Gaussian Processes of Nonlinear Diffusion Filtering

Ramūnas Girdziušas and Jorma Laaksonen
Laboratory of Computer and Information Science
Helsinki University of Technology
P.O.Box 5400, FI-02015 HUT, Espoo, FINLAND
E-mail: Ramunas.Girdziusas@hut.fi, Jorma.Laaksonen@hut.fi

***Abstract*— Nonlinear diffusion filtering can be improved if viewed as Bayesian Gaussian process regression. We relate the covariance functions of the diffusion process outcome to the spatial diffusion operator and show how Bayesian evidence criterion can be utilized to determine the parameters of the nonlinear diffusivity and the optimal diffusion stopping time. Computational example is given where the nonlinear diffusion filtering outperforms typical Gaussian process regression.**

I. Introduction

Nonlinear diffusion of noisy discontinuous data is especially useful when filtering very large observation sets with low input dimension. It avoids $O(n^2 \dots n^3)$ complexity common in support vector machines and Gaussian process regression in general. Additive operator splitting (AOS) implementations require $O(n)$ number of multiplications to estimate the optimal regressor, and they scale well with increasing input dimensionality [14], [3]. The values of the signal or its derivatives on the boundary of the data domain, the periodicity and discontinuities can be utilized very effectively.

Many existing regression techniques, including multilayer perceptron networks, can be viewed as parametric representations of the mean and covariance functions of the corresponding Gaussian process (GP) model. In this light the paradigm of nonlinear diffusion filtering is a source of efficient covariance functions and an extension to the GP regression. Intuitively, we can think of diffusion filtering as either (i) an iterated Gaussian process regression with the covariance as the Green's function of the spatial part of the diffusion operator or (ii) the standard Gaussian process regression, but with a nonlinear covariance function.

The analysis of nonlinear diffusion, however, is very limited as neither Green's functions, nor the source solutions of the nonlinear diffusion equation are generally known. Few exceptions exist, e.g. *fast diffusions*, where the diffusivity is a polynomial function of the diffused quantity. As a consequence, correlation properties of the nonlinear diffusion outcome are unknown and obtaining good regression results is often difficult. Careful experimenting is required when determining optimal time evolution of a variety of key parameters which strongly affect the regression outcome. The absence of rigorous criteria for these parameters is not the only problem. Diffusion filtering is always iterative, and whereas the outcome of linear filtering might tend towards the average value of the signal, the issue of convergence of the nonlinear diffusion outcome to the true signal is more subtle [6]. Therefore, determination of the stopping time is crucial.

There are two principally different approaches in constructing nonlinear diffusion filtering. The classical method solves the boundary value problem of the second order parabolic partial differential equation (PDE) whose initial solution is formed by a *complete set* of noisy observations [14]. Numerous stable and efficient algorithms that work on massive meshes are known, but the diffusion in this framework is usually performed until the steady state is reached or it is stopped by utilizing heuristic means [9].

Another view focuses on the diffusion of a *single observation*. For this purpose, one postulates the infinitesimal drift and variance functions and then utilizes the framework of stochastic differential equations (SDE) [2]. The classical diffusion equations now have to be solved for the probability density of the values of a chosen observation. Such a methodology is important in applications which demand higher-dimensional inputs and would be especially valuable when performing regression at few locations without obtaining a complete solution. A particular improvement to this framework is the application of Girsanov's formula in deriving Bayes factors for the identification of the drift and variance functions of several well-established diffusion processes [11]. However, this methodology cannot be easily extended to the case of nonlinear diffusion filtering.

We believe that diffusion filtering can be optimized by considering the joint probability density of both: the diffusion outcome and its parameters. This can be achieved by viewing the classical diffusion filtering as Bayesian GP regression [10]. For this purpose, Section II-A first states a common numerical implementation of the diffusion equations based on the AOS scheme. Sect. II-B then briefly discusses the GP framework and provides its simple interpretation from an unconstrained quadratic minimization viewpoint. Sect. III bridges the GP regression and diffusion filtering by providing a relationship between the GP kernel and the propagated-in-time discrete version of the spatial diffusion operator. Sect. IV shows an example of the correlation functions of the linear and nonlinear diffusion outcomes. The Bayesian evidence criterion is then employed to determine the optimal diffusion stopping time. Speed and accuracy comparison between the typical GP regression models and nonlinear diffusion filtering is also presented. Conclusions are drawn in Sect. V.

