Reservoir Modelling Under Uncertainty - A Kernel Learning Approach

Vasily Demyanov¹, Mike Christie², Mikhail Kanevski³, Alexey Pozdnoukhov⁴

**Abstract.** We propose a novel approach to tackle the problem of matching the data from different sources and keeping the model realism to produce reliable reservoir prediction. This high level overview presents a way to bring together prior geological knowledge and relevant data under uncertainty by means of intelligent data fusion through kernel learning. The kernel learning approach balances model complexity with goodness of fit by selecting and reproducing relevant model features. Impact from data at different sources is combined by means of kernels functions, which represent spatial correlation. Kernel regression in the hyper space of features ensures representation of non-linearity and overcomes heavy stationarity assumption. Use of the kernel transformation allows us to reproduce continuity along data manifolds and represent correlation at multiple scales. Support vector formalism ensures the data influence the pattern according to their relevance to the modelled phenomenon (spatial dependency) and the target problem (matching the data). Furthermore, the impact of noisy and atypical data is handled in a rigorous way to prevent loss of predictive capability of the model.

**Introduction**

The success in modelling diversity and natural variability of geological media depends on the capability of a particular algorithm to reproduce realism of complex multi-scale reservoir characteristics and agree with data at the same time. Adequate representation of uncertainty in spatial properties is determined by the choice of appropriate model parameterisation and its ability to match the data. The problem of agreement between geostatistical simulations and preservation of the a priori spatial features is subject to uncertainty in conditioning data and assumptions about the model properties (e.g. connectivity, continuity, correlation, stationarity, etc.).

---

1 Institute of Petroleum Engineering, Heriot-Watt University, Riccarton , EH144AS, UK vasily.demyanov@pet.hw.ac.uk, http://www.pet.hw.ac.uk/research/uncertainty-quantification.htm
2 Institute of Petroleum Engineering, Heriot-Watt University, Riccarton , EH144AS, UK mike.christie@pet.hw.ac.uk, http://www.pet.hw.ac.uk/research/uncertainty-quantification.htm
3 Institute of Geomatics and Analysis of Risk, University of Lausanne, www.unil.ch/igar
4 National Centre for Geocomputation, National University of Ireland Maynooth, Ireland, http://ncg.nuim.ie

A variety of geostatistical algorithms used in reservoir modeling are based on different concepts of representing spatial correlation and continuity. Object-based algorithms traditionally provide very interpretive models using a variety of geologically tailored shapes placed according to the Boolean rules and conditioning [Deutsch 2002]. Although, object-based models can retain a high level or realism, they may be difficult to condition to all the hard and soft data, which involves sophisticated non-unique optimization. Uncertainty quantification with object-based models may be also difficult because of their increasing complexity with a large number and variability of the object shapes. Traditional variogram-based geostatistical stochastic simulation algorithms are widely used and are straightforward to condition to available data. However, they have known limitation in reproduction of connectivity features, e.g. in fluvial environment. More recently proposed multi-point statistics simulations are based on the higher-order statistical moments (training image) and have proved their ability to represent variability of complex connected patterns and honour the data following the geostatistical sequential simulation paradigm [Caers, 2005]. One of the difficulties in the above approaches is in robust control of connectivity and its relation with the model control parameters, such as spatial correlation and distribution, which can be complex and non-unique.

A different approach for modeling spatial distribution of reservoir properties with kernel transformation was used in [Demyanov et.al. 2008]. Spatial correlation model in this approach is described with a kernel regression in the metric space, where the metric controls the complexity of the model. The model complexity is balanced by the goodness of fit according to the statistical learning theory described in [Vapnik, 1995]. Kernel functions are represented by support vectors assigned based on learning from data. Support vector formalism has been applied for reservoir property modeling for example in [Kanevski, et.al. 2001; Al Anazi & Gates 2010].

