Probabilistic Reconstruction of Lithofacies with Support Vector Machines

Nutchapol Dendumrongsup
Energy Resource Engineering Department
Stanford University, Stanford, CA 94305, USA

Daniel M. Tartakovsky
Professor, Energy Resource Engineering Department
Stanford University, Stanford, CA 94305, USA

Abstract—**Delineation of geological features from limited hard and/or soft data is crucial to predicting subsurface phenomena. Ubiquitous sparsity of available data implies that the reliability of any delineation effort is inherently uncertain. We present probabilistic support vector machines (pSVM) as a viable method for both lithofacies delineation from sparse data and quantification of the corresponding predictive uncertainty. Our numerical experiments demonstrate an agreement between the probability of a pixel classifier predicted with pSVM and indicator Kriging. While the latter requires manual inference of a variogram (two-point correlation function) from spatial observations, pSVM are highly automated and less data intensive. We also investigate the robustness of pSVM with respect to its hyper-parameters and the number of measurements.**

I. Introduction

The need to delineate geological features, e.g., lithofacies, is ubiquitous in subsurface application. Understanding subsurface geology is crucial to exploration of mineral resources and oil and gas reservoirs. It is also of central importance in subsurface hydrology since geologic makeup of the subsurface plays a crucial role in fluid flow and contaminant transport. A typical example is a problem of locating permeable zones in an aquiclude that separates two aquifers, the upper aquifer contaminated with industrial and/or agricultural pollutants, and the lower aquifer used for municipal water supplies [1].

Geostatistics has long been used to gain insight into spatial distributions of physical properties of geologic formations [2]. By adapting the probabilistic framework it accounts for inherent predictive uncertainty that arises from subsurface heterogeneity and data sparsity. In doing so, geostatistics relies on the ergodicity hypothesis, which postulates the equivalence between spatial statistics and its ensemble counterpart. This hypothesis cannot be proven, but it does require a subsurface environment to be (weakly) stationary, i.e., relevant subsurface properties to have both constant means and variances and translation-invariant correlation functions. Tools of statistical learning theory [3], such as support vector machines (SVM) [4], provide an attractive alternative to geostatistics because it does not invoke ergodicity, is highly automated, and requires fewer measurements to remain viable [5].

The latter feature is a key advantage that distinguishes SVM from other machine learning techniques, e.g., deep neural networks, used for facies delineation [6]. Unlike neural networks, SVM possess rigorous performance guarantees and error bounds [3]. Moreover, robust uncertainty quantification for neural networks remains elusive, due to their limited interpretation capacity. For these reasons, we adapt probabilistic SVM (pSVM) [7] as a means to delineate geological facies from sparse data to quantify corresponding predictive uncertainty.

Like many other machine learning techniques, standard SVM [5] provides only a "best" estimate of the spatial arrangement of geological features consistent with available data. The use of pSVM enables us to quantify uncertainty inherent in such reconstructions and to identify facies with a required degree of fidelity. While the original pSVM [7] were designed to quantify the probability of SVM misclassifying pixels of a complete image, ours aim to cope with the sparsity of subsurface data, i.e., with the task of reconstructing an image from a few pixels.

In section II, we formulate the problem of subsurface facies delineation from sparse data. Section III contains a brief overview of the standard SVM approach to this problem and introduces pSVM. The latter is used in section IV to probabilistically reconstruct facies of a synthetic dataset, which enables us to study the method's accuracy, robustness, and data requirements. Main conclusions drawn from this study are summarized in section V.

II. Problem of Facies Delineation

Consider a subsurface environment $D \in \mathbb{R}^2$ consisting of two lithofacies M_1 and M_2 ($D =$

$M_1 \cup M_2$), e.g., high- and low-permeability heterogeneous geologic materials. Our goal is to reconstruct a (single- or multi-connected) boundary between these facies from continuous parameter data $K_i = K(\mathbf{x}_i)$ collected at N locations, $\mathbf{x}_i = (x_i, y_i)^\top$ with $i \in \{1, \cdots, N\}$, throughout the domain D. These locations form a set $T = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$.

To transform this task into a classification problem, we convert values K_i of the continuous function $K(\mathbf{x}_i)$ into values of an indicator function (aka categorical variable)

$$J(\mathbf{x}_i) = \begin{cases} 1 & \mathbf{x}_i \in M_1, \\ -1 & \mathbf{x}_i \in M_2. \end{cases} \qquad (1)$$

This step assumes that the data $\{K_i\}_{i=1}^N$ are well differentiated, i.e., each measurement K_i unambiguously identifies the measurement location $\mathbf{x}_i$ as belonging either to facies M_1 or M_2. In the case of poorly differentiated data, this step could be preceded by a nearest neighbor classifier [8].

III. Support Vector Machines

SVMs are often considered one of the best "out of the box" classifiers that yield great performance in a variety of settings [9]. In their simplest form, SVMs are applicable to linearly separable data, e.g., data collected from perfectly stratified geologic media in which different geologic facies are separated by either planes in three dimensions or straight lines in two dimensions. Nonlinear SVMs enable one to deal with general subsurface environments by projecting them into higher-dimensional space in which the data are linearly separable by a hyperplane. Linear and nonlinear SVMs are briefly reviewed in sections III-A and III-B for the sake of completeness.

A. Linear SVM

A linearly separable data set $\{J_i\}_{i=1}^N$ implies the existence of a straight line, $\mathbf{a} \cdot \mathbf{x} + b = 0$, that separates the locations $\mathbf{x}_i$ at which $J = -1$ from those at which $J = 1$ (Fig. 1). The unknown constants $\mathbf{a} = (a_1, a_2)^\top$ and b are computed by maximizing the distance (margin) between $\mathbf{a} \cdot \mathbf{x} + b = 0$ and the locations at which $J_i = -1$ and $J_i = -1$.

The solid line in Fig. 1 represents the optimum classifier line $\mathbf{a} \cdot \mathbf{x} + b = 0$, while the dotted lines indicate the extent of the margin, the region within which the boundary could be shifted orthogonally while preserving the perfect classifying accuracy [9]. (The points used to construct the margin are called the support vectors.) These two lines have the same slope as the classifier line (the vector $\mathbf{a}$) but differ by the intercepts, i.e., their equations are written as $\mathbf{a} \cdot \mathbf{x} + b = \pm 1$.