II. Massive Nonlinear Diffusion Filtering and Fixed-Input Gaussian Process Regression

A. Semi-Implicit AOS-based Diffusion Filtering

For the sake of brevity, let us directly start with the semi-implicit AOS-based numerical algorithm of the isotropic nonlinear diffusion filtering [14]. Initially, it sets the model output equal to the observations, i.e. $u_0(\cdot) = y(\cdot)$, and then proceeds by iterating three basic steps:

$$\nabla u_\sigma = \nabla(u_k * h_\sigma), \tag{1}$$

$$\mathbf{D} = \big[1 - \exp(-c(\frac{\lambda}{||\nabla u_\sigma||})^s)\big]\mathbf{I}, \tag{2}$$

$$u_{k+1} = u_k + \tau\nabla\cdot\mathbf{D}\nabla u_{k+1}, \tag{3}$$

$$\text{s.t. } (\mathbf{D}\nabla u_{k+1}\cdot\mathbf{n}) = 0 \text{ on } \partial\Omega. \tag{4}$$

Here the signal u_t is first pre-filtered with a Gaussian filter h_σ of variance σ in order to obtain an estimate of its gradient. The diffusivity $\mathbf{D} \equiv \mathbf{D}(\nabla u_\sigma) \in \mathbb{R}^{d\times d}$ is next evaluated, whose parameter c can always be chosen so that the smoothing of u will take place only in the regions of a small Euclidean norm of the gradient, i.e. $||\nabla u_\sigma|| < \lambda$, thus preserving fronts and discontinuities [3], [9]. The specific choice of the boundary conditions in Eq. (4) will be clarified in Sect. III-A, but otherwise it can be arbitrary. The variance σ, the contrast parameter λ, the sharpness of nonlinearity s, the time step size τ and the termination time $m \equiv k_{\text{term}}$ should be specified beforehand.

Eqs. (1)-(3) state a paradoxical way to sharpen the edges of the signal by applying the diffusion equations, which are normally thought of as the process of smoothing, obtainable in the case of $\mathbf{D} = \mathbf{I}$ [4]. Analytical solutions are very much unknown, and the remaining text is devoted to the interpretation and better utilization of Eqs. (1)-(3).

Eq. (3) can be solved by employing the AOS principle [14]. If we gather the values of the signal $\mathbf{u} = \{u_i | i = 1, 2, \ldots, n, u_i \in \mathbb{R}^1\}$ given at n spatial locations $\mathbf{x} = \{\mathbf{x}_i | i = 1, 2, \ldots, n, \mathbf{x}_i \in \mathbb{R}^d\}$, then Eq. (3) becomes

$$\mathbf{u}_{k+1} = \frac{1}{d}\sum_{l=1}^{d}\big(\mathbf{I} - \tau d\mathbf{A}_l(\mathbf{u}_k)\big)^{-1}\mathbf{u}_k, \tag{5}$$

The matrix $\mathbf{A}_l$ approximates the second order derivatives of the function u along the space axis l. Numerical derivatives are weighted according to the diffusivity values and they essentially depend on the discrete space step h. Eq. (5) approximates Eq. (3) with $o(\tau)$ and $o(h^2)$ accuracy. It is efficient, because the matrices $\mathbf{A}_l$ are tri-diagonal, and the products with their inverses can be estimated by applying only $O(n)$ number of multiplications via the Thomas algorithm [14], [3]. The numerical stability holds for all step sizes $\tau > 0$, but the condition number of the diagonally dominant matrices $\mathbf{I}-\tau\mathbf{A}_l$ increases at large values of τ [14].

B. Gaussian Process Model

Let us now postulate the joint probability density of the observation vector $\mathbf{y} \in \mathbb{R}^n$ and the diffusion outcome $\mathbf{u} \in \mathbb{R}^n$:

$$\begin{pmatrix} \mathbf{u} \\ \mathbf{y} \end{pmatrix} \sim \mathcal{N}\Big(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{K}_\theta & \mathbf{K}_\theta \\ \mathbf{K}_\theta & \mathbf{K}_\theta + \theta_0\mathbf{I} \end{pmatrix}\Big). \tag{6}$$

Eq. (6) assumes that each value of the measurement vector $\mathbf{y}$ is contaminated by additive Gaussian noise, whose variance is denoted by θ_0, and the marginal distribution $p(\mathbf{u})$ is a zero mean Gaussian process [10], defined by a symmetric matrix $\mathbf{K}_\theta \in \mathbb{R}^{n\times n}$. The next section will give the precise meaning to this matrix whose parameters $\boldsymbol{\theta}$ will appear as the unknown diffusion quantities.

