{(1)}), \ldots, G_{E,k}(\beta^{(N)})\right]^{\mathsf{T}}$, $\boldsymbol{a} \in \mathbb{R}^N$ and $\boldsymbol{c} \in \mathbb{R}^{p+1}$.

The interpolation matrix on the left-hand side of equation (21) is invertible if and only if the rank of $\boldsymbol{P}$ is $p+1$ (Powell 1992). For the case of a cubic RBF with a linear tail, this holds if and only if the set of (distinct) design points contains $p+1$ points that are *affinely independent*. For stability purposes, we solve equation (21) by means of matrix factorizations, as described in Powell (1996).

Notice that, for all k, the linear systems of equations have the same interpolation matrices, and, consequently, only a single matrix factorization is required to simultaneously solve the interpolating equations for multiple right-hand sides.

A.3 Definitions of the Posterior Densities and of G_E

From Section 5 we have that $[Y|w, \gamma] = MVN(Hw, \Sigma_Y)$ and that $[w|\gamma] = MVN(\mu_w, \Sigma_w)$. The prior $[\gamma]$ will be specified later.

Using the standard trick of completing the square, we get that

$$-2 \cdot \log[w|Y, \gamma] + c = (w-b)^{\mathsf{T}}\Psi^{-1}(w-b) - b^{\mathsf{T}}\Psi^{-1}b + Y^{\mathsf{T}}\Sigma_Y^{-1}Y + \mu_w^{\mathsf{T}}\Sigma_w^{-1}\mu_w,$$

where

$$\Psi^{-1} = H^{\mathsf{T}}\Sigma_Y^{-1}H + \Sigma_w^{-1} \tag{22}$$

and

$$b = \left\{H^{\mathsf{T}}\Sigma_Y^{-1}H + \Sigma_w^{-1}\right\}^{-1}(H^{\mathsf{T}}\Sigma_Y^{-1}Y + \Sigma_w^{-1}\mu_w), \tag{23}$$

so that $[w|\gamma, Y]$ is $MVN(b, \Psi)$.

By using the identity

$$[Y, w, \gamma] = [Y|w, \gamma][w|\gamma][\gamma] = [w|\gamma, Y][\gamma|Y][Y]$$

and manipulating the expressions, it can be seen that

$$\begin{aligned} \log[\gamma|Y] = c + \log[\gamma] - \frac{1}{2}\log|\Sigma_Y| - \frac{1}{2}\log|\Sigma_w| - \frac{1}{2}\log|\Psi^{-1}| \\ - \frac{1}{2}\left(-b^{\mathsf{T}}\Psi^{-1}b + Y^{\mathsf{T}}\Sigma_Y^{-1}Y + \mu_w^{\mathsf{T}}\Sigma_w^{-1}\mu_w\right) \end{aligned} \tag{24}$$

where c do not depend on w or γ. We assume that $\mu_w = 0$, so the last term in $\log[\gamma|Y]$ vanishes.

To finish the specification of the model, we put a uniform priors on σ's (non-uniform on σ^2) as suggested by Gelman (2006), and uniform priors on the correlation parameters. The priors are proper since the parameter space for γ is bounded.

The output of the expensive computation to evaluate $[\gamma|Y]$ under this specification consists of the third, (possibly, fourth), fifth, sixth and seventh terms in equation (24) (counting c). Since we also need to evaluate $[w|\gamma, Y]$ in order to compute $[w, \gamma|Y]$, we need to save b and a Cholesky factor of Ψ^{-1}; however, this makes saving of some of the terms in the above expression for $\log[\gamma|Y]$ unnecessary.

A.4 Estimation of the Total Variation Norm by Importance Sampling

A Monte Carlo (MC) method to estimate the total variation (TV) norm is presented in this section.