Additional kernel distortion is introduced on the estimation by the manifold of unlabelled data. Unlabelled data form a manifold described by a set of relevant locations without available measurement values of the modelled property. For instance, a shape of a possible sand body relates the observation of sand porosity to the given sand facie geometry, and in this case the relevant manifold provides additional conditioning to the estimation model. Learning algorithms that take into account unlabelled data are called semi-supervised – they use not only the data with known answers but also data without answers for predictions. An application of semi-supervised support vector regression (SVR) was applied to map air pollution in [Sindhwani V et. al. 2005]. The idea of the manifold is to enforce continuity along the manifold as additional restriction to the estimation models showed significant improvement in predictions over conventional supervised-learning algorithms [Pozdnoukhov & Bengio, 2006].

A generalization of support vector formalism has been developed in multiple kernel learning theory by [Scholkopf & Smola, 2002]. Multiple kernel learning (MKL) provides a way to integrate information more accurately from multiple
features, which represent different signal patterns. Application of MKL to reservoir property modelling was demonstrated in [Demyanov et.al., 2010], where spatial features at different scale were fused together in a self-consistent way balancing the goodness of fit and the model complexity. Further development of MKL reservoir modeling is presented in [Backhouse et.al. 2011].

Kernel PCA technique in the metric space was proposed to evaluate uncertainty of reservoir models in [Caers et.al 2010], where it was used to discover the non-parametric relationship between reservoir realisations in metric space. Distance based approach with kernel transformation in the metric space was applied in the multi-point statistics simulations [Honarkhah & Caers 2010]. Also, kernel learning algorithms were used to model relationship between the geomorphological parameters to derive prior probability density functions for reservoir models in [Rojas et.al. 2012].

In the present work we develop further the application of semi-supervised SVR reservoir property model and demonstrate its performance for representing uncertainty of reservoir predictions using different manifolds based on raw seismic data and multi-point statistics realization.

Kernel Learning Methods

Kernel learning methods unite a large family of data driven algorithms that use kernel functional formalism. Among them are neural networks based on radial basis functions and Naradaya-Watson kernel estimator [Haykin, 2008]. More advance kernel learning algorithms namely support vector machines were proposed by Vapnik and Chervoninkins to solve classification problems. Their statistical learning theory [Vapnik & Chervonenkins 1974] explains the nature of many data driven algorithms and lays a solid mathematical foundation for the training process. It proposes a rigorous way to balance the goodness of model fit with the model complexity, described as VC-dimension, to make sure the model does not become over complex and fail to generalise to make good predictions.

Support vector regression

Support vector regression (SVR) was designed as an extension of support vector machine to solve a regression problem [Vapnik et.al. 1997]. The idea behind SVR is to perform a linear robust regression in a high-dimensional metric space that provides the right balance between the goodness of fit (empirical risk) and the model complexity. In an original space the result corresponds to nonlinear modeling. Robustness is achieved by applying an ε-insensitive cost function, i.e. data inside the ε-tube are not penalized (see Fig. 1).
The SVR estimate is computed as a kernel regression (eq. 1) based on the defined support vectors (see Fig. 1):

\[ f(x) = \sum_{i}^{L} (\alpha_i^* - \alpha_i) K(x, x_i) + b \]  

(1)

where \( K(x, x_i) \) is a kernel function (Gaussian is used in this study) imposed on each of the data location \( x_i \); \( \alpha_i^* \) and \( \alpha_i \) are the coefficients that are assigned to data: \( \alpha > 0 \) for support vectors \( \alpha = C \) for atypical/noisy data and \( \alpha = 0 \) to the rest of the data.

SVR is controlled by several tuning hyper-parameters including the kernel size, the regularisation term \( C \), which imposes the bound on the support vectors, and the \( \varepsilon \)-tube (see Fig. 1). These parameters can be tuned using cross-validation [Kanevski et al 2009] or Bayesian inference [Demyanov et al, 2008].

Figure 1. Support vector regression performs a linear regression in the kernel space into which the problem is transformed by a kernel transformation \( K \) using assigned SVs with \( \alpha > 0 \) in Eq. (1) from [Kanevski, et. al. 2009]

**Semi-supervised support vector regression**

One of the burdens of data driven methods is its demand to large amounts of data. In reservoir model there are usually not many hard conditioning data, which usually come from processing and interpreting information from the well logs and core. In this case extra relevant information becomes extremely valuable and can improve the quality of the models. Therefore the idea of unlabelled data – the relevant location with no corresponding function measurement available – becomes appealing.