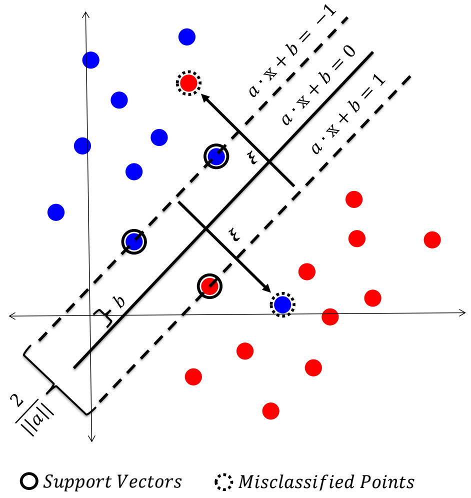
$a \cdot \mathbb{x} + b = -1$
$a \cdot \mathbb{x} + b = 0$
$a \cdot \mathbb{x} + b = 1$
ξ
ξ
b
$\frac{2}{||a||}$
○ Support Vectors
⁘ Misclassified Points

Fig. 1: A linear SVM classifier maximizes the margin between the reconstructed boundary (straight line), $\mathbf{a} \cdot \mathbf{x} + b = 0$, and the locations $\mathbf{x}_i$ at which $J_i = -1$ (blue circles) and $J_i = -1$ (red circles).

All the data points $\mathbf{x}_i$ ($i = 1, \cdots, N$) satisfy either $\mathbf{a} \cdot \mathbf{x} + b \geq 1$ or $\mathbf{a} \cdot \mathbf{x} + b \geq -1$. These inequalities are combined into one,

$$(\mathbf{a} \cdot \mathbf{x}_i + b) J_i \geq 1, \quad i = 1, \cdots, N. \qquad (2)$$

The inequality (2) becomes an equality for the support vectors $\mathbf{x}_i$. Let $a = |\mathbf{a}|$ denote the Euclidean length of $\mathbf{a}$. One can show (e.g., [5]) that the margin d is given by $d = 2/a$. The SVM identifies the values of $\mathbf{a}$ and b by maximizing the margin d or, equivalently, by minimizing $a/2$ subject to the linear constraints (2). Introducing Lagrange multipliers $\boldsymbol{\gamma} = \{\gamma_1, \cdots, \gamma_N\}$, this yields an optimization problem $\{\mathbf{a}^\star, b^*\} = \text{argmin}_{\mathbf{a},b,\boldsymbol{\gamma}} L$, where the objective function $L(\mathbf{a}, b, \boldsymbol{\gamma})$ is given by

$$L = \frac{a}{2} - \sum_{i=1}^{N} \gamma_i [(\mathbf{a} \cdot \mathbf{x}_i + b) J_i - 1]. \qquad (3)$$

The indicator function at points $\mathbf{x}$ where measurements are absent is given by

$$J(\mathbf{x}) = \text{sgn}(\mathbf{a}^\star \cdot \mathbf{x} + b^\star). \qquad (4)$$

It is usually referred to as a decision function in the SVM literature.

The linear SVM can be augmented to account for slight deviations from a perfectly linear classification boundary by introducing slack variables $\xi_i \geq 0$ ($i = 1, \cdots, N$). The linear SVM minimization problem is replaced with the problem of minimizing the objective loss function $a/2 + C\sum_{i=1}^{N} \xi_i$ subject to the constraints $(\mathbf{a} \cdot \mathbf{x}_i + b)J_i \geq 1 - \xi_i$ with $i = 1, \ldots, N$. Magnitude of the constant C determines the strength of the slack penalty. Introducing Lagrange multipliers $\gamma = \{\gamma_1, \cdots, \gamma_N\}$ and $\delta = \{\delta_1, \cdots, \delta_N\}$ for $i \in 1, \ldots, N$ gives an objective function similar to (3). This optimization problem is rewritten as $\{\mathbf{a}^\star, b^*, \boldsymbol{\xi}^\star\} = \text{argmin}_{\mathbf{a}, b, \boldsymbol{\xi}, \gamma, \delta} L_{\xi}$, where $L_{\xi}(\mathbf{a}, b, \boldsymbol{\xi}, \gamma, \delta)$ is defined as

$$L_{\xi} = L - \sum_{i=1}^{N} (\gamma_i + \delta_i - C)\xi_i, \tag{5}$$

with L given by (3). To facilitate the solution of this optimization problem, one converts it into its dual,

$$\gamma^\star = \text{argmax}_{\gamma}[\sum_{i=1}^{N} \gamma_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \gamma_i \gamma_j J_j J_j \mathbf{x}_i \cdot \mathbf{x}_j] \tag{6}$$

subject to constraints $0 \leq \gamma_i \leq C$ and $\sum_{i=1}^{N} \gamma_i J_i = 0$. Once $\gamma^\star$ is obtained, the solution of $\partial L_{\xi} / \partial a_k = 0$ ($k = 1, 2$) in (5) is

$$\mathbf{a}^\star = \sum_{i=1}^{N} \gamma_i^\star J_i \mathbf{x}_i. \tag{7}$$

Let $\mathbf{x}_n = \mathbf{x}_+$ and $\mathbf{x}_k = \mathbf{x}_-$, for some n and k, denote support vectors for which $J = 1$ and $J = -1$, respectively. For these support vectors, the SVM inequality constraints $(\mathbf{a} \cdot \mathbf{x}_i + b)J_i \geq 1 - \xi_i$ turn into equations, $\pm(\mathbf{a} \cdot \mathbf{x}_\pm + b) = 1 - \xi_\pm$ with $\xi_n = \xi_+$ and $\xi_k = \xi_-$. Their solution is

$$b^\star = -\frac{1}{2}(\mathbf{a}^\star \cdot (\mathbf{x}_+ + \mathbf{x}_-) - \xi_- + \xi_+). \tag{8}$$

