Kernel Adaptive Metropolis-Hastings

Dino Sejdinovic[⋆] DINO@GATSBY.UCL.AC.UK
Heiko Strathmann[⋆] UCABHST@GATSBY.UCL.AC.UK
Maria Lomeli Garcia[⋆] MLOMELI@GATSBY.UCL.AC.UK
Christophe Andrieu[‡] C.ANDRIEU@BRISTOL.AC.UK
Arthur Gretton[⋆] ARTHUR.GRETTON@GMAIL.COM
[⋆]Gatsby Unit, CSML, University College London, UK and [‡]School of Mathematics, University of Bristol, UK

Abstract

A Kernel Adaptive Metropolis-Hastings algorithm is introduced, for the purpose of sampling from a target distribution with strongly nonlinear support. The algorithm embeds the trajectory of the Markov chain into a reproducing kernel Hilbert space (RKHS), such that the feature space covariance of the samples informs the choice of proposal. The procedure is computationally efficient and straightforward to implement, since the RKHS moves can be integrated out analytically: our proposal distribution in the original space is a normal distribution whose mean and covariance depend on where the current sample lies in the support of the target distribution, and adapts to its local covariance structure. Furthermore, the procedure requires neither gradients nor any other higher order information about the target, making it particularly attractive for contexts such as Pseudo-Marginal MCMC. Kernel Adaptive Metropolis-Hastings outperforms competing fixed and adaptive samplers on multivariate, highly nonlinear target distributions, arising in both real-world and synthetic examples.

1. Introduction

The choice of the proposal distribution is known to be crucial for the design of Metropolis-Hastings algorithms, and methods for adapting the proposal distribution to increase the sampler's efficiency based on the history of the Markov chain have been widely studied. These methods often aim to learn the covariance structure of the target distribution, and adapt the proposal accordingly. Adaptive MCMC samplers were first studied by Haario et al. (1999; 2001), where the authors propose to update the proposal distribution along the sampling process. Based on the chain history, they estimate the covariance of the target distribution and construct a Gaussian proposal centered at the current chain state, with a particular choice of the scaling factor from Gelman et al. (1996). More sophisticated schemes are presented by Andrieu & Thoms (2008), e.g., adaptive scaling, component-wise scaling, and principal component updates.

While these strategies are beneficial for distributions that show high anisotropy (e.g., by ensuring the proposal uses the right scaling in all principal directions), they may still suffer from low acceptance probability and slow mixing when the target distributions are strongly nonlinear, and the directions of large variance depend on the current location of the sampler in the support. In the present work, we develop an adaptive Metropolis-Hastings algorithm in which samples are mapped to a reproducing kernel Hilbert space, and the proposal distribution is chosen according to the covariance in this feature space (Schölkopf et al., 1998; Smola et al., 2001). Unlike earlier adaptive approaches, the resulting proposal distributions are locally adaptive in input space, and oriented towards nearby regions of high density, rather than simply matching the global covariance structure of the distribution. Our approach combines a move in the feature space with a stochastic step towards the nearest input space point, where the feature space move can be analytically integrated out. Thus, the implementation of the procedure is straightforward: the proposal is simply a multivariate Gaussian in the input space, with location-dependent covariance which is informed by the feature space representation of the target. Furthermore, the resulting Metropolis-Hastings sampler only requires the ability to evaluate the unnormalized density of the target (or its unbiased estimate, as in Pseudo-Marginal MCMC of Andrieu & Roberts, 2009), and no gradient evaluation is needed, making it applicable to situations where more sophisticated schemes based on Hamiltonian Monte Carlo (HMC) or Metropolis Adjusted Langevin Algorithms

Proceedings of the 31[st] International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

(MALA) (Roberts & Stramer, 2003; Girolami & Calderhead, 2011) cannot be applied.

We begin our presentation in Section 2, with a brief overview of existing adaptive Metropolis approaches; we also review covariance operators in the RKHS. Based on these operators, we describe a sampling strategy for Gaussian measures in the RKHS in Section 3, and introduce a cost function for constructing proposal distributions. In Section 4, we outline our main algorithm, termed Kernel Adaptive Metropolis-Hastings (MCMC Kameleon). We provide experimental comparisons with other fixed and adaptive samplers in Section 5, where we show superior performance in the context of Pseudo-Marginal MCMC for Bayesian classification, and on synthetic target distributions with highly nonlinear shape.

2. Background

Adaptive Metropolis Algorithms. Let $\mathcal{X} = \mathbb{R}^d$ be the domain of interest, and denote the unnormalized target density on $\mathcal{X}$ by π. Additionally, let $\Sigma_t = \Sigma_t(x_0, x_1, \ldots, x_{t-1})$ denote an estimate of the covariance matrix of the target density based on the chain history $\{x_i\}_{i=0}^{t-1}$. The original adaptive Metropolis at the current state of the chain state $x_t = y$ uses the proposal

$$q_t(\cdot|y) = \mathcal{N}(y, \nu^2\Sigma_t), \tag{1}$$

where $\nu = 2.38/\sqrt{d}$ is a fixed scaling factor from Gelman et al. (1996). This choice of scaling factor was shown to be optimal (in terms of efficiency measures) for the usual Metropolis algorithm. While this optimality result does not hold for Adaptive Metropolis, it can nevertheless be used as a heuristic. Alternatively, the scale ν can also be adapted at each step as in Andrieu & Thoms (2008, Algorithm 4) to obtain the acceptance rate from Gelman et al. (1996), $a^* = 0.234$.