GP regression then proceeds by estimating a Gaussian conditional density function $p(\mathbf{u}|\mathbf{y})$ with the following mean and covariance [8]:

$$\bar{\mathbf{u}} \equiv E[\mathbf{u}|\mathbf{y}] = \mathbf{K}_\theta(\mathbf{K}_\theta + \theta_0\mathbf{I})^{-1}\mathbf{y}, \tag{7}$$

$$Cov[\mathbf{u}|\mathbf{y}] = \mathbf{K}_\theta - \mathbf{K}_\theta(\mathbf{K}_\theta + \theta_0\mathbf{I})^{-1}\mathbf{K}_\theta. \tag{8}$$

Here we assume that the same input locations correspond to the measurements and estimated signal. This is not a limitation of GP regression, but it simplifies the equations. It is important to notice that Eq. (7) also solves

$$\bar{\mathbf{u}} = \arg\min_{\mathbf{u}}\big(\frac{1}{2\theta_0}||\mathbf{u}-\mathbf{y}||^2 + \frac{1}{2}\mathbf{u}^T\mathbf{K}_\theta^{-1}\mathbf{u}\big). \tag{9}$$

Therefore, an intuition behind the GP regression is simply to solve the regularization problem by minimizing the Euclidean $||\cdot||^2$ norm between the data and model output while maintaining certain regularity properties of the model outputs via their quadratic form determined by the inverse kernel matrix $\mathbf{K}_\theta^{-1}$. As we will see later, this is directly related to the variational form of the diffusion equations.

In general, the model Eq. (6) should be completed so that the averaging in Eq. (7) could be performed also w.r.t. the density of the unknown diffusion parameters $\boldsymbol{\theta}$ or simply *hyperparameters*. However, as their number is small, it is convenient to postulate a noninformative (possibly improper) prior and perform the approximate Bayesian regression with a single parameter setting $\boldsymbol{\theta}^*$:

$$p(\mathbf{u}|\mathbf{y}) = \int p(\mathbf{u}|\mathbf{y},\boldsymbol{\theta})p(\boldsymbol{\theta})\mathbf{d}\boldsymbol{\theta} \approx p(\mathbf{u}|\mathbf{y},\boldsymbol{\theta}^*), \tag{10}$$

where the hyperparameters $\boldsymbol{\theta}^*$ can be chosen to maximize the loglikelihood (logevidence of the observations) [13]:

$$\ln p(\boldsymbol{\theta}|\mathbf{y}) = -\frac{1}{2}\ln|\mathbf{C}_\theta| - \frac{1}{2}\mathbf{y}^T\mathbf{C}_\theta^{-1}\mathbf{y} - \frac{n}{2}\ln 2\pi. \tag{11}$$

Here $\mathbf{C}_\theta \equiv \mathbf{K}_\theta + \theta_0\mathbf{I}$. The gradient-based optimization circumvents the evaluation of the log-determinant in Eq. (11), but also the Monte Carlo sampling of the determinant values can be employed [13]. Section IV-B will attempt to describe the role of Eq. (11) in optimal diffusion stopping, but we have observed that it is also useful in determining the diffusivity parameters λ and s as well. Normally, they would have to be chosen in an *ad hoc* way.

III. Two Views on Nonlinear Diffusion Filtering

A. Green's Function Viewpoint

A greater picture comes into view after interpreting the diffusion as an energy balance. For this purpose, Eq. (5) can be alternatively expressed as

$$\mathbf{u}_{k+1} = \frac{1}{d}\sum_{l=1}^{d}\mathbf{u}_{k+1,l}, \tag{12}$$

where $\mathbf{u}_{k+1,l} \in \mathbb{R}^n$ denotes $\mathbf{u}_k$, smoothed along the l-th axis:

$$\mathbf{u}_{k+1,l} = \arg\min_{\mathbf{u}}\big(\frac{1}{2\tau}||\mathbf{u}-\mathbf{u}_k||^2 - \frac{1}{2}\mathbf{u}^T\mathbf{A}_l(\mathbf{u}_k)\mathbf{u}\big). \tag{13}$$

This can be obtained by applying basic rules of matrix calculus to Eq. (5), but its understanding comes through multiplication of the Eq. (3) by u_{k+1} followed by the integration over the domain $\mathbf{x} \in \Omega$ with the employment of the Green's identity:

$$\int_\Omega u\nabla\cdot(\mathbf{D}\nabla u) = \int_{\partial\Omega} u(\mathbf{D}\nabla u\cdot\mathbf{n}) - \int_\Omega(\nabla u)^T\mathbf{D}\nabla u. \tag{14}$$

The key point here is that the negative term in Eq. (13) should not be mistaken for 'de-regularizing' as in the total least squares framework. Under the boundary conditions, stated in Eq. (4), the quadratic form in Eq. (13) is discrete counterpart of its continuous form in Eq. (14), whose purpose is to penalize non-Euclidean norm of the gradient ∇u.

Comparing Eq. (13) to Eq. (9), we can see that the diffusion filtering step is the same as the GP regression, applied iteratively along each space dimension l with the kernel

$$\tilde{\mathbf{K}}_l^{-1} = -\mathbf{A}_l(\mathbf{u}_k). \tag{15}$$

Eq. (15) indicates that the covariance functions in the matrix $\tilde{\mathbf{K}}_l$ are determined by the nonuniform Green's functions of the l-th component of the spatial diffusion operator $\mathcal{L} = \partial_{x_l}D_{ll}\partial_{x_l}$ with the corresponding boundary conditions in Eq. (4). This can be seen by applying the analytical formulae of the inverses of the tridiagonal matrices [16].