Thus, a solution for the indicator function (4) is

$$J(\mathbf{x}) = \text{sgn}(\sum_{i=1}^{N} \gamma_i^\star J_i \mathbf{x}_i \cdot \mathbf{x} + b^\star). \tag{9}$$

B. Nonlinear SVM

Boundaries of lithofacies in the subsurface are rarely, if ever, planes (straight lines). Hence, parameter data $\{K_i\}_{i=1}^{N}$ or its indicator counterpart $\{J_i\}_{i=1}^{N}$ belonging to different lithofacies cannot be separated by the line $\mathbf{a}^\star \cdot \mathbf{x} + b^\star = 0$ in $d = 2$ or 3 spatial dimensions. Fortunately, it has been proven [3] that there exists a higher-dimensional space (whose dimension m is generally unknown) in which the data become linearly separable. Let $\mathbf{F} : \mathbb{R}^d \to \mathbb{R}^m$ denote a map of the d-dimensional physical space onto that m-dimensional space (known as a feature space). In other words, every point $\mathbf{x} \in \mathbb{R}^d$ corresponds to a point $\hat{\mathbf{x}} \in \mathbb{R}^m$, such that $\hat{\mathbf{x}} = \mathbf{F}(\mathbf{x})$. The linear SVM in $\mathbb{R}^m$ separates the data by a hyperplane $\hat{\mathbf{a}}^\star \cdot \hat{\mathbf{x}} + b^\star = 0$, whose coefficients $\hat{\mathbf{a}}^\star \in \mathbb{R}^m$ and $b^\star \in \mathbb{R}$ are determined from the transformed dataset $\{\hat{\mathbf{x}}_i, J_i\}_{i=1}^{N}$. This is accomplished by solving the quadratic optimization of the linear SVM (3) and (5) in which $\mathbf{a}$ and $\mathbf{x}$ are replaced with $\hat{\mathbf{a}}$ and $\hat{\mathbf{x}}$. Similar to (4), the indicator function is given by $J(\mathbf{x}) = \text{sgn}(\hat{\mathbf{a}}^\star \cdot \mathbf{F}(\mathbf{x}) + b^\star)$.

While this indicator function is linear in the m-dimensional feature space, it corresponds to a nonlinear function in the physical space, whose specific form being determined by the mapping $\mathbf{F}$. The latter is proven to exist, but its form is generally known and, hence, $J(\mathbf{x})$ is not directly computable. Instead, one solves the dual constrained optimization problem (6) with $\mathbf{x}_i$ and $\mathbf{x}_j$ replaced by $\hat{\mathbf{x}}_i = \mathbf{F}(\mathbf{x}_i)$ and $\hat{\mathbf{x}}_j = \mathbf{F}(\mathbf{x}_j)$. The resulting inner product of the mapping functions, $\mathbf{F}(\mathbf{x}_i) \cdot \mathbf{F}(\mathbf{x}_j)$, remains uncomputable and is replaced with an empirical function called a Mercer kernel, $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) \equiv \mathbf{F}(\mathbf{x}_i) \cdot \mathbf{F}(\mathbf{x}_j)$. Examples of Mercer kernels include polynomials, sigmoid functions (e.g., hyperbolic tangent), and polynomials, and Gaussian functions [9]. In our numerical experiments, we use the exponential radial basis function kernel

$$\mathcal{K}_{\text{ERB}}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-|\mathbf{x}_i - \mathbf{x}_j| / \ell), \tag{10}$$

where ℓ denotes the kernel's width or the radius of influence of the samples selected to be support vectors. Once a functional form for the Mercer kernel $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$ has been selected, the dual optimization problem

$$\gamma^\star = \text{argmax}_{\gamma}[\sum_{i=1}^{N} \gamma_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \gamma_i \gamma_j J_j J_j \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)] \tag{11}$$

is solved subject to constraints $0 \leq \gamma_i \leq C$ and $\sum_{i=1}^{N} \gamma_i J_i = 0$.

In analogy to (9), the indicator function is written as

$$J(\mathbf{x}) = \text{sgn}\, g(\mathbf{x}), \quad g(\mathbf{x}) = \sum_{i=1}^{N} \gamma_i^\star J_i K(\mathbf{x}_i, \mathbf{x}) + b^\star. \tag{12}$$

Combining (7) and (8), both written for their counterparts in $\mathbb{R}^m$, yields a computable expression for the constant $b^\star$,

$$b^\star = -\frac{1}{2} \sum_{i=1}^{N} \gamma_i^\star J_i [\mathcal{K}(\mathbf{x}_i, \mathbf{x}_+) + \mathcal{K}(\mathbf{x}_i, \mathbf{x}_-)]. \tag{13}$$

C. Uncertainty Quantification for SVM Predictions

To quantify uncertainty in SVM reconstruction, we adapt the probabilistic SVM [7] originally developed in the context of the classification of nonperfectly separable data. The original SVM classifier J in (12) is binary, defined by the sign of the function $g(\mathbf{x})$. Instead, we use a value of $g(\mathbf{x})$ to estimate probability $\mathbb{P}[\mathbf{x} \in M_1]$ of the point $\mathbf{x}$ belonging to the material M_1. Let $g_i = g(\mathbf{x}_i)$ with $i = 1, \cdots, N$ constitute a training set. These numbers are thought of as realizations of the corresponding random variables $G_1, \cdots, G_N$, which are characterized by the (unknown) class-conditioned probability $\mathbb{P}[G_i \le g^\star | J(\mathbf{x}_i) = 1]$ where $g^\star$ is a value of $g(\mathbf{x}_i)$ at the point of interest $\mathbf{x}_i$. Instead of estimating this probability, we estimate the conditional probability $\mathbb{P}(J(\mathbf{x}_i) = 1 | G_i = g^\star)$ and extend this probability to any point $\mathbf{x}$ where measurements are not available.