RKHS Embeddings and Covariance Operators. According to the Moore-Aronszajn theorem (Berlinet & Thomas-Agnan, 2004, p. 19), for every symmetric, positive definite function (*kernel*) $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there is an associated reproducing kernel Hilbert space $\mathcal{H}_k$ of real-valued functions on $\mathcal{X}$ with reproducing kernel k. The map $\varphi : \mathcal{X} \to \mathcal{H}_k$, $\varphi : x \mapsto k(\cdot, x)$ is called the canonical feature map of k. This feature map or embedding of a single point can be extended to that of a probability measure P on $\mathcal{X}$: its kernel embedding is an element $\mu_P \in \mathcal{H}_k$, given by $\mu_P = \int k(\cdot, x)\, dP(x)$ (Berlinet & Thomas-Agnan, 2004; Fukumizu et al., 2004; Smola et al., 2007). If a measurable kernel k is bounded, it is straightforward to show using the Riesz representation theorem that the mean embedding $\mu_k(P)$ exists for all probability measures on $\mathcal{X}$. For many interesting bounded kernels k, including the Gaussian, Laplacian and inverse multi-quadratics, the kernel embedding $P \mapsto \mu_P$ is injective. Such kernels are said to be *characteristic* (Sriperumbudur et al., 2010; 2011), since each distribution is uniquely characterized by its embedding (in the same way that every probability distribution has a unique characteristic function). The kernel embedding μ_P is the representer of expectations of smooth functions w.r.t. P, i.e., $\forall f \in \mathcal{H}_k$, $\langle f, \mu_P \rangle_{\mathcal{H}_k} = \int f(x) dP(x)$. Given samples $\mathbf{z} = \{z_i\}_{i=1}^n \sim P$, the embedding of the empirical measure is $\mu_{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^n k(\cdot, z_i)$.

Next, the covariance operator $C_P : \mathcal{H}_k \to \mathcal{H}_k$ for a probability measure P is given by $C_P = \int k(\cdot, x) \otimes k(\cdot, x)\, dP(x) - \mu_P \otimes \mu_P$ (Baker, 1973; Fukumizu et al., 2004), where for $a, b, c \in \mathcal{H}_k$ the tensor product is defined as $(a \otimes b)c = \langle b, c \rangle_{\mathcal{H}_k} a$. The covariance operator has the property that $\forall f, g \in \mathcal{H}_k$, $\langle f, C_P g \rangle_{\mathcal{H}_k} = \mathbb{E}_P(fg) - \mathbb{E}_P f \mathbb{E}_P g$.

Our approach is based on the idea that the nonlinear support of a target density may be learned using Kernel Principal Component Analysis (Kernel PCA) (Schölkopf et al., 1998; Smola et al., 2001), this being linear PCA on the empirical covariance operator in the RKHS, $C_{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^n k(\cdot, z_i) \otimes k(\cdot, z_i) - \mu_{\mathbf{z}} \otimes \mu_{\mathbf{z}}$, computed on the sample $\mathbf{z}$ defined above. The empirical covariance operator behaves as expected: applying the tensor product definition gives $\langle f, C_{\mathbf{z}} g \rangle_{\mathcal{H}_k} = \frac{1}{n}\sum_{i=1}^n f(z_i)g(z_i) - \left(\frac{1}{n}\sum_{i=1}^n f(z_i)\right)\left(\frac{1}{n}\sum_{i=1}^n g(z_i)\right)$. By analogy with algorithms which use linear PCA directions to inform M-H proposals (Andrieu & Thoms, 2008, Algorithm 8), nonlinear PCA directions can be encoded in the proposal construction, as described in Appendix C. Alternatively, one can focus on a Gaussian measure on the RKHS determined by the empirical covariance operator $C_{\mathbf{z}}$ rather than extracting its eigendirections, which is the approach we pursue in this contribution. This generalizes the proposal (1), which considers the Gaussian measure induced by the empirical covariance matrix on the original space.

3. Sampling in RKHS

We next describe the proposal distribution at iteration t of the MCMC chain. We will assume that a subset of the chain history, denoted $\mathbf{z} = \{z_i\}_{i=1}^n$, $n \leq t-1$, is available. Our proposal is constructed by first considering the samples in the RKHS associated to the empirical covariance operator, and then performing a gradient descent step on a cost function associated with those samples.

Gaussian Measure of the Covariance Operator. We will work with the Gaussian measure on the RKHS $\mathcal{H}_k$ with mean $k(\cdot, y)$ and covariance $\nu^2 C_{\mathbf{z}}$, where $\mathbf{z} = \{z_i\}_{i=1}^n$ is the subset of the chain history. While there is no analogue of a Lebesgue measure in an infinite dimensional RKHS, it is instructive (albeit with some abuse of notation) to denote

this measure in the "density form" $\mathcal{N}(f\,;\,k(\cdot,y),\nu^2 C_{\mathbf{z}}) \propto \exp\left(-\frac{1}{2\nu^2}\left\langle f-k(\cdot,y), C_{\mathbf{z}}^{-1}(f-k(\cdot,y))\right\rangle_{\mathcal{H}_k}\right)$. As $C_{\mathbf{z}}$ is a finite-rank operator, this measure is supported only on a finite-dimensional affine space $k(\cdot,y)+\mathcal{H}_{\mathbf{z}}$, where $\mathcal{H}_{\mathbf{z}} = \operatorname{span}\{k(\cdot,z_i)\}_{i=1}^n$ is the subspace spanned by the canonical features of $\mathbf{z}$. It can be shown that a sample from this measure has the form $f = k(\cdot,y) + \sum_{i=1}^n \beta_i \left[k(\cdot,z_i)-\mu_{\mathbf{z}}\right]$, where $\beta \sim \mathcal{N}(0, \frac{\nu^2}{n} I)$ is isotropic. Indeed, to see that f has the correct covariance structure, note that:

$$\begin{aligned}
&\mathbb{E}\left[(f-k(\cdot,y))\otimes(f-k(\cdot,y))\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n\sum_{j=1}^n \beta_i\beta_j \left(k(\cdot,z_i)-\mu_{\mathbf{z}}\right)\otimes\left(k(\cdot,z_j)-\mu_{\mathbf{z}}\right)\right] \\
&= \frac{\nu^2}{n}\sum_{i=1}^n \left(k(\cdot,z_i)-\mu_{\mathbf{z}}\right)\otimes\left(k(\cdot,z_i)-\mu_{\mathbf{z}}\right) = \nu^2 C_{\mathbf{z}}.
\end{aligned}$$

Due to the equivalence in the RKHS between a Gaussian measure and a Gaussian Process (GP) (Berlinet & Thomas-Agnan, 2004, Ch. 4), we can think of the RKHS samples f as trajectories of the GP with mean $m(x) = k(x,y)$ and covariance function

$$\begin{aligned}
\kappa(x,x') &= \operatorname{cov}\left[f(x), f(x')\right] \\
&= \frac{\nu^2}{n}\sum_{i=1}^n \left(k(x,z_i)-\mu_{\mathbf{z}}(x)\right)\left(k(x',z_i)-\mu_{\mathbf{z}}(x')\right).
\end{aligned}$$

The covariance function κ of this GP is therefore the kernel k convolved with itself with respect to the empirical measure associated to the samples $\mathbf{z}$, and draws from this GP therefore lie in a smaller RKHS; see Saitoh (1997, p. 21) for details.