This connection emphasizes the classical notion of 'diffusion kernel' and it seems that it is more useful to the GP regression than to the nonlinear diffusion filtering on two major accounts. Green's functions present a method to choose GP kernels in the boundary value problems, and a single regression with the GP kernel $\tilde{\mathbf{K}}_l$ is very efficient. The difference in computational complexity between the typical GP regression and the regression with diffusion kernel is huge: $O(n^2\dots n^3)$ vs. $O(n)$ number of scalar multiplications.

In addition, the principal contribution of Eq. (15) lies in the possibility to iterate the GP regression and as we will indicate in Section IV-A, the covariance properties of the diffusion outcome can be very different from the ones seen in the Green's functions. Here it is also worth mentioning that the diffusion kernels find even broader range of applications than filtering [5].

B. A More Complete View of Diffusion Filtering

The notion of the 'diffusion kernel' as the Green's function is different from the general meaning of the kernel in the GP framework. The first case assumes the GP regression applied in a repetitive manner which smooths the initial data along each data dimension and at each time iteration according to Eq. (13). On the other hand, we can ask the question about the corresponding GP kernel which embodies the covariance between the outputs of the finally diffused quantity at two arbitrary spatial locations.

Such GP model can be deduced by first expressing the final diffusion outcome at the time index $k = m$ through the initial solution $u_0 = y$ according to Eq. (5) and then relating the corresponding operator to Eq. (7). The precise connection between the GP regression and nonlinear diffusion filtering can then be written compactly:

$$\mathbf{K}_\theta(\mathbf{K}_\theta + \theta_0\mathbf{I})^{-1} = \prod_{k=m}^{1}\frac{1}{d}\sum_{l=1}^{d}\big(\mathbf{I} - \tau d\mathbf{A}_l(\mathbf{u}_k)\big)^{-1}. \tag{16}$$

Therefore, nonlinear diffusion filtering is a GP regression with a particular kernel which is related to the weighted finite difference matrices $\mathbf{A}_l$ or the diffusion kernel in Eq. (15) in a nonlinear way.

Eq. (16) has infinitely many solutions, in particular, it is easy to see that any scaling of the solution, i.e. the kernel matrix $\mathbf{K}_\theta$ and the noise level θ_0 leaves Eq. (16) invariant. However, if we assume that the noise level can be reasonably guessed or estimated, then the corresponding GP kernel matrix can be computed according to

$$\mathbf{K}_\theta = \theta_0(\mathbf{I} - \mathbf{A})^{-1}\mathbf{A}. \tag{17}$$

Here the matrix $\mathbf{A}$ denotes the right side of Eq. (16). This equation is important for two reasons: (i) it directly links the regularization operators of linear and nonlinear diffusion with the correlation properties of the diffusion outcome, and (ii) it indicates that diffusion filtering is suboptimal if its parameters do not optimize any sensible probabilistic criteria.

An important problem in diffusion filtering is determining the optimal diffusion stopping time [9], [2]. The iteration according to Eq. (5) creates the so called *scale space* of the diffusion outcomes with progressively smaller details (linear diffusion) or smoothed 'staircasing' (if Eq. (2) is utilized). The latter does not necessarily converge to the true solution.

A particular recently discussed insight [9] to optimal diffusion stopping is based on the assumption that the true signal and the noise is uncorrelated. Therefore, the diffusion can be stopped when the correlation between the vectors $\mathbf{u}_0 - \mathbf{u}_m$ and $\mathbf{u}_m$ is minimal. Eq. (11) will also embody the 'decorrelation' if one notices that $\mathbf{y}^T\mathbf{C}_\theta^{-1}\mathbf{y} = \frac{1}{\theta_0}\bar{\mathbf{u}}^T\mathbf{C}_\theta\mathbf{K}_\theta^{-1}(\mathbf{y} - \bar{\mathbf{u}})$. The main difference is in that Bayesian evidence Eq. (11) will 'decorrelate' only at small values of noise variance θ_0. We can see that Eq. (11) clarifies and extends the decorrelation-based stopping by using the explicit probabilistic assumptions of Section II-B.