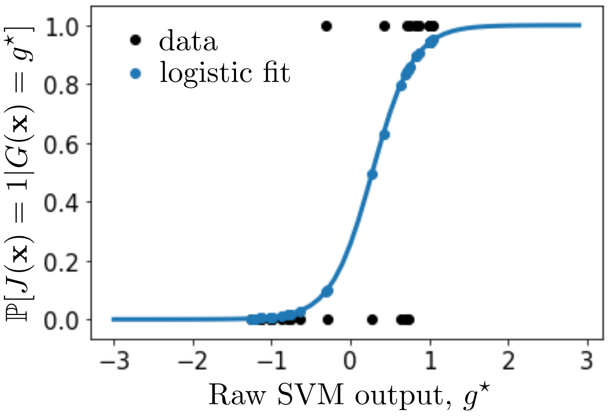

Fig. 2: The logistic fit of raw SVM output represents a probability $\mathbb{P}[\mathbf{x} \in M_1]$ conditioned on the SVM raw output being g^*, i.e., $\mathbb{P}[J(\mathbf{x}) = 1 | G(\mathbf{x}) = g^\star]$.

Our parametric estimation strategy relies on the assumed functional form of $\mathbb{P}(J(\mathbf{x}) = 1 | G(\mathbf{x}) = g^\star)$. By way of example, we consider a sigmoidal function in Fig. 2,

$$\mathbb{P}[J(\mathbf{x}) = 1 | G(\mathbf{x}) = g^\star] = \frac{1}{1 + \exp(Ag^\star + B)}. \quad (14)$$

The fitting parameters A and B are found by minimizing the negative log-likelihood of the training data. First, we map the training set $\{\mathbf{x}_i, J_i\}_{i=1}^N$ onto a training set $\{g_i, t_i\}_{i=1}^N$, where $t_i = (J_i + 1)/2$ are target probabilities. Then, the negative log-likelihood function or a "cross-entropy error function" is minimized to find optimal values $A^\star$ and $B^\star$,

$$\{A^\star, B^\star\} = \text{argmin}_{A,B} \Big[-\sum_i^N t_i \ln p_i + (1 - t_i) \ln(1 - p_i) \Big], \quad (15)$$

where $p_i = \mathbb{P}(J(\mathbf{x}_i) = 1 | G(\mathbf{x}_i) = g^\star)$. We use the trust-region Newton algorithm [10], [11] to solve this two-parameter minimization problem. Finally, (14) with $A = A^\star$ and $B = B^\star$ is used in place of $\mathbb{P}[\mathbf{x} \in M_1]$.

IV. Simulation Results

A. Synthetic Dataset

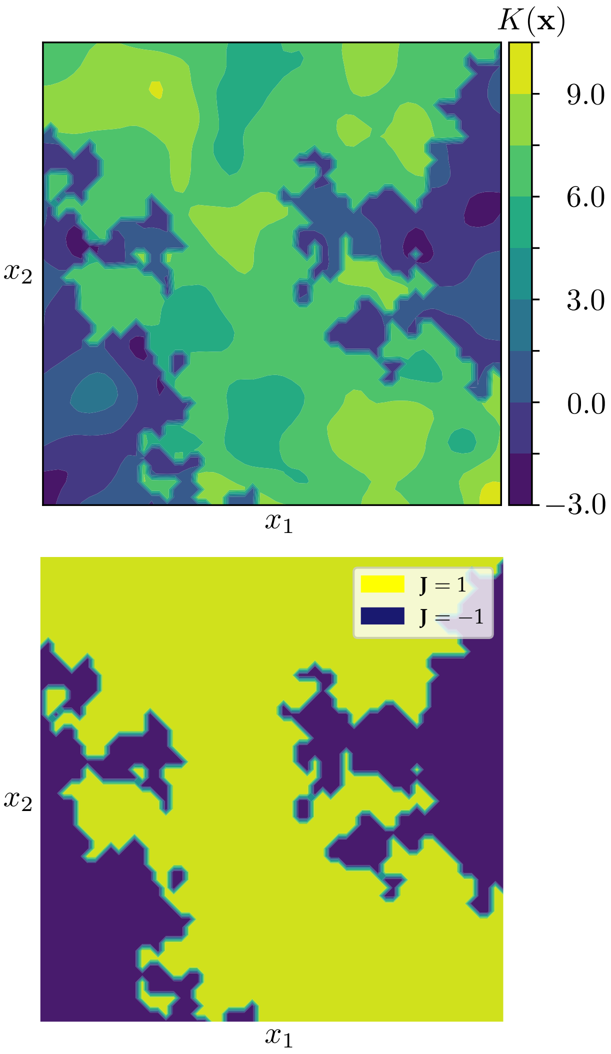

Fig. 3: Hydraulic conductivity field $K(\mathbf{x})$ (top) and the corresponding indicator function $J(\mathbf{x})$ (bottom) used in our numerical experiments. The two heterogeneous facies are sufficiently distinct for the mapping $K(\mathbf{x}) \to J(\mathbf{x})$ not to introduce a classification error.

Our goal is to reconstruct probabilistically the lithofacies defined by, e.g., the hydraulic conductivity field $K(\mathbf{x})$ in Figure 3 from N measurements $K_i = K(\mathbf{x}_i)$ collected at randomly selected locations $\mathbf{x}_i$ $(i = 1, \cdots, N)$. The field $K(\mathbf{x})$, originally used for similar purpose in [5], is constructed by superimposing two autocorrelated, weakly stationary, normally distributed random fields, representing two distinct spatial distributions of log-conductivity with the ensemble means of 0.1 and 7.0. When hydraulic conductivities are expressed in centimeters per day, this corresponds to clayey and sandy materials, respectively. The two log-conductivity distributions are mutually uncorrelated, have unit variance and Gaussian autocorrelation with unit correlation scale. SGSIM software [12] is used to generate both fields on a 60×60 grid, using a grid spacing of 1/5 of the log-conductivity correlation length. Next, the composite porous medium is constructed by randomly choosing the shape of the internal boundary. The corresponding indicator field $J(\mathbf{x})$ is constructed by assigning to each pixel either $+1$ or -1, i.e., identifying its membership in either facies M_1 or M_2, using a threshold value of 4.0. Given the vast difference between the means, this assignment is free of classification error.