Obtaining Target Samples through Gradient Descent. We have seen how to obtain the RKHS sample $f = k(\cdot,y) + \sum_{i=1}^n \beta_i \left[k(\cdot,z_i)-\mu_{\mathbf{z}}\right]$ from the Gaussian measure in the RKHS. This sample does not in general have a corresponding pre-image in the original domain $\mathcal{X} = \mathbb{R}^d$; i.e., there is no point $x_* \in \mathcal{X}$ such that $f = k(\cdot,x_*)$. If there were such a point, then we could use it as a proposal in the original domain. Therefore, we are ideally looking for a point $x^* \in \mathcal{X}$ whose canonical feature map $k(\cdot,x^*)$ is close to f in the RKHS norm. We consider the optimization problem

$$\begin{aligned}
&\arg\min_{x\in\mathcal{X}} \left\|k(\cdot,x) - f\right\|_{\mathcal{H}_k}^2 = \\
&\arg\min_{x\in\mathcal{X}} \left\{k(x,x) - 2k(x,y) - 2\sum_{i=1}^n \beta_i \left[k(x,z_i)-\mu_{\mathbf{z}}(x)\right]\right\}.
\end{aligned}$$

In general, this is a non-convex minimization problem, and may be difficult to solve (Bakir et al., 2003). Rather than solving it for every new vector of coefficients β, which would lead to an excessive computational burden for every proposal made, we simply make a single descent step along the gradient of the cost function,

$$g(x) = k(x,x) - 2k(x,y) - 2\sum_{i=1}^n \beta_i \left[k(x,z_i)-\mu_{\mathbf{z}}(x)\right], \tag{2}$$

i.e., the proposed new point is

$$x^* = y - \eta \nabla_x g(x)|_{x=y} + \xi,$$

where η is a gradient step size parameter and $\xi \sim \mathcal{N}(0,\gamma^2 I)$ is an additional isotropic 'exploration' term *after* the gradient step. It will be useful to split the scaled gradient at y into two terms as $\eta\nabla_x g(x)|_{x=y} = \eta\left(a_y - M_{\mathbf{z},y}H\beta\right)$, where $a_y = \nabla_x k(x,x)|_{x=y} - 2\nabla_x k(x,y)|_{x=y}$,

$$M_{\mathbf{z},y} = 2\left[\nabla_x k(x,z_1)|_{x=y}, \ldots, \nabla_x k(x,z_n)|_{x=y}\right] \tag{3}$$

is a $d\times n$ matrix, and $H = I - \frac{1}{n}\mathbf{1}_{n\times n}$ is the $n\times n$ centering matrix.

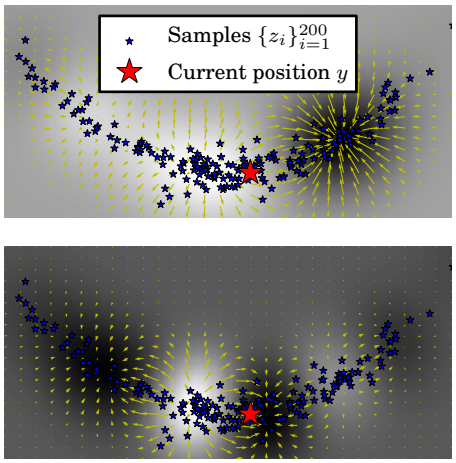

Figure 1. Heatmaps (white denotes large) and gradients of $g(x)$ for two samples of β and fixed $\mathbf{z}$.

Figure 1 plots $g(x)$ and its gradients for several samples of β-coefficients, in the case where the underlying $\mathbf{z}$-samples are from the two-dimensional nonlinear Banana target distribution of Haario et al. (1999). It can be seen that g may have multiple local minima, and that it varies most along the high-density regions of the Banana distribution.

4. MCMC Kameleon Algorithm

4.1. Proposal Distribution

We now have a recipe to construct a proposal that is able to adapt to the local covariance structure for the current chain

MCMC Kameleon

Input: unnormalized target π, subsample size n, scaling parameters ν, γ, adaptation probabilities $\{p_t\}_{t=0}^{\infty}$, kernel k,

- At iteration $t+1$,
  1. With probability p_t, update a random subsample $\mathbf{z} = \{z_i\}_{i=1}^{\min(n,t)}$ of the chain history $\{x_i\}_{i=0}^{t-1}$,
  2. Sample proposed point x^* from $q_{\mathbf{z}}(\cdot|x_t) = \mathcal{N}(x_t, \gamma^2 I + \nu^2 M_{\mathbf{z},x_t} H M_{\mathbf{z},x_t}^\top)$, where $M_{\mathbf{z},x_t}$ is given in Eq. (3) and $H = I - \frac{1}{n}\mathbf{1}_{n\times n}$ is the centering matrix,
  3. Accept/Reject with the Metropolis-Hastings acceptance probability $A(x_t, x^*)$ in Eq. (4),

$$x_{t+1} = \begin{cases} x^*, & \text{w.p. } A(x_t, x^*), \\ x_t, & \text{w.p. } 1 - A(x_t, x^*). \end{cases}$$

state y. This proposal depends on a subset of the chain history $\mathbf{z}$, and is denoted by $q_{\mathbf{z}}(\cdot|y)$. While we will later simplify this proposal by integrating out the moves in the RKHS, it is instructive to think of the proposal generating process as:

1. Sample $\beta \sim \mathcal{N}(0, \nu^2 I)$ ($n \times 1$ normal of RKHS coefficients).
   - This represents an RKHS sample $f = k(\cdot, y) + \sum_{i=1}^{n} \beta_i [k(\cdot, z_i) - \mu_{\mathbf{z}}]$ which is the goal of the cost function $g(x)$.
2. Move along the gradient of g: $x^* = y - \eta \nabla_x g(x)|_{x=y} + \xi$.
   - This gives a proposal $x^*|y, \beta \sim \mathcal{N}(y - \eta a_y + \eta M_{\mathbf{z},y} H \beta, \gamma^2 I)$ ($d \times 1$ normal in the original space).