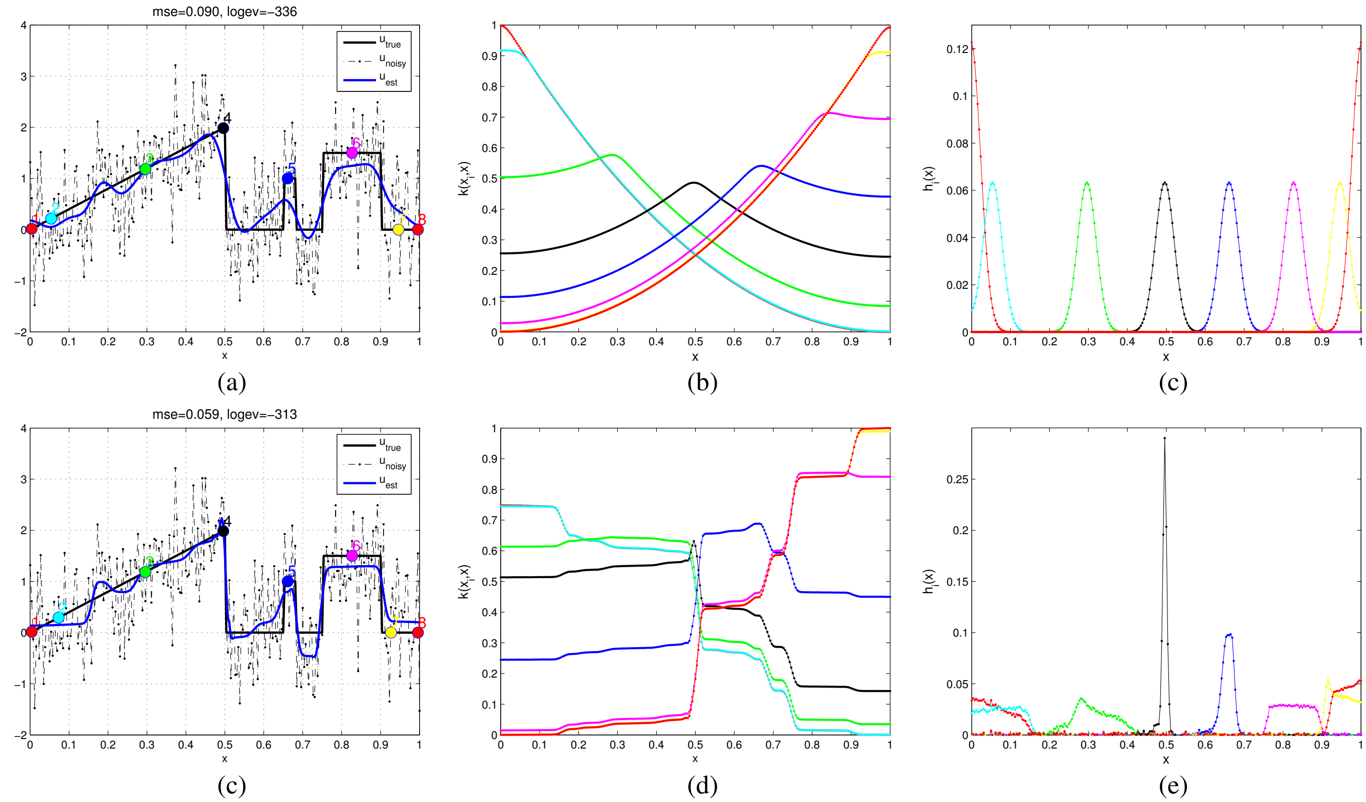


Fig. 1. Diffusion filtering from Gaussian process viewpoint: (a) indicates the outcome of the linear diffusion, (b) the covariance functions at eight locations indicated in the plot (a), (c) the corresponding smoothing factors, which in the linear diffusion filtering represent a simple weighted averaging, (d) indicates the result of the non-linear diffusion filtering, (e) shows the corresponding covariance kernels of the nonlinear diffusion, whereas (f) states its smoothing factors which depend on the discontinuities of the signal.

IV. Computational Analysis

A. What Do GP Kernels of Diffusion Filtering Look Like?

Consider the test problem of estimating discontinuous signal from its observations contaminated with additive Gaussian noise. The analysis of similar problem in the non-optimized diffusion case can be found in [7]. The outcome of linear and nonlinear diffusion u_m, a sample of the corresponding GP covariances $\mathbf{K}_\theta$ at eight chosen locations and their smoothing factors $\mathbf{h} = \mathbf{K}_\theta(\mathbf{K}_\theta + \theta_0\mathbf{I})^{-1}$ are shown in Fig. 1. The actual noise level of the signal is set to $\theta_0 = 0.44$, but we assume that in a realistic situation it would be unknown. Therefore, we set it to the estimate $\tilde{\theta}_0 \approx 0.538$ which can be obtained by a typical GP regression model with the isotropic Gaussian kernel function. This value is utilized when applying Eq. (16) and (17). Fig. 1 indicates that by imposing the reflecting boundary conditions stated in Eq. (3), one assumes that the outcome of the linear and nonlinear diffusion filtering will have more correlated values in the vicinity of the boundaries. Notice that the smoothing factors of nonlinear diffusivity Eq. (2) follow the discontinuities of the signal. The correlation functions of both processes are positive, which can be advocated to the fact that diffusions are Markovian [11].