B. Probabilistic Facies Reconstruction

The boundaries between materials M_1 and M_2 reconstructed by the standard (deterministic) SVM from $N = 50$ data points (out of the total of $N_{\text{tot}} = 3600$ pixels) are shown in Fig. 4. Even with this relatively sparse sampling, SVM mislabels some the data points in order to prevent the overfitting of the model (blue dots are corresponding to geological facie $J = -1$ that end up in the area in the middle area). Such mislabeling of pixels of a full image gave impetus to the original pSVM [7]. Here we use as a probabilistic classifier of incomplete images with the majority of pixels missing.

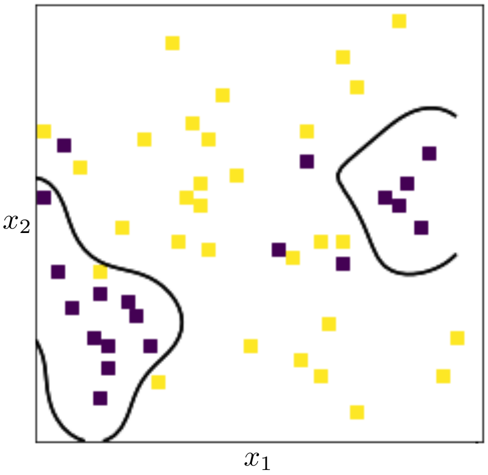

Fig. 4: Boundaries drawn by SVM with the slack penalty constant $C = 1$ and the kernel width $\ell = 2$. Blue and orange pixels represent samples from materials M_1 ($J = 1$) and M_2 ($J = -1$), respectively.

Figure 5 exhibits a representative probability map of facies M_1 reconstructed by pSVM from $N = 360$ measurements. It indicates the confidence in identifying each pixel as a member of M_1. The dark blue areas represent subdomains where the probability $\mathbb{P}[x \in M_1]$ is close to zero, i.e, these areas are highly likely to consist of material M_2. On the other hand, the yellow and light green areas represent subdomains that likely belong to material M_1. SVM classification of the remaining parts of the simulation domain is highly uncertain. Although not shown here, and as expected, this transition zone increases as the sampling density, N/N_{tot}, decreases. We also found that the measurement locations play smaller role on the confidence maps as the sampling density increases.

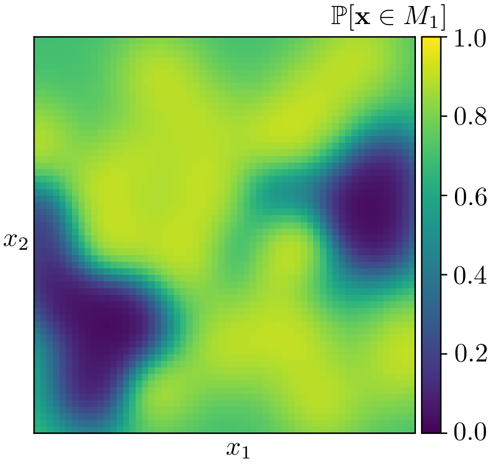

Fig. 5: Probability map of facies M_1 reconstructed by pSVM with the slack penalty constant $C = 1$ and the kernel width $\ell = 0.5$ from $N = 50$ pixels (out of the total of $N_{\text{tot}} = 3600$).

One metric of pSVM output is the fractional number of uncertain pixels, $N_{\text{unc}}/N_{\text{tot}}$, defined for a given probability threshold $P > 0.5$. A pixel $\mathbf{x}$ is deemed uncertain with confidence P if its membership probability $\mathbb{P}[\mathbf{x} \in M_1]$ falls within the interval $[1 - P, P]$. Figure 6 shows $N_{\text{unc}}/N_{\text{tot}}$ as function of the sampling density N/N_{tot} for two degrees of certainty, $P = 0.75$ and 0.95; this result represents an average over 20 realizations of the set $T = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ of randomly selected measurement locations, each of which yields a probability map similar to the one in Fig 5. As expected, the number of uncertain pixels increases with the probability threshold P and decreases with the sampling density.

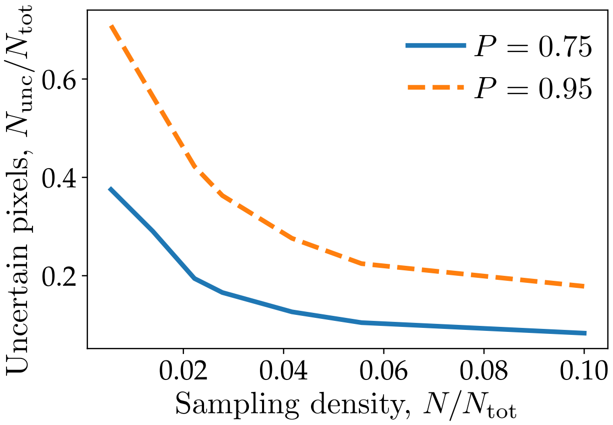

Fig. 6: Relative number of uncertain pixels, $N_{\text{unc}}/N_{\text{tot}}$, as function of the sampling density, N/N_{tot}, for two degrees of certainty, $P = 0.75$ and 0.95.