Our first step in the derivation of the explicit proposal density is to show that as long as k is a differentiable positive definite kernel, the term a_y vanishes.

Proposition 1. *Let k be a differentiable positive definite kernel. Then $a_y = \nabla_x k(x,x)|_{x=y} - 2\nabla_x k(x,y)|_{x=y} = 0$.*

Since $a_y = 0$, the gradient step size η always appears together with β, so we merge η and the scale ν of the β-coefficients into a single scale parameter, and set $\eta = 1$ henceforth. Furthermore, since both $p(\beta)$ and $p_{\mathbf{z}}(x^*|y,\beta)$ are multivariate Gaussian densities, the proposal density $q_{\mathbf{z}}(x^*|y) = \int p(\beta) p_{\mathbf{z}}(x^*|y,\beta) d\beta$ can be computed analytically. We therefore get the following closed form expression for the proposal distribution.

Proposition 2. $q_{\mathbf{z}}(\cdot|y) = \mathcal{N}(y, \gamma^2 I + \nu^2 M_{\mathbf{z},y} H M_{\mathbf{z},y}^\top)$.

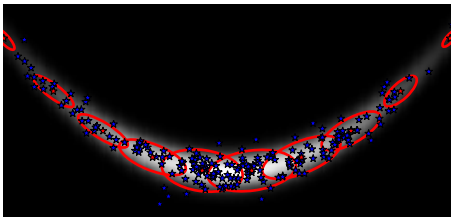

Figure 2. 95% contours (red) of proposal distributions evaluated at a number of points, for the first two dimensions of the banana target of Haario et al. (1999). Underneath is the density heatmap, and the samples (blue) used to construct the proposals.

Proofs of the above Propositions are given in Appendix A.

With the derived proposal distribution, we proceed with the standard Metropolis-Hastings accept/reject scheme, where the proposed sample x^* is accepted with probability

$$A(x_t, x^*) = \min\left\{1, \frac{\pi(x^*) q_{\mathbf{z}}(x_t|x^*)}{\pi(x_t) q_{\mathbf{z}}(x^*|x_t)}\right\}, \quad (4)$$

giving rise to the MCMC Kameleon Algorithm. Note that each $\pi(x^*)$ and $\pi(x_t)$ could be replaced by their unbiased estimates without impacting the invariant distribution (Andrieu & Roberts, 2009).

The constructed family of proposals encodes local structure of the target distribution, which is learned based on the subsample $\mathbf{z}$. Figure 2 depicts the regions that contain 95% of the mass of the proposal distribution $q_{\mathbf{z}}(\cdot|y)$ at various states y for a fixed subsample $\mathbf{z}$, where the Banana target is used (details in Section 5). More examples of proposal contours can be found in Appendix B.

4.2. Properties of the Algorithm

The update schedule and convergence. MCMC Kameleon requires a subsample $\mathbf{z} = \{z_i\}_{i=1}^{n}$ at each iteration of the algorithm, and the proposal distribution $q_{\mathbf{z}}(\cdot|y)$ is updated each time a new subsample $\mathbf{z}$ is obtained. It is well known that a chain which keeps adapting the proposal distribution need not converge to the correct target (Andrieu & Thoms, 2008). To guarantee convergence, we introduce adaptation probabilities $\{p_t\}_{t=0}^{\infty}$, such that $p_t \to 0$ and $\sum_{t=1}^{\infty} p_t = \infty$, and at iteration t we update the subsample $\mathbf{z}$ with probability p_t. As adaptations occur with decreasing probability, Theorem 1 of Roberts & Rosenthal (2007) implies that the resulting algorithm is ergodic and converges to the correct target. Another straightforward way to guarantee convergence is to fix the set $\mathbf{z} = \{z_i\}_{i=1}^{n}$ after a "burn-in" phase; i.e., to stop adapting Roberts & Rosenthal (2007, Proposition 2). In this case, a "burn-in" phase is used to get a rough sketch of the shape of the distribution: the initial samples need not

come from a converged or even valid MCMC chain, and it suffices to have a scheme with good exploratory properties, e.g., Welling & Teh (2011). In MCMC Kameleon, the term γ allows exploration in the initial iterations of the chain (while the subsample $\mathbf{z}$ is still not informative about the structure of the target) and provides regularization of the proposal covariance in cases where it might become ill-conditioned. Intuitively, a good approach to setting γ is to slowly decrease it with each adaptation, such that the learned covariance progressively dominates the proposal.

Symmetry of the proposal. In Haario et al. (2001), the proposal distribution is asymptotically symmetric due to the vanishing adaptation property. Therefore, the authors compute the standard Metropolis acceptance probability. In our case, the proposal distribution is a Gaussian with mean at the current state of the chain $x_t = y$ and covariance $\gamma^2 I + \nu^2 M_{\mathbf{z},y} H M_{\mathbf{z},y}^\top$, where $M_{\mathbf{z},y}$ depends both on the current state y and a random subsample $\mathbf{z} = \{z_i\}_{i=1}^n$ of the chain history $\{x_i\}_{i=0}^{t-1}$. This proposal distribution is never symmetric (as covariance of the proposal always depends on the current state of the chain), and therefore we use the Metropolis-Hastings acceptance probability to reflect this.