B. Evidence Maximizing Diffusion Stopping

Fig. 2 indicates that Bayesian evidence criterion is helpful when determining the optimal diffusion stopping time. Notice how the outcome of the nonlinear diffusion, mean squared error (MSE) values and the logevidence match each other. MSE achieves its minimum and Bayesian evidence its maximum at $m \in [10, 20]$. We have also observed that the sharpness of nonlinearity $s = 2$ and the gradient threshold $\lambda = 0.01$ represent global maximum of the evidence criterion in this case. A more detailed analysis of the optimal diffusion stopping is summarized in Figs. 3a-c. We have observed that at the adequate noise levels, such as $\theta_0 = 0.25 \cdots 0.5$, the evidence criterion yields best filtering results, whereas the decorrelation criterion [9] overestimates the optimal stopping time, especially at larger noise variances, where it tends towards oversmoothing. As Section III-B has already predicted, when the noise variance is small, decorrelation can be close to Bayesian evidence maximization. We observe that both criteria are rather flat in this case.

C. Comparison with Typical GP Regression Models

A comparison of several frequently used GP regression models [12] and the diffusion filtering is presented in Table I. Here we first apply Bayesian GP regression with the Gaussian kernel, Brownian motion covariance $\min(x_i, x_j)$, the integrated step-function kernel $1 + \theta_1 - 2\theta_1|x_i - x_j|$ and the Laplacian kernel $\exp(-\theta_1|x_i - x_j|)$. The latter free GP models yield similar results, with the best one indicated in Table I. The question whether the chosen covariance function

Fig. 2. Nonlinear diffusion outcome with different terminating times m. Filtering can be stopped at the maximum of the Bayesian evidence criterion.