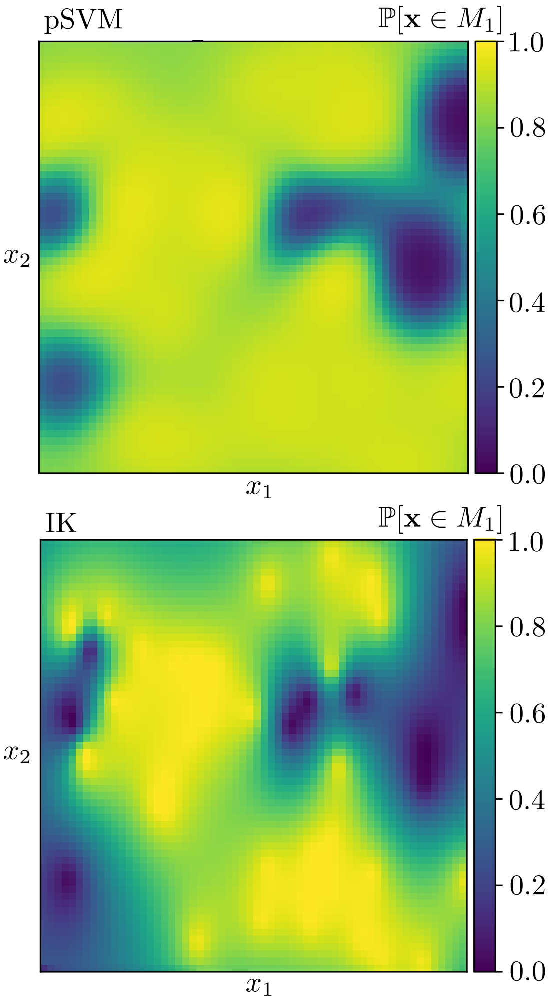

Fig. 7: Probability maps of facies M_1 generated by pSVM (top) and indicator Kriging (bottom).

C. Comparison with Indicator Kriging

Indicator Kriging (IK) [2] provides an alternative means for probabilistic reconstruction of lithofacies [1]. This method defines the indicator function as $I(\mathbf{x}_i) = 1$ for $\mathbf{x}_i \in M_1$ and $= 0$ for $\mathbf{x}_i \in M_2$, and treats it as a stationary random field. The best linear unbiased estimator (aka Kriging) interpolates between the measurement points, yielding $\mathbb{E}[I(\mathbf{x})] = \mathbb{P}[\mathbf{x} \in M_1]$. The correlation function (variogram) of $I(\mathbf{x})$, which is inferred from the spatial data $I_i = I(\mathbf{x}_i)$ with $i = 1, \cdots, N$, determines the interpolation weights.

Figure 7 provides probability maps of M_1 alternatively identified with pSVM and IK from $N = 50$ measurements. (The measurement locations used to generate this figure differ from those used in Fig. 5; this allows us to illustrate their impact on the quality and reliability of facies reconstruction.) The probability maps generated with pSVM and IK are qualitatively similar, even though pSVM generates a smoother map than that produced by IK. This suggests that pSVM provides a more conservative facies reconstruction (larger areas with probabilities other than 0 and 1).

The qualitative similitude between pSVM and IK argues in favor of the former. That is because IK is more complex and possesses more tunable parameters, such as lag, lag separation, lag tolerance, azimuth, dip, tolerance, and bandwidth. Manual fitting of data to an experimental variogram is highly subjective, requiring one to visually identify an appropriate nugget effect and sill. Finally, construction of a variogram requires a large number of samples collected at various degrees of spatial separation, while SVM theoretically works with as few as two data points (support vectors).

Figure 8 serves to quantify the discrepancy between the two probability maps in Fig. 7, and to compare them with their empirical counterpart for the ground truth in Fig. 3. A metric for this assessment is computed as follows. First, we construct a histogram of pixels in each of the probability maps in Fig. 7 whose probabilities fall within bins of size $\Delta p = 0.1$. Second, we compute the average probability for each bin: for example, let $\{\mathbf{x}_k\}_{k \in \mathcal{I}_{0.1,0.2}}$ denote a set of $N_{0.1,0.2}$ pixels with probabilities $\{p_k\}_{k \in \mathcal{I}_{0.1,0.2}}$ that fall within the bin $[0.1, 0.2)$; then the average probability for that bin is $\bar{p}_{[0.1,0.2)} = (1/N_{0.1,0.2}) \sum_{k \in \mathcal{I}_{0.1,0.2}} p_k$. The values of $\bar{p}$ for each bin, inferred from the probability maps in Fig. 7, are plotted in Fig. 8 against the corresponding probabilities p^{true}. These are computed as the fraction of the pixels in a given bin labeled as $J = 1$ in Fig. 3. For example, if the number of the $J = 1$ pixels in the set $\{\mathbf{x}_k\}_{k\in\mathcal{I}_{0.1,0.2}}$ is $N^{J=1}_{0.1,0.2}$, then $p^{\text{true}}_{[0.1,0.2)} = N^{J=1}_{0.1,0.2}/N_{0.1,0.2}$. The 45-degree line in Fig. 8 corresponds to the perfect agreement between the average probability $\bar{p}$ and the corresponding empirical probability p^{true} inferred from the ground truth. The average probabilities $\bar{p}$ predicted by either pSVM or Kriging exhibit comparable deviations from the 45 degree line; this suggests that, according to this metric, the two methods for probabilistic reconstruction of geologic facies have comparable accuracy.

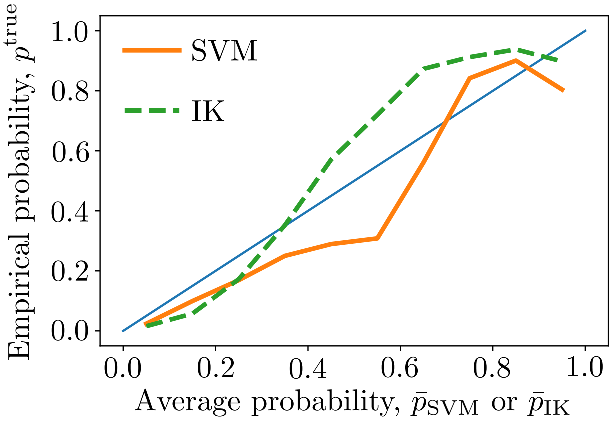

Fig. 8: Comparison of the verage probability $\bar{p}$ predicted by either pSVM ($\bar{p}_{\text{SVM}}$) or indicator Kriging ($\bar{p}_{\text{IK}}$) and the corresponding empirical probability p^{true} inferred from the ground truth. The 45-degree line corresponds to the perfect agreement between the two.

D. *Sensitivity to SVM Parameters*

Performance of SVM is controlled by two parameters: the slack penalty constant C in (5) and the kernel width ℓ in (10). The regularization parameter C provides a trade-off between the correct classification of training data and the minimization of generalization error. Larger values of C allow smaller margins if the decision function is better at correctly classifying all training points. Smaller values of C promote larger margins and, hence, a simpler decision function at the cost of training accuracy.