For probability measures G_X and G_Y with densities g_X and g_Y the TV norm is defined as

$$TV(G_X, G_Y) = \sup_{A \in \mathbb{R}} |G_X(A) - G_Y(A)| = \frac{1}{2} \int_{\mathbb{R}} |g_X(t) - g_Y(t)| dt.$$

Notice that

$$\int_{\mathbb{R}} |g_X(t) - g_Y(t)| dt = \int_{\mathbb{R}} \frac{|g_X(t) - g_Y(t)|}{g(t)} g(t) dt \approx \frac{1}{M} \sum_{i=1}^{M} \frac{|g_X(V_i) - g_Y(V_i)|}{g(V_i)},$$

where $V_1, \ldots, V_M$ are *i.i.d.* from g. If the importance density is $g = \frac{1}{2} g_X + \frac{1}{2} g_Y$, the random variable $|g_X(V_i) - g_Y(V_i)|/g(V_i)$ is supported on the interval $[0, 2]$ and, as a consequence, its variance is bounded by 1 from above. (The variance is much lower if the true TV norm is small.) Hence, an MC estimate of the TV norm to a desired accuracy can be easily obtained.

If the densities g_X and g_Y are unknown, but the respective univariate samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ are available, estimates of g_X and g_Y can be used as in Algorithm 1 below. (It is assumed that the sample quantiles for the two samples are consistent; independence is not necessary.)

We use a pilot run to estimate the variance of Z_i and choose M to make the MC error of the estimated TV norm negligible. In our applications, x_i's and y_i's are produced by MCMC runs from the cheap-to-evaluate approximate posterior densities whose length can be chosen by the user to control the accuracy of $\widetilde{g}_X$ and $\widetilde{g}_Y$.

Algorithm 1 ESTIMATE TV NORM

Require: $x_1, \ldots, x_n \sim g_X$; $y_1, \ldots, y_m \sim g_Y$; M

estimate g_X and g_Y using kernel smoothing by $\widetilde{g}_X$ and $\widetilde{g}_Y$ from $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$
for $i = 1, \ldots, M$ **do**
    draw $B_i \sim Bernoulli(1/2)$
    if $B_i = 0$ **then**
        set $V_i \leftarrow x_j$ with probability $1/n$ for $j = 1, \ldots, n$
    else
        set $V_i \leftarrow y_j$ with probability $1/m$ for $j = 1, \ldots, m$
    end if
    set

$$Z_i \leftarrow \frac{|\widetilde{g}_X(V_i) - \widetilde{g}_Y(V_i)|}{\widetilde{g}_X(V_i) + \widetilde{g}_Y(V_i)}$$

end for
return sample mean and sample variance of $Z_1, \ldots, Z_M$

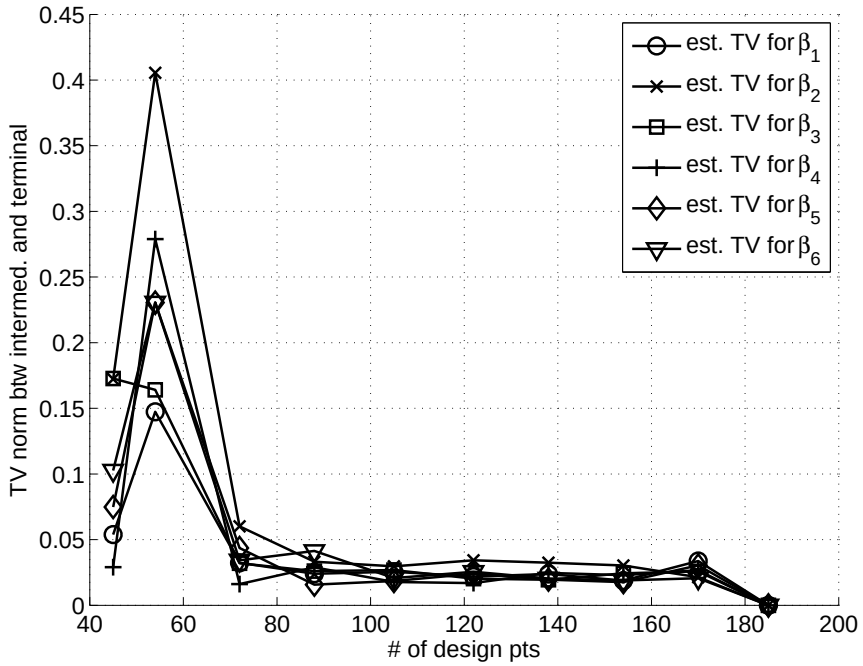

Figure 4: Summaries for `GRIMA`: estimated TV norms between samples from RBF approximations to $\log[\beta|Y]$ with 185 knots and with smaller numbers of knots. The sample size is $3 \cdot 10^4$.