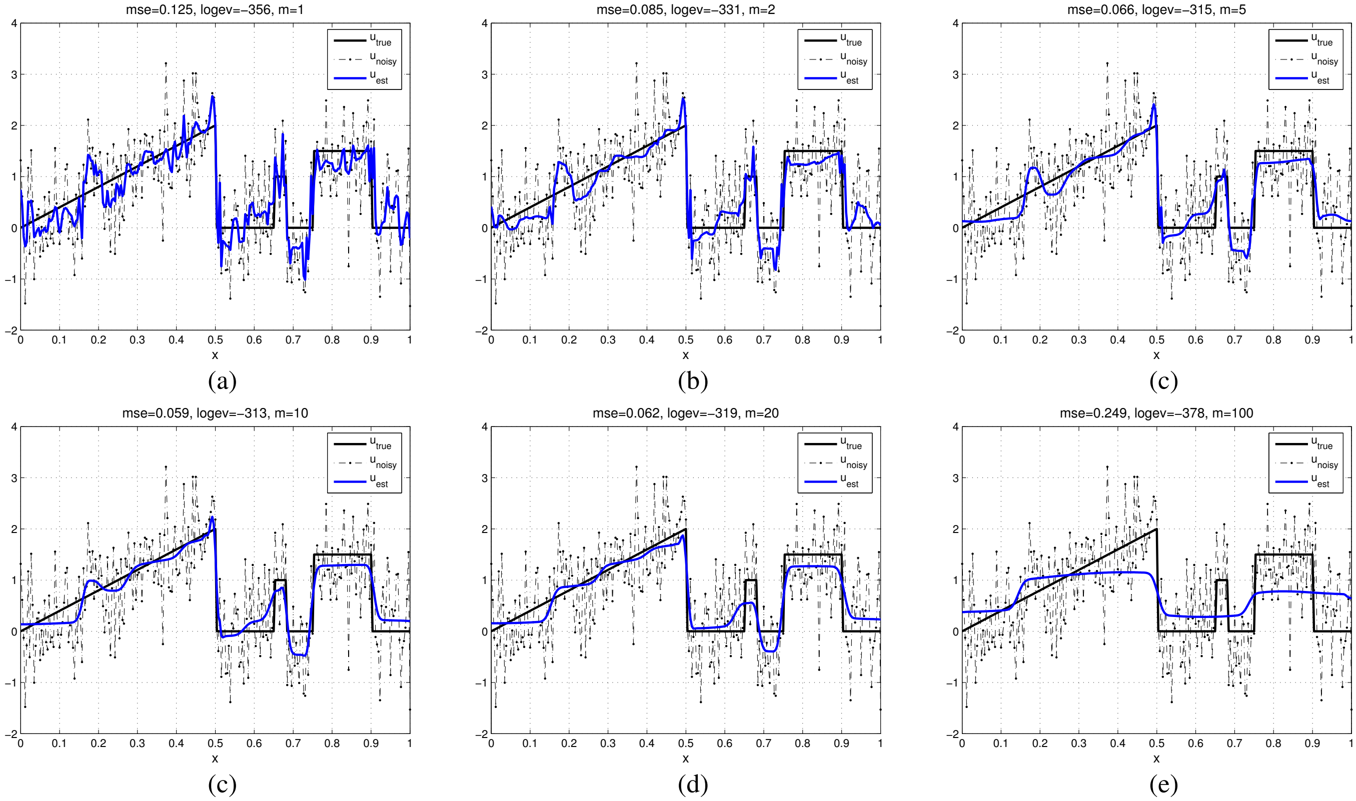


TABLE I

REGRESSION RESULTS

	$n = 260$		$n = 1000$	
Method	mse	time, s.	mse	time, s.
GP, Gaussian kernel	0.0932	3.87	0.0501	113
GP, neural network kernel [15]	0.0789	8.48	0.0382	218
Linear diffusion	0.0905	0.20	0.0378	0.23
GP, Laplacian kernel	0.0792	0.32	0.0375	0.4
Nonlinear diffusion	0.0588	0.21	0.0247	0.25

is stationary (Laplacian kernel) or nonstationary (covariance of the Brownian motion) can be irrelevant to the regression task. More importantly, the covariance functions should approximately follow sharp changes of the true signal hidden in its noisy observations. To make a comparison more complete, we have also experimented with the covariance function of the multilayer perceptron, which can be obtained in the limit of an infinite number of hidden neurons with the 'erf' activation functions [15]. Table I shows that increasing the number of measurements from $n = 260$ to $n = 1000$ results in a significant increase of the computational complexity of the GP regression in the case of gaussian and the neural network kernel functions.

The regression outcome of the best representative of the GP models, namely the one with the Laplacian kernel, and also the optimized linear and nonlinear diffusion filtering can be seen in Fig. 3d. Clearly, the diffusion filtering better preserves the sharp changes of the signal, which is also supported by the MSE values, shown in Table I.

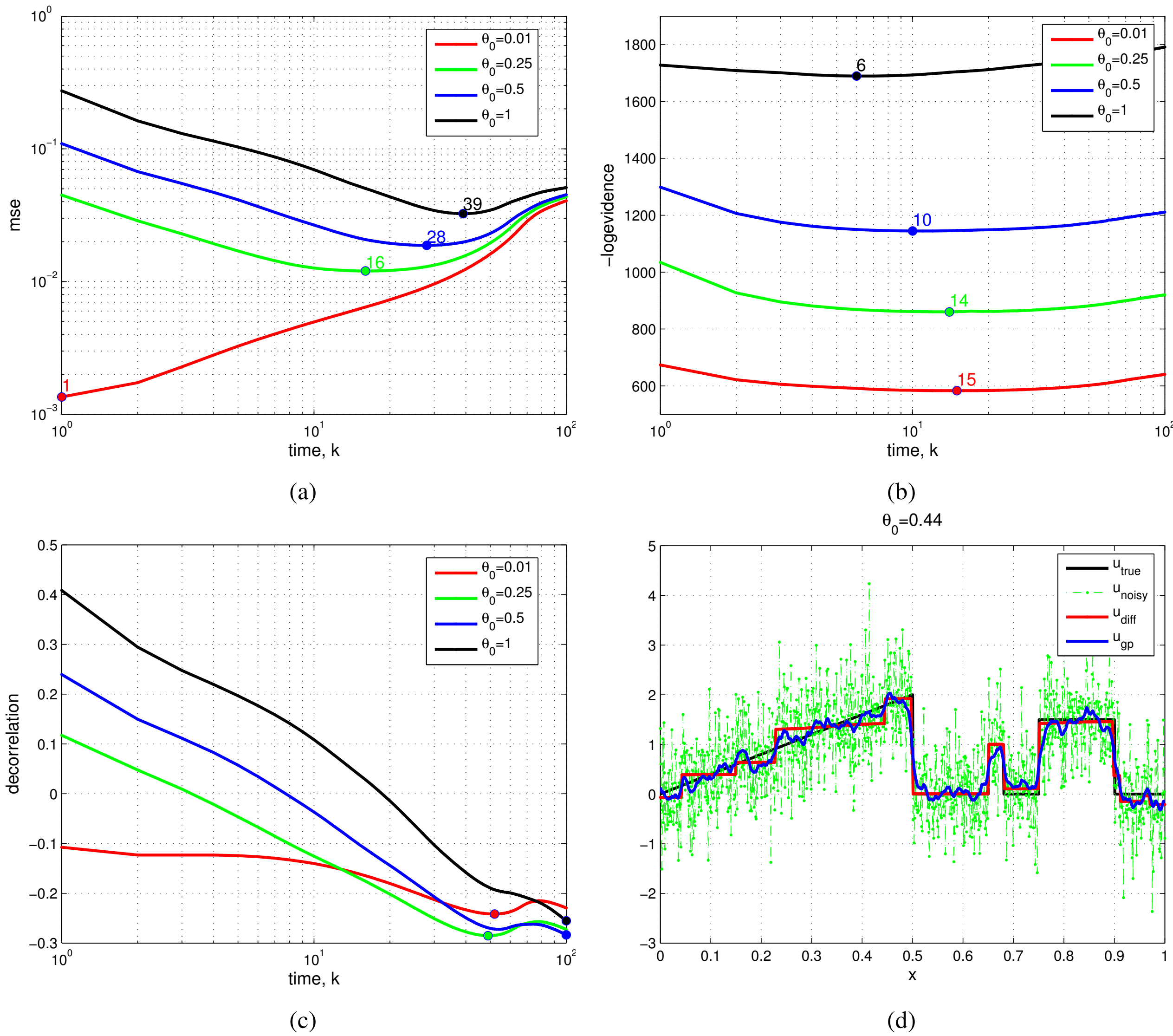
The diffusion filtering could be continued even in the cases $n > 10^5$, but the evaluation of the log-evidence requires further numerical analysis. The condition number of the AOS matrix $\mathbf{A}$ is already 10^{15} in the case $n = 260$. Therefore, Eq. (17) requires careful implementation.