Small values of the parameter ℓ indicate a long-range influence of each observation, while its high values limit the overall impact of each data point. If ℓ is too small, the radius of influence of the support vectors includes only the support vector itself and no amount of regularization with C would prevent overfitting. When ℓ is very large, the model is too constrained and cannot capture the complexity or "shape" of the data. The region of influence of any selected support vector would include the whole training set. The resulting model will behave similarly to a linear model with a set of hyperplanes separating the centers of two classes.

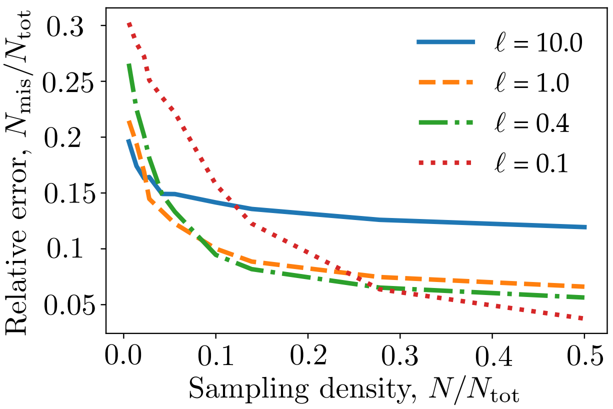

Fig. 9: Relative number of misclassified pixels, $N_{\text{mis}}/N_{\text{tot}}$, as function of sampling density N/N_{tot}, for several values of the kernel width ℓ.

Figure 9 shows the impact of ℓ on the SVM accuracy, i.e., on the relative number of misclassified pixels, $N_{\text{mis}}/N_{\text{tot}}$. Small values of ℓ are beneficial when sampling density is high, while its large values yield a better performance when data are very sparse. In the latter case, small values of ℓ lead to overfit and result in low prediction accuracy.

Figure 10 exhibits representative reconstructions of geological facies obtained with two choices of the parameter ℓ (and $C = 1.0$) for two sampling densities. Large values of ℓ yield boundaries that are too smooth, while small values of ℓ cause boundaries to follow the training points too closely. Selecting a right value for ℓ is more crucial for low sample density ($N/N_{\text{tot}} = 50/3600$). As expected, this situation also gives rise to appreciable variation between realizations (different sample locations). Both the importance of selecting a value for ℓ and the between-realizations variability diminish when the sampling density increases to $N/N_{\text{tot}} = 360/3600$.

E. *Strategy for Parameter Selection*

Unlike IK, SVM possesses a well-established framework for tuning its hyperparameter. Specifically, optimal values of C and ℓ are chosen with the following algorithm.

- Identify a region in the two-dimensional SVM parameter space (spanned by parameters C and ℓ), over which C and ℓ are allowed to vary.

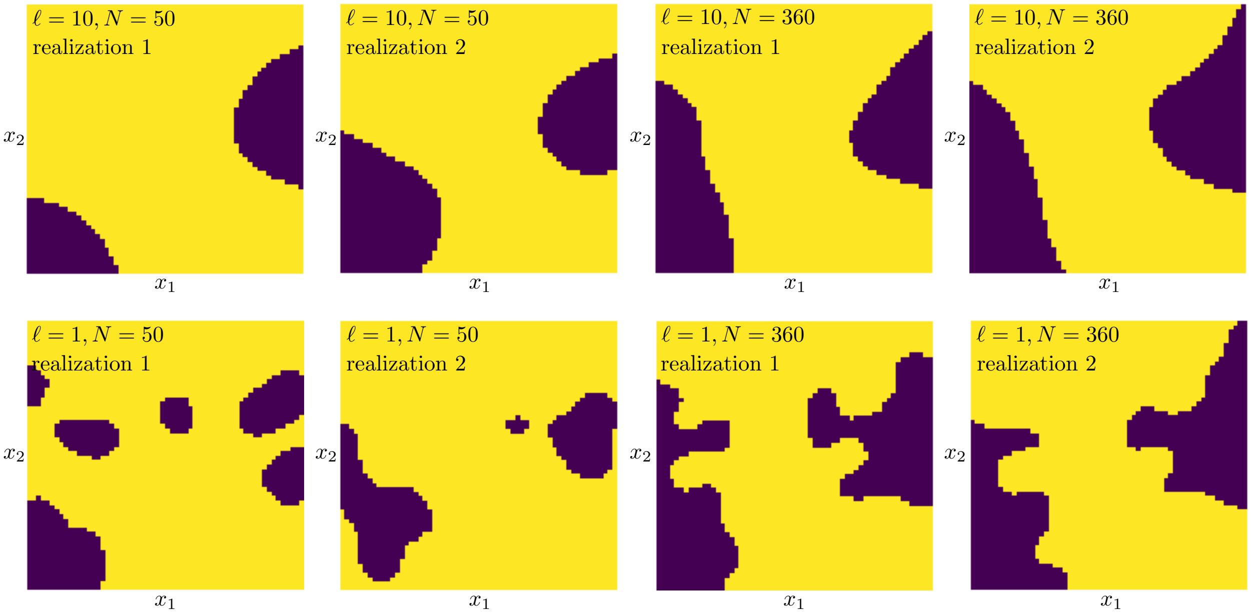

Fig. 10: Representative lithofacies reconstructions via SVM with $C = 1$ and either $\ell = 10.0$ (top row) or $\ell = 1.0$ (bottom row), for two sets of N samples (realizations 1 and 2).

In our computational examples, we chose the intervals $C \in [10^{-2}, 10^3]$ and $\ell \in [10^{-2}, 10^3]$ to exclude parameters for which poor performance is expected a priori.