Relationship to MALA and Manifold MALA. The Metropolis Adjusted Langevin Algorithm (MALA) algorithm uses information about the gradient of the log-target density at the current chain state to construct a proposed point for the Metropolis step. Our approach does not require that the log-target density gradient be available or computable. Kernel gradients in the matrix $M_{\mathbf{z},y}$ are easily obtained for commonly used kernels, including the Gaussian kernel (see section 4.3), for which the computational complexity is equal to evaluating the kernel itself. Moreover, while standard MALA simply shifts the mean of the proposal distribution along the gradient and then adds an isotropic exploration term, our proposal is centered at the current state, and it is the covariance structure of the proposal distribution that coerces the proposed points to belong to the high-density regions of the target. It would be straightforward to modify our approach to include a drift term along the gradient of the log-density, should such information be available, but it is unclear whether this would provide additional performance gains. Further work is required to elucidate possible connections between our approach and the use of a preconditioning matrix (Roberts & Stramer, 2003) in the MALA proposal; i.e., where the exploration term is scaled with appropriate metric tensor information, as in Riemannian manifold MALA (Girolami & Calderhead, 2011).

4.3. Examples of Covariance Structure for Standard Kernels

The proposal distributions in MCMC Kameleon are dependant on the choice of the kernel k. To gain intuition regarding their covariance structure, we give two examples below.

Linear kernel. In the case of a linear kernel $k(x, x') = x^\top x'$, we obtain $M_{\mathbf{z},y} = 2\left[\nabla_x x^\top z_1|_{x=y}, \ldots, \nabla_x x^\top z_n|_{x=y}\right] = 2\mathbf{Z}^\top$, so the proposal is given by $q_{\mathbf{z}}(\cdot|y) = \mathcal{N}(y, \gamma^2 I + 4\nu^2 \mathbf{Z}^\top H \mathbf{Z})$; thus, the proposal simply uses the scaled empirical covariance $\mathbf{Z}^\top H \mathbf{Z}$ just like standard Adaptive Metropolis (Haario et al., 1999), with an additional isotropic exploration component, and depends on y only through the mean.

Gaussian kernel. In the case of a Gaussian kernel $k(x, x') = \exp\left(-\frac{\|x-x'\|_2^2}{2\sigma^2}\right)$, since $\nabla_x k(x, x') = \frac{1}{\sigma^2} k(x, x')(x' - x)$, we obtain

$$M_{\mathbf{z},y} = \frac{2}{\sigma^2}\left[k(y, z_1)(z_1 - y), \ldots, k(y, z_n)(z_n - y)\right].$$

Consider how this encodes the covariance structure of the target distribution:

$$\begin{aligned} R_{ij} &= \gamma^2 \delta_{ij} \\ &+ \frac{4\nu^2(n-1)}{\sigma^4 n} \sum_{a=1}^{n} \left[k(y, z_a)\right]^2 (z_{a,i} - y_i)(z_{a,j} - y_j) \\ &- \frac{4\nu^2}{\sigma^4 n} \sum_{a \neq b} k(y, z_a) k(y, z_b)(z_{a,i} - y_i)(z_{b,j} - y_j). \quad (5) \end{aligned}$$

As the first two terms dominate, the previous points z_a which are close to the current state y (for which $k(y, z_a)$ is large) have larger weights, and thus they have more influence in determining the covariance of the proposal at y.

Matérn kernel. In the Matérn family of kernels $k_{\vartheta,\rho}(x, x') = \frac{2^{1-\vartheta}}{\Gamma(\vartheta)} \left(\frac{\|x-x'\|_2}{\rho}\right)^\vartheta K_\vartheta \left(\frac{\|x-x'\|_2}{\rho}\right)$, where K_ϑ is the modified Bessel function of the second kind, we obtain a form of the covariance structure very similar to that of the Gaussain kernel. In this case, $\nabla_x k_{\vartheta,\rho}(x, x') = \frac{1}{2\rho^2(\vartheta-1)} k_{\vartheta-1,\rho}(x, x')(x' - x)$, so the only difference (apart from the scalings) to (5) is that the weights are now determined by a "rougher" kernel $k_{\vartheta-1,\rho}$ of the same family.

5. Experiments

In the experiments, we compare the following samplers: **(SM)** Standard Metropolis with the isotropic proposal $q(\cdot|y) = \mathcal{N}(y, \nu^2 I)$ and scaling $\nu = 2.38/\sqrt{d}$, **(AM-FS)** Adaptive Metropolis with a learned covariance matrix and fixed scaling $\nu = 2.38/\sqrt{d}$, **(AM-LS)** Adaptive Metropolis with a learned covariance matrix and scaling learned to bring the acceptance rate close to $\alpha^* = 0.234$ as described in Andrieu & Thoms (2008, Algorithm 4), and **(KAMH-LS)** MCMC Kameleon with the scaling ν learned

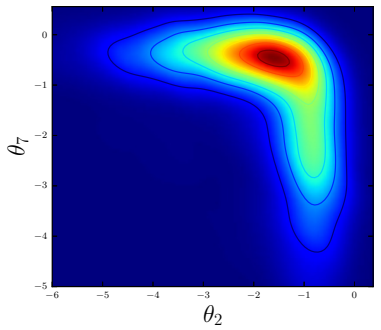

Figure 3. Dimensions 2 and 7 of the marginal hyperparameter posterior on the UCI Glass dataset

in the same fashion (γ was fixed to 0.2), and which also stops adapting the proposal after the burn-in of the chain (in all experiments, we use a random subsample $\mathbf{z}$ of size $n = 1000$, and a Gaussian kernel with bandwidth selected according to the median heuristic). We consider the following nonlinear targets: (1) the posterior distribution of Gaussian Process (GP) classification hyperparameters (Filippone & Girolami, 2014) on the UCI glass dataset, and (2) the synthetic banana-shaped distribution of Haario et al. (1999) and a flower-shaped disribution concentrated on a circle with a periodic perturbation.