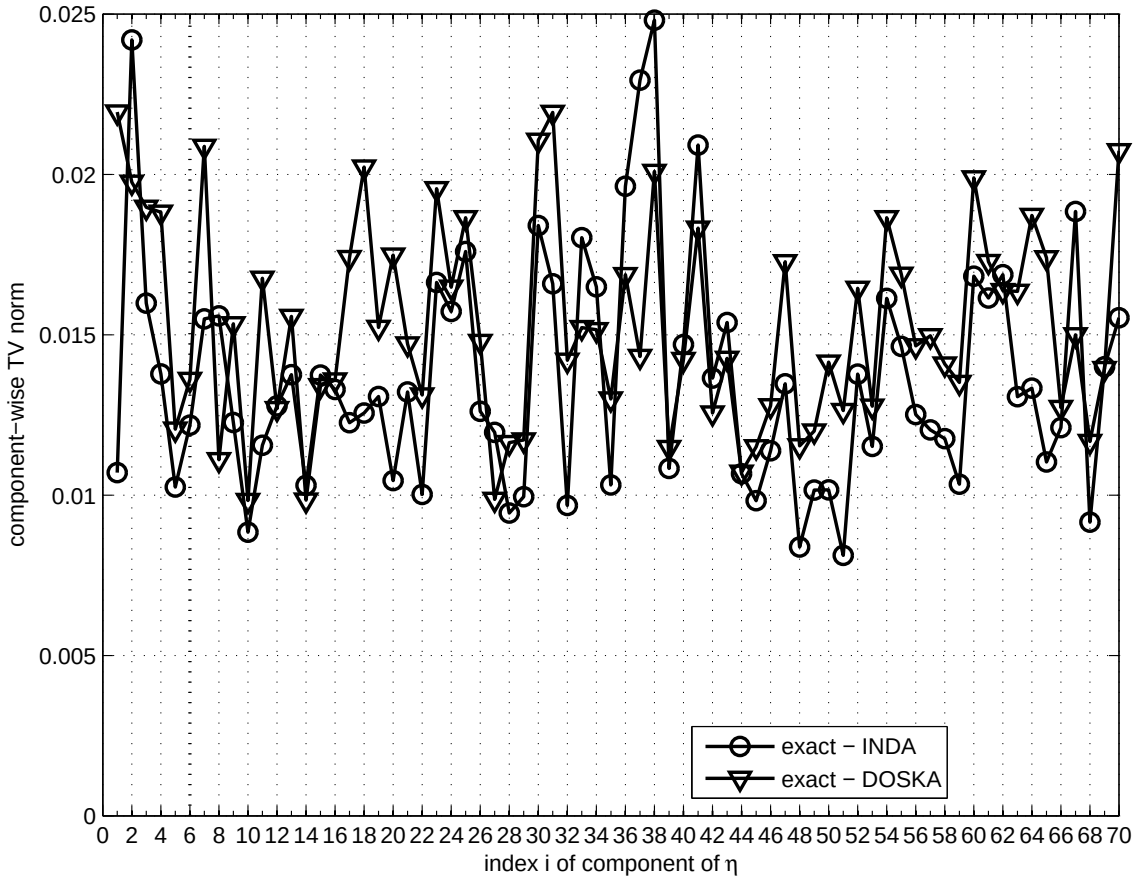

Figure 5: Summaries for the approximation with 185 β-knots: estimated component-wise TV norms between samples from the exact and approximate densities for `DOSKA` (∇) and `INDA` (o). MCMC sample size is 10^5.