Unlabelled data can be used in semi-supervised learning algorithms, which unlike conventional supervised learning algorithms (e.g. multi-layer perceptron) do not require all the input data to have the corresponding output answers for training (as for the labelled data) [Chapelle, 2006]. Unlabelled data are introduced in semi-supervised algorithms through kernel modifications imposed by the manifold of unlabelled data. Unlabelled data bring the information on the data
manifold to incorporate natural similarity between likely related locations in a similar way to other types of spatial correlation/continuity models (variogram, training image, Boolean objects). Predictive manifold learning with kernels incorporates the geometry induced by the low-dimensional manifold (see Fig. 2a). This is generally done by first constructing a graph model of the manifold by connecting the adjacent (both labelled and unlabelled) data samples. Then, the model smoothness is enforced along the manifold by employing the properties of the kernels as the dot products in some Hilbert space (Pozdnoukhov and Bengio, 2006). The original kernel function $K(x,x')$ is modified implementing the smoothness of the kernel with respect to the geometrical structure of the manifold. It is achieved by using the notion of the graph Laplacian. Though semi-supervised SVR provides good results, it includes a costly matrix inversion and introduces additional parameters to tune [Demyanov 2008]. The influence of the manifold on the predictions around the labeled points can be seen in Fig. 2b (zoomed in from the left), where unlabelled data enforces the continuity of the kernel imposed over the labeled data points (support vector). The estimated spatial field shows diversity and increase to the peak values beyond the neighbouring observation due to the complexity of the kernel regression controlled by the dimension of the Hilbert metric space. Similar pattern can be possibly obtained with a Gaussian process with proper conditioning, although the spatial small scale variability of a Gaussian process is related to a non-physical seed, whereas in SVR it is controlled by the manifold, which can be described by interpretative parameters. The impact of the manifold is controlled by the dedicated parameter, which balances the distortion of the SV kernels, and also depends on the density of the unlabelled points in it.

Figure 2. Semi-supervised kernel learning: a) manifold of unlabelled data (+) defines the region of relevance for the kernel model; b) unlabelled data (°) influence the estimate around the labelled data point (●) by distorting the kernel and imposing continuity along the manifold.
One of the important questions is the source of the manifold will be addressed in this work.

**Semi-supervised Support Vector Regression Reservoir Model with Different Manifolds**

A semi-supervised SVR model was designed to populated petrophysical properties (porosity and permeability) in a 2D reservoir model. A layer with sinuous channels from Stanford VI synthetic reservoir model [Castro, et al, 2005] was selected for the case study. The grid consists of 30 thousand cells and includes four facie types: channel sand, point bar sand, channel boundary mud and the shale (see Fig. 3a). Conventional geostatistical modelling implies distributing the facies and then populating them separately with the petrophysical properties. Semi-supervised SVR models were designed separately for the sands and the shales combining pairs of facies with similar ranges of porosity and permeability. The model is conditioned to just 30 hard data coming from the wells and to the soft data provided by synthetic seismic (see Fig. 3b). Also, a noisy version of the synthetic seismic was produced by adding normally distributed noise to the smoothed seismic in order to blur the clearly defined channel boundaries (see Fig. 3c). The reservoir produces from multiple wells in the middle (not all the wells penetrate the channels in the selected layer) with water injected through the wells on the top and left edges.

Several options were exercised to design the manifold for the semi-supervised SVR model. A manifold can be derived from the synthetic seismic data by applying a cut-off to separate between the sand and the shale, as it was done in [Demyanov et al, 2008]. In case of clean synthetic seismic it clearly separates and identifies channel (see Fig. 4a). However, in case of noisy seismic the manifold can be quite uncertain and a flexible cut-off may be needed. Also, in this case
there is no much control over the manifold apart from the cut-off value for seismic "interpretation" and the impact of the manifold on the SVR estimates.