5.1. Pseudo-Marginal MCMC for GP Classification

In the first experiment, we illustrate usefulness of the MCMC Kameleon sampler in the context of Bayesian classification with GPs (Williams & Barber, 1998). Consider the joint distribution of latent variables $\mathbf{f}$, labels $\mathbf{y}$ (with covariate matrix X), and hyperparameters θ, given by

$$p(\mathbf{f}, \mathbf{y}, \theta) = p(\theta)p(\mathbf{f}|\theta)p(\mathbf{y}|\mathbf{f}),$$

where $\mathbf{f}|\theta \sim \mathcal{N}(0, \mathcal{K}_\theta)$, with $\mathcal{K}_\theta$ modeling the covariance between latent variables evaluated at the input covariates: $(\mathcal{K}_\theta)_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}'_j|\theta) = \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\frac{(x_{i,d}-x'_{j,d})^2}{\ell_d^2}\right)$ and $\theta_d = \log \ell_d^2$. We restrict our attention to the binary logistic classifier; i.e., the likelihood is given by $p(y_i|f_i) = \frac{1}{1-\exp(-y_i f_i)}$ where $y_i \in \{-1, 1\}$. We pursue a fully Bayesian treatment, and estimate the posterior of the hyperparameters θ. As observed by Murray & Adams (2012), a Gibbs sampler on $p(\theta, \mathbf{f}|y)$, which samples from $p(\mathbf{f}|\theta, y)$ and $p(\theta|\mathbf{f}, y)$ in turn, is problematic, as $p(\theta|\mathbf{f}, y)$ is extremely sharp, drastically limiting the amount that any Markov chain can update $\theta|\mathbf{f}, y$. On the other hand, if we directly consider the marginal posterior $p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta)$ of the hyperparameters, a much less peaked distribution can be obtained. However, the marginal likelihood $p(\mathbf{y}|\theta)$ is intractable for non-Gaussian likelihoods $p(\mathbf{y}|\mathbf{f})$, so it is not possible to analytically integrate out the latent variables. Recently developed pseudo-marginal MCMC methods (Andrieu & Roberts, 2009) enable *exact* inference on this problem (Filippone & Girolami, 2014), by replacing $p(\mathbf{y}|\theta)$ with an unbiased estimate

$$\hat{p}(\mathbf{y}|\theta) := \frac{1}{n_{\text{imp}}}\sum_{i=1}^{n_{\text{imp}}} p(\mathbf{y}|\mathbf{f}^{(i)})\frac{p(\mathbf{f}^{(i)}|\theta)}{q(\mathbf{f}^{(i)}|\theta)}, \tag{6}$$

where $\left\{\mathbf{f}^{(i)}\right\}_{i=1}^{n_{\text{imp}}} \sim q(\mathbf{f}|\theta)$ are n_{imp} importance samples. In Filippone & Girolami (2014), the importance distribution $q(\mathbf{f}|\theta)$ is chosen as the Laplacian or as the Expectation Propagation (EP) approximation of $p(\mathbf{f}|\mathbf{y}, \theta) \propto p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\theta)$, leading to state-of-the-art results.

We consider the UCI Glass dataset (Bache & Lichman, 2013), where classification of window against non-window glass is sought. Due to the heterogeneous structure of each of the classes (i.e., non-window glass consists of containers, tableware and headlamps), there is no single consistent set of lengthscales determining the decision boundary, so one expects the posterior of the covariance bandwidths θ_d to have a complicated (nonlinear) shape. This is illustrated by the plot of the posterior projections to the dimensions 2 and 7 (out of 9) in Figure 3. Since the ground truth for the hyperparameter posterior is not available, we initially ran 30 Standard Metropolis chains for 500,000 iterations (with a 100,000 burn-in), kept every 1000-th sample in each of the chains, and combined them. The resulting samples were used as a benchmark, to evaluate the performance of shorter single-chain runs of **SM**, **AM-FS**, **AM-LS** and **KAMH-LS**. Each of these algorithms was run for 100,000 iterations (with a 20,000 burnin) and every 20-th sample was kept. Two metrics were used in evaluating the performance of the four samplers, relative to the large-scale benchmark. First, the distance $\left\|\hat{\mu}_\theta - \mu_\theta^b\right\|_2$ was computed between the mean $\hat{\mu}_\theta$ estimated from each of the four sampler outputs, and the mean μ_θ^b on the benchmark sample (Fig. 4, left), as a function of sample size. Second, the MMD (Borgwardt et al., 2006; Gretton et al., 2007) was computed between each sampler output and the benchmark sample, using the polynomial kernel $(1 + \langle\theta, \theta'\rangle)^3$; i.e., the comparison was made in terms of all mixed moments of order up to 3 (Fig. 4, right). The figures indicate that **KAMH-LS** approximates the benchmark sample better than the competing approaches, where the effect is especially pronounced in the high order moments, indicating that **KAMH-LS** thoroughly explores the distribution support in a relatively small number of samples.

We emphasise that, as for *any* pseudo-marginal MCMC scheme, neither the likelihood itself, nor any higher-order information about the marginal posterior target $p(\theta|\mathbf{y})$, are available. This makes HMC or MALA based approaches such as (Roberts & Stramer, 2003; Girolami & Calderhead, 2011) unsuitable for this problem, so it is very difficult to deal with strongly nonlinear posterior targets. In contrast, as indicated in this example, the MCMC Kameleon scheme is able to effectively sample from such nonlinear targets,

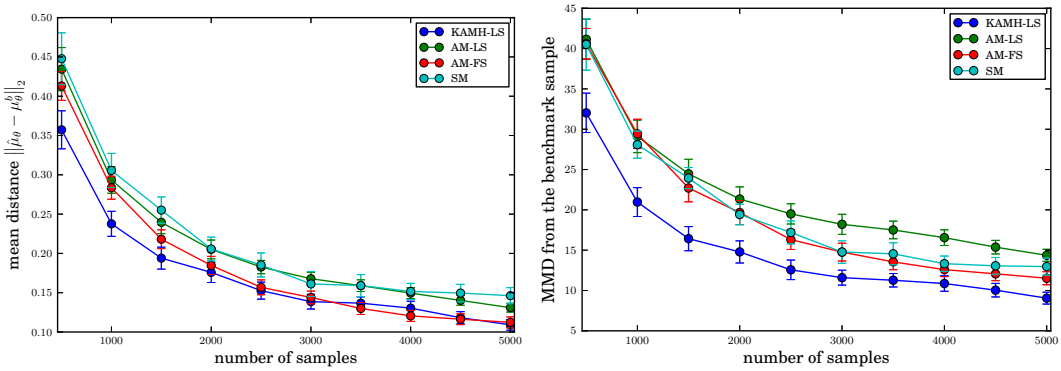

Figure 4. The comparison of **SM**, **AM-FS**, **AM-LS** and **KAMH-LS** in terms of the distance between the estimated mean and the mean on the benchmark sample (left) and in terms of the maximum mean discrepancy to the benchmark sample (right). The results are averaged over 30 chains for each sampler. Error bars represent 80%-confidence intervals.