Finally, it is important to emphasize that a typical GP regression with Gaussian covariance functions will be very hard to implement in a large scale problem, whereas Laplacian kernel does not present significant difficulty as it yields tridiagonal inverse covariance matrix, e.g. see [1]. Nonlinear diffusion filtering, on the other hand, represents a much broader class of efficiently implementable covariance functions. They can be Green's functions of Eq. (16) or, more generally, iterated Green's functions in Eq. (17), which are especially useful when recovering signals with discontinuities.

V. CONCLUSIONS

Nonlinear diffusion filtering presents a source of efficient covariance functions useful in regression analysis. In this work we have explored the connection between the discrete AOS regularization operators of the nonlinear diffusion filtering and their corresponding Gaussian process kernels. Depiction of these kernels provides a visual insight into mathematics of diffusion filtering, whereas the utilization of Bayesian evidence helps to clarify the model assumptions and is particularly useful in determination of the parameters of the nonlinear

Fig. 3. Optimal diffusion stopping according to: (a) MSE evolution when a true signal is known, (b) minima of the negative logevidence and (c) decorrelation criterion. The comparative results with the best typical GP regression model (d).



diffusivity function and optimal diffusion stopping. It improves heuristic practical principles such as diffusion stopping by decorrelation of the signal and noise values.

References

[1] P. Attard, O. G. Jepps, and S. Marčelja. Information content of signals using correlation function expansions of entropy. *Physical Review E*, 56(4):4052–4067, 1997.

[2] Y. Bao and H. Krim. Smart nonlinear diffusion: A probabilistic approach. *IEEE Trans. on PAMI*, 26(1):63–72, January 2004.

[3] Frederico D'Almeida. Nonlinear diffusion toolbox. MATLAB Central.

[4] S. Kichenassamy. The Perona–Malik paradox. *SIAM J. Appl. Math*, 57(5):1328–1342, October 1997.

[5] J. Lafferty and G. Lebanon. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6:129–163, 2005.

[6] M. Lassas and S. Siltanen. Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Problems*, 20:1537–1563, October 2004.

[7] A. Marquina and S. Osher. Explicit algorithms for a new time dependent model based on level set motion for nonlinear deblurring and noise removal. *SIAM J. SCi. Comput.*, 22(2):387–405, 2000.

[8] T. P. Minka. From hidden Markov models to linear dynamical systems. Technical report no. 531, Carnegie Mellon University, 1999.

[9] P. Mrázek and M. Navara. Selection of optimal stopping time for nonlinear diffusion filtering. *Int. Journal of Computer Vision*, 52(2):189–203, 2003.

[10] A. O'Hagan and J. F. C. Kingman. Curve fitting and optimal design for prediction. *J. R. Statist. Soc. B*, 40(1):1–42, 1978.

[11] N. G. Polson and G. O. Roberts. Bayes factors for discrete observations from diffusion processes. *Biometrika*, 8(1):11–26, March 1994.

[12] C.E. Rasmussen. *Evaluation of Gaussian Processes and Other Methods for Non-linear Regression*. Ph.d. thesis, University of Toronto, 1996.

[13] J. Skilling. Bayesian numerical analysis. In *Physics and Probability*, pages 207–221. Cambridge Univ. Press, 1993.

[14] J. Weickert, B. M. ter Haar Romeny, and M. A.Viergever. Efficient and reliable schemes for nonlinear diffusion filtering. *IEEE Trans. on Image Processing*, 7(3):398–410, March 1998.

[15] C. K. I. Williams. Computation with infinite neural networks. *Neural Computation*, 10(5):1203–1216, 1998.

[16] T. Yamamoto. Inversion formulas for tridiagonal matrices with applications to boundary value problems. *Num. Func. Anal. and Opt.*, 22(3):357–385, 2001.