Another way to create the manifold can be proposed to reproduce the continuity of the fluvial channels. Multi-points statistics (MPS) simulated realisations based on the correct training image very well represents connected spatial structures [Caers 2005]. Although, multi-point statistics realisations reproduce well global spatial structure based on a training image, they can exhibit significant variability in connectivity due to high local uncertainty of stochastic simulations. This uncertainty is inherent and corresponds to the random seed, which is not a meaningful geological parameter but reproduce some uncorrelated variability. A manifold derived from the multi-points statistics realisation represents the region of likely location of continuous geological bodies. Continuity of the bodies in this case is imposed by the kernel nature of how the manifold is integrated into the SVR model. Thus, semi-supervised SVR is able to ensure better continuity of the property pattern than the one based on the MPS facies model with the petrophysical properties distributed with sequential Gaussian simulations. The latter tends to separate high and low tail values due to its maximum entropy nature [Deustsch, 2002]. Using the MPS model to produce the manifold also provides more flexibility for its continuity and correlation with control parameters of Ti transformations and proportions in addition to the points’ density and the semi-supervised SVR parameters of the manifold’s impact. Figure 4 shows manifolds generated based on seismic (a) and based on the SNESIM realisation (b).

![Figure 4. Manifold for semi-supervised learning based on: seismic (a); multipoint statistics realisation (b)](image)

SNESIM MPS algorithm was designed to model spatial distribution categorical variables and is used commonly for facies simulation [Strebelle, 2002]. Similarly we used SNESIM to generate manifolds for the sand and shale facies. A training image with sinuous parallel channels with 2 facies for sand and shale was used.
SNESIM realisation conditioned on the hard data from the wells and the soft probability derived from the synthetic seismic were computed (see Fig. 5b,c). The scaling of the training image was considered uncertain to allow the variation of channel width and wavelength (Remy et al, 2009). Affinity parameter used in SNESIM algorithm was varied to produce a different channel configuration (see Fig. 5d).

The manifold was designed by generating points at random within the facie region given by the MPS realisation. The number of points in the manifold varied to provide different density across the manifold.

Separate semi-supervised SVR models were designed for porosity and permeability for sand and shale and then merged on the reservoir grid using a cookie-cut with the corresponding manifolds.

The semi-supervised SVR porosity patterns with different manifold models are presented in Figure 6. The model based on the manifolds from the synthetic “clean” seismic features very clear separation of the channel structure (see Fig. 6a). The model with MPS-based manifold shows more variability of the spatial distribution of the petrophysical
properties along the manifold with some low porosity sands located on the side of the channels and in between (see Fig. 6b). Once the “noisy” seismic is used to condition the model the channel boundaries get more blurry (see Fig. 6c).

**History Matching Semi-supervised SVR Model**

The semi-supervised SVR reservoir model was history matched to the truth case production data – oil and water production. The reservoir production was run under the pressure control with constant water injection for the pressure support. The truth case data provided for the period of 6.5 years were polluted with noise at a constant level (~10% for each well) to form the production history. The misfit function was defined as a standard least squares norm:

$$M = \sum_{i=1}^{T} \frac{(q_{\text{obs}}^i - q_{\text{sim}}^i)^2}{2\sigma^2}$$

where $T$ is the number of observations, $q$ is the rate (oil or water), superscripts $\text{obs}$ and $\text{sim}$ refer to observed and simulated, and $\sigma^2$ is the variance of the observed data equal to the variance of the Gaussian noise added to the truth case production.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prior range</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Kernel width for channel sand porosity</td>
<td>5 : 20</td>
<td></td>
</tr>
<tr>
<td>Kernel shale porosity</td>
<td>10 : 40</td>
<td></td>
</tr>
<tr>
<td>Kernel width for channel sand permeability</td>
<td>5 : 20</td>
<td></td>
</tr>
<tr>
<td>Kernel shale permeability</td>
<td>10 : 40</td>
<td></td>
</tr>
<tr>
<td>SVR regularisation constant for sand porosity (log C)</td>
<td>2 : 4</td>
<td></td>
</tr>
<tr>
<td>SVR regularisation constant for shale porosity (log C)</td>
<td>0 : 3</td>
<td></td>
</tr>
<tr>
<td>SVR regularisation constant for sand permeability (log C)</td>
<td>3 : 8</td>
<td></td>
</tr>
<tr>
<td>SVR regularisation constant for shale permeability (log C)</td>
<td>2 : 4</td>
<td></td>
</tr>
<tr>
<td>Number of points in the sand manifold</td>
<td>200 : 1200</td>
<td></td>
</tr>
<tr>
<td>Number of points in the sand manifold</td>
<td>200 : 1200</td>
<td></td>
</tr>
<tr>
<td>Scaling for porosity data</td>
<td>5 : 100</td>
<td></td>
</tr>
<tr>
<td>Scaling for permeability data</td>
<td>10 : 1000</td>
<td></td>
</tr>
<tr>
<td>TI horizontal (x) affinity parameter</td>
<td>0.1 : 5</td>
<td></td>
</tr>
<tr>
<td>TI vertical (y) affinity parameter</td>
<td>0.1 : 5</td>
<td></td>
</tr>
</tbody>
</table>

History matching was performed using stochastic sampling, in particular – Particle Swarm Optimisation algorithm used in [Mohammed et al, 2010]. Uncertain model parameters are listed in Table 1. Uninformative flat priors were defined with intervals based on the sensitivity runs (see Table 1).
The optimisation resulted in generating multiple history matched models that form ensembles to represent the range of uncertainty of the mode predictions. This uncertainty can be quantified by inferring the models from the generated ensembles in a similar way presented in [Christie et al 2006].

Production predictions for 5 best models for each of the four presented cases are plotted in Figure 7. It can be noticed that semi-supervised SVR model based on the “clean” synthetic seismic produces almost identical history matches because it is well conditioned by the data with no noise, which strongly restricts the manifold shape (see Fig. 7d). Use of MPS realisation to design the manifold results in different production predictions (see Fig. 7a), which provide slightly worse fit in case of no TI transformation (affinity parameter is fixed and the TI channel geometry proportions do not match with the ones of the truth case, see Fig. 5a,d). Once the affinity parameter is tuned in history matching, the model fits
better and the spread of the predictions is wider due to the variability between the realisations (see Fig. 7b). The worst fits and the widest spread of the predictions came from the model conditioned to the noisy seismic (see Fig. 7c).

Conclusions

The work demonstrates an application of semi-supervised SVR for modeling petrophysical reservoir properties. Various aspects of the model uncertainty were analyzed, in particular – the uncertainty in spatial continuity and connectivity described by the manifold of unlabelled data. Manifold provides a way to represent continuous geological features in the model. Uncertainty of the manifold need to represent geological heterogeneity and is modelled with multipoint statistics.

Different sources of the manifold were considered: synthetic seismic data and MPS realization. MPS provides a good solution for the manifold because the scheme combines the reproduction of the complex multi-point spatial continuity based on the TI and SVR enforces connectivity along the manifold.

Ensemble of history matched semi-supervised SVR reservoir models with MPS based manifold was generated with stochastic sampling technique. A number of model parameters were tuned to obtain multiple solutions for the history matching problem. Among them were SVR parameters (kernel size and regularisation), TI transformation global affinity multiplier in the MPS algorithm parameters, strength of the manifold and some other.

Further work is seen in inferring the obtained ensemble of models to approximate the posterior probability and compute the confidence intervals for the predictions.

Acknowledgement

Funding of this work was provided by and by the industrial sponsors of the Heriot-Watt Uncertainty Project, phase 3. The authors would like to acknowledge Swiss National Science Foundation for funding “GeoKernels: kernel-based methods for geo- and environmental sciences (Phase II)” (No 200020-121835/1) and “Analysis and Modelling of Space.Time Patterns in Complex Regions” (200021_140658). The authors would like to thank J. Caers and S. Castro for providing Stanford VI case study and Stanford University of the use of SGeMS.
Bibliography