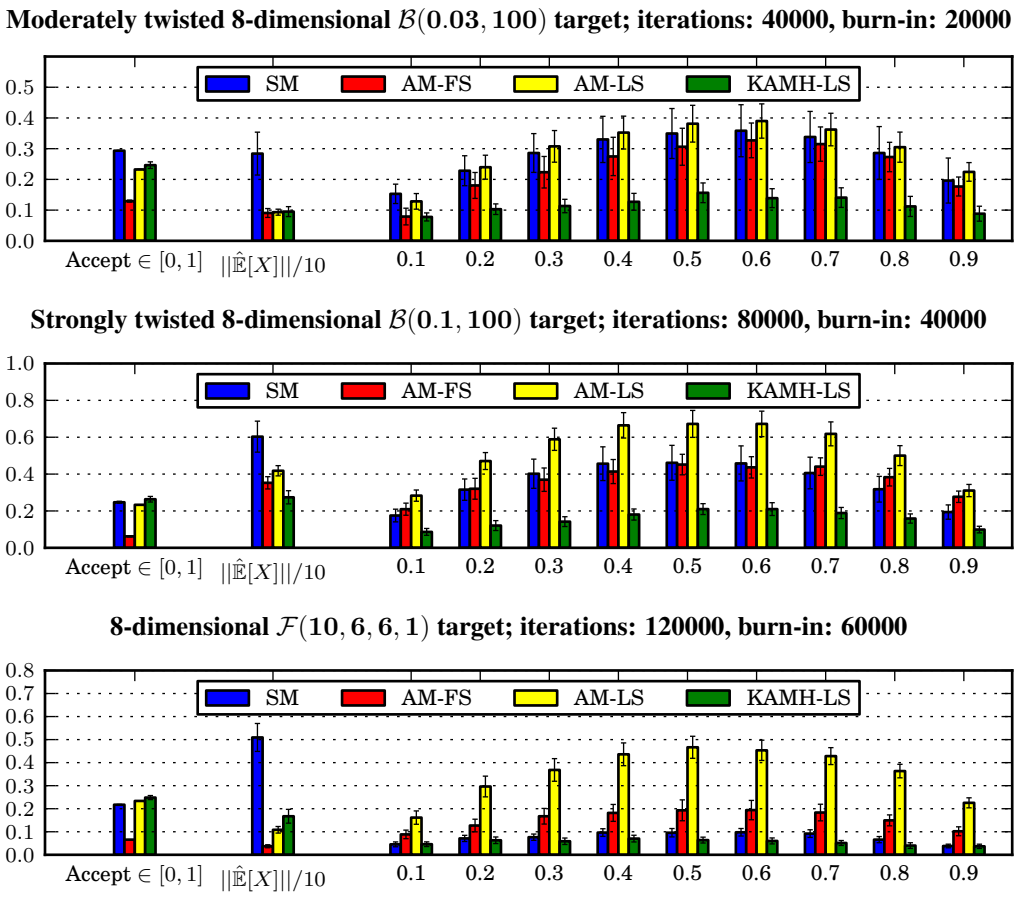

Figure 5. Results for three nonlinear targets, averaged over 20 chains for each sampler. *Accept* is the acceptance rate scaled to the interval $[0, 1]$. The norm of the mean $||\hat{\mathbb{E}}[X]||$ is scaled by 1/10 to fit into the figure scalling, and the bars over the $0.1, \ldots, 0.9$-quantiles represent the deviation from the exact quantiles, scaled by 10; i.e., 0.1 corresponds to 1% deviation. Error bars represent 80%-confidence intervals.

and outperforms the vanilla Metropolis methods, which are the *only* competing choices in the pseudo-marginal context.

In addition, since the bulk of the cost for pseudo-marginal MCMC is in importance sampling in order to obtain the acceptance ratio, the additional cost imposed by **KAMH-LS** is negligible. Indeed, we observed that there is an increase of only 2-3% in terms of effective computation time in comparison to all other samplers, for the chosen size of the chain history subsample ($n = 1000$).

5.2. Synthetic examples

Banana target. In Haario et al. (1999), the following family of nonlinear target distributions is considered. Let $X \sim \mathcal{N}(0, \Sigma)$ be a multivariate normal in $d \geq 2$ dimensions, with $\Sigma = \mathrm{diag}(v, 1, \ldots, 1)$, which undergoes the transformation $X \to Y$, where $Y_2 = X_2 + b(X_1^2 - v)$, and $Y_i = X_i$ for $i \neq 2$. We will write $Y \sim \mathcal{B}(b, v)$. It is clear that $\mathbb{E}Y = 0$, and that

$$\mathcal{B}(y; b, v) = \mathcal{N}(y_1; 0, v)\mathcal{N}(y_2; b(y_1^2 - v), 1) \prod_{j=3}^{d} \mathcal{N}(y_j; 0, 1).$$

Flower target. The second target distribution we consider is the d-dimensional flower target $\mathcal{F}(r_0, A, \omega, \sigma)$, with

$$\mathcal{F}(x; r_0, A, \omega, \sigma) = \exp\left(-\frac{\sqrt{x_1^2 + x_2^2} - r_0 - A\cos\left(\omega \mathrm{atan2}\left(x_2, x_1\right)\right)}{2\sigma^2}\right) \times \prod_{j=3}^{d} \mathcal{N}(x_j; 0, 1).$$

This distribution concentrates around the r_0-circle with a periodic perturbation (with amplitude A and frequency ω) in the first two dimensions.