- Discretize this region with a regular grid and carry out the SVM reconstruction for all pairs of the parameter values. We used the mesh size $\Delta(\lg C) = 1$ and $\Delta(\lg \ell) = 1$, which results in 36 reconstructions.
- Perform five-fold cross-validation to evaluate the test error of each parameter combination. The data are split into $k = 5$ subsets or "folds" $\mathcal{F}_i$ with $i = 1, \cdots, k$. For every i, the model is trained on all folds except for the ith fold. The test error on the ith fold is computed as

$$\mathcal{E}_i = \frac{1}{k} \sum_{n \in \mathcal{F}_i} \mathbb{1}(J_n \neq \hat{J}_{n/i}),$$

where J_n is a true label of pixel $\mathbf{x}_n$, and $\hat{J}_{n/i}$ is a prediction for the pixel $\mathbf{x}_n$ obtained without using the fold $\mathcal{F}_i$ of the dataset to fit the model. Next, cross-validation error is obtained by averaging the test errors of individual folds,

$$\mathcal{E}_{\text{cv}} = \frac{1}{k} \sum_{i=1}^{k} \mathcal{E}_i.$$

- Finally, the classification accuracy is defined as $\mathcal{A} = 1 - \mathcal{E}_{\text{cv}}$.

Figure 11 shows the classification accuracy $\mathcal{A}$ for the 36 combinations of the SVM parameters C and ℓ. The best-performing SVM models are parameterized with the values of C and ℓ that lie on the diagonal of the plot. Possible spurious variations of the classification accuracy $\mathcal{A}$ between different chosen parameters can be smoothed out by increasing the number of folds, k, used in cross-validation at the expense of computational time.

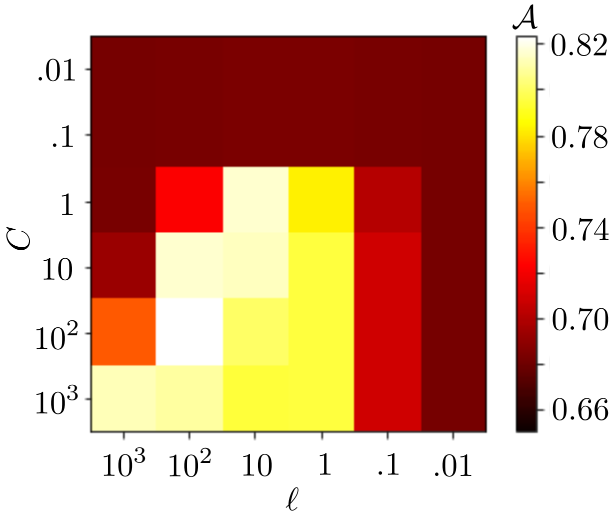

Fig. 11: Classification accuracy $\mathcal{A}$ computed from the five-fold cross-validation. Light color elements correspond to the combinations of the SVM parameters C and ℓ that lead to well-performing models.

V. Conclusions

We introduced probabilistic support vector machines (pSVM) as a means of delineation of subsurface hydrofacies from sparse data. The method replaces the binary classifier with its continuous counterpart that is constructed by fitting a logistic curve to observations. The result is a probability map that provides the likelihood of a pixel belonging to a facies, rather than a deterministic pixel label provided by standard SVM. Our numerical experiments lead to the following major conclusions.

1) Probability maps generated with pSVM and indicator Kriging (IK), the current method of choice for probabilistic forecasting, are qualitatively similar.
2) pSVM generates smoother probability maps than those produced by IK, suggesting that pSVM provides a more conservative facies reconstruction (larger areas with probabilities other than 0 and 1).
3) The qualitative similitude between pSVM and IK argues in favor of the former, because IK is more complex, has more tunable parameters, and has higher data requirements.
4) Performance of SVM is controlled by two parameters: the slack penalty constant C and the kernel width ℓ.
5) Small values of ℓ are beneficial when sampling density is high, while its large values yield a better performance for sparse data.

More work remains to be done in the area of probabilistic image reconstruction from sparse data. In future studies, we will alternatives to pSVM for uncertainty quantification, such as conformal prediction [13]. The latter uses past experience to determine precise levels of confidence in new predictions. It is designed for an on-line setting in which labels are predicted successively, each one being revealed before the next is predicted [14]. Such a strategy might indicate regions in space where a prediction with a required degree of certainty is not possible due to the lack of information.

References

[1] L. Guadagnini, A. Guadagnini, and D. M. Tartakovsky, "Probabilistic reconstruction of geologic facies," *J. Hydrol.*, vol. 294, pp. 57–67, 2004.

[2] E. H. Isaaks and R. M. Srivastava, *An Introduction to Applied Geostatistics*. New York: Oxford Univ. Press, 1990.

[3] V. N. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.

[4] D. M. Tartakovsky and B. E. Wohlberg, "Delineation of geologic facies with statistical learning theory," *Geophys. Res. Lett.*, vol. 31, no. 18, p. L18502, 2004.

[5] B. Wohlberg, D. M. Tartakovsky, and A. Guadagnini, "Subsurface characterization with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 1, pp. 47–57, 2006.

[6] Y. Zeng, K. Jiang, and J. Chen, "Automatic Seismic Salt Interpretation with Deep Convolutional Neural Networks," *arXiv e-prints*, p. arXiv:1812.01101, Nov 2018.

[7] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 2000, pp. 61–74.

[8] B. E. Wohlberg and D. M. Tartakovsky, "Delineation of geological facies from poorly differentiated data," *Adv. Water Resour.*, vol. 32, no. 2, p. 225230, 2009.

[9] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. Springer, 2014.

[10] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.

[11] C.-J. Lin, R. C. Weng, and S. S. Keerthi, "Trust region newton methods for large-scale logistic regression," in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML '07. New York, NY, USA: ACM, 2007, pp. 561–568. [Online]. Available: http://doi.acm.org/10.1145/1273496.1273567

[12] C. V. Deutsch and A. G. Journel, *GSLIB Geostatistical Software Library and User's Guide*, 2nd ed. New York: Oxford Univ. Press, 1998.

[13] Y. Hechtlinger, B. Pczos, and L. Wasserman, "Cautious deep learning," 2018.

[14] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *J. Mach. Learn. Res.*, vol. 9, pp. 371–421, Jun. 2008. [Online]. Available: http://dl.acm.org/citation.cfm?id=1390681.1390693