In these examples, exact quantile regions of the targets can be computed analytically, so we can directly assess performance without the need to estimate distribution distances on the basis of samples (i.e., by estimating MMD to the benchmark sample). We compute the following measures of performance (similarly as in Haario et al. (1999); Andrieu & Thoms (2008)) based on the chain after burn-in: average acceptance rate, norm of the empirical mean (the true mean is by construction zero for all targets), and the deviation of the empirical quantiles from the true quantiles. We consider 8-dimensional target distributions: the moderately twisted $\mathcal{B}(0.03, 100)$ banana target (Figure 5, top) and the strongly twisted $\mathcal{B}(0.1, 100)$ banana target (Figure 5, middle) and $\mathcal{F}(10, 6, 6, 1)$ flower target (Figure 5, bottom).

The results show that MCMC Kameleon is superior to the competing samplers. Since the covariance of the proposal adapts to the local structure of the target at the current chain state, as illustrated in Figure 2, MCMC Kameleon does not suffer from wrongly scaled proposal distributions. The result is a significantly improved quantile performance in comparison to all competing samplers, as well as a comparable or superior norm of the empirical mean. **SM** has a significantly larger norm of the empirical mean, due to its purely random walk behavior (e.g., the chain tends to get stuck in one part of the space, and is not able to traverse both tails of the banana target equally well). **AM** with fixed scale has a low acceptance rate (indicating that the scaling of the proposal is too large), and even though the norm of the empirical mean is much closer to the true value, quantile performance of the chain is poor. Even if the estimated covariance matrix closely resembles the true global covariance matrix of the target, using it to construct proposal distributions at every state of the chain may not be the best choice. For example, **AM** correctly captures scalings along individual dimensions for the flower target (the norm of its empirical mean is close to its true value of zero) but fails to capture local dependence structure. The flower target, due to its symmetry, has an isotropic covariance in the first two dimensions – even though they are highly dependent. This leads to a mismatch in the scale of the covariance and the scale of the target, which concentrates on a thin band in the joint space. **AM-LS** has the "correct" acceptance rate, but the quantile performance is even worse, as the scaling now becomes too small to traverse high-density regions of the target.

6. Conclusions

We have constructed a simple, versatile, adaptive, gradient-free MCMC sampler that constructs a family of proposal distributions based on the sample history of the chain. These proposal distributions automatically conform to the local covariance structure of the target distribution at the current chain state. In experiments, the sampler outperforms existing approaches on nonlinear target distributions, both by exploring the entire support of these distributions, and by returning accurate empirical quantiles, indicating faster mixing. Possible extensions include incorporating additional parametric information about the target densities, and exploring the tradeoff between the degree of subsampling of the chain history and convergence of the sampler.

Software. Python implementation of MCMC Kameleon is available at `https://github.com/karlnapf/kameleon-mcmc`.

Acknowledgments. D.S., H.S., M.L.G. and A.G. acknowledge support of the Gatsby Charitable Foundation. We thank Mark Girolami for insightful discussions and the anonymous reviewers for useful comments.

References

Andrieu, C. and Roberts, G.O. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, 37(2): 697–725, 2009.

Andrieu, C. and Thoms, J. A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373, 2008.

Bache, K. and Lichman, M. UCI Machine Learning Repository, 2013. URL http://archive.ics.uci.edu/ml.

Baker, C. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

Bakir, G., Weston, J., and Schölkopf, B. Learning to find pre-images. In *Advances in Neural Information Processing Systems 16*. MIT Press, 2003.

Berlinet, A. and Thomas-Agnan, C. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. J. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics (ISMB)*, 22(14):e49–e57, 2006.

Filippone, M. and Girolami, M. Pseudo-marginal Bayesian inference for Gaussian Processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014. doi: TPAMI.2014.2316530.

Fukumizu, K., Bach, F. R., and Jordan, M. I. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Mach. Learn. Res.*, 5:73–99, 2004.

Gelman, A., Roberts, G. O., and Gilks, W. R. Efficient Metropolis jumping rules. In *Bayesian statistics, 5 (Alicante, 1994)*, Oxford Sci. Publ., pp. 599–607. 1996.

Girolami, M. and Calderhead, B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 73(2):123–214, 2011.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. A kernel method for the two-sample problem. In *Advances in Neural Information Processing Systems 19*, pp. 513–520, 2007.

Haario, H., Saksman, E., and Tamminen, J. Adaptive Proposal Distribution for Random Walk Metropolis Algorithm. *Comput. Stat.*, 14(3):375–395, 1999.

Haario, H., Saksman, E., and Tamminen, J. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

Murray, I. and Adams, R.P. Slice sampling covariance hyperparameters of latent Gaussian models. In *Advances in Neural Information Processing Systems 23*. 2012.

Roberts, G.O. and Rosenthal, J.S. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Probab.*, 44(2):458–475, 03 2007.

Roberts, G.O. and Stramer, O. Langevin diffusions and Metropolis-Hastings algorithms. *Methodol. Comput. Appl. Probab.*, 4:337–358, 2003.

Saitoh, S. *Integral transforms, reproducing kernels, and their applications*. Pitman Research Notes in Mathematics 369, Longman Scientific & Techn., 1997.

Schölkopf, B., Smola, A. J., and Müller, K.-R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10:1299–1319, 1998.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A Hilbert space embedding for distributions. In *Proceedings of the Conference on Algorithmic Learning Theory (ALT)*, pp. 13–31. Springer, 2007.

Smola, A. J., Mika, S., Schölkopf, B., and Williamson, R. C. Regularized principal manifolds. *J. Mach. Learn. Res.*, 1:179–209, 2001.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Lanckriet, G., and Schölkopf, B. Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 11:1517–1561, 2010.

Sriperumbudur, B., Fukumizu, K., and Lanckriet, G. Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.*, 12:2389–2410, 2011.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Springer, 2008.

Welling, M. and Teh, Y.W. Bayesian learning via stochastic gradient Langevin dynamics. In *Proc. of the 28th International Conference on Machine Learning (ICML)*, pp. 681–688, 2011.

Williams, C.K.I. and Barber, D. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998.