IDIAP RESEARCH REPORT

Estimation of Conditional Distributions using Gaussian Mixture Models

Nicolas Gilardi [a] gilardi@idiap.ch
Samy Bengio [b] bengio@idiap.ch
Mikhail Kanevski [b] kanevski@idiap.ch

IDIAP–RR 02-03

March 2002

SUBMITTED FOR PUBLICATION

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] University of Lausanne, Mineralogy Institute, BFSH2, 1015 Lausanne, Switzerland
[b] Dalle Molle Institute for Perceptual Artificial Intelligence CP 592, rue du Simplon 4, 1920 Martigny, Switzerland

IDIAP Research Report 02-03

Estimation of Conditional Distributions using Gaussian Mixture Models

Nicolas Gilardi gilardi@idiap.ch Samy Bengio bengio@idiap.ch
Mikhail Kanevski kanevski@idiap.ch

March 2002

submitted for publication

Abstract. This paper proposes the use of Gaussian Mixture Models to estimate conditional probability density functions. A conditional Gaussian Mixture Model has been compared to the geostatistical method of Sequential Gaussian Simulations. The data set used is a part of the digital elevation model of Switzerland.

1 Introduction

Environmental survey needs very reliable tools in order to facilitate decision making. An important category of these tools is called "Risk Maps". It consists of drawing various kinds of probability maps, such as "indicator maps" (probability of exceeding a threshold), the "value at risk" (quantile map), etc.

These problems can be solved using classical regression models such as K-Nearest Neighbors, Inverse Distance, Indicator Kriging, Artificial Neural Networks, etc. However, it is known that regression models based on minimization of the expected error have a smoothing effect and do not recover the variability of data. In the case of risk mapping, this smoothing effect is not acceptable as one is especially interested in unusual events, i.e. events that are not necessarily extreme but often far from the mean value. It was thus necessary to develop alternate prediction methods which would concentrate on reconstructing not only the mean but also, at least, the variability of data.

In Geostatistics, Stochastic Simulations [3] were developed to solve these particular problems. However, these methods have some drawbacks. The modelization process is usually very complicated and necessitates a strong expert knowledge. In addition, they are often based on some assumptions about data distribution (stationarity, normality, ...), and they do not provide any analytical model of the local distribution of a sample point which can be reused for other tasks.

In this paper, we propose a method that can estimate the local probability density function (PDF) for each data point, without making any assumption of the distribution of data. It is based on the use of Gaussian Mixture Models (GMM) for conditional density estimation, by conditioning a global PDF model on the sample location.

To evaluate the relative performance of this method, we compare it to the well-known Geostatistical method of Sequential Gaussian Simulations (SGS).

We will first present the principles of conditional GMM and SGS algorithms. Then, we will describe the methodology used to build, use and compare the models during the experiments. Finally, we will present the experiments themselves, the results and we will conclude on the efficiency of conditional GMM for local PDF estimation.

2 Algorithms Description

2.1 Gaussian Mixture Models

A mixture of Gaussians is a natural extension of a Gaussian distribution. It has the property of being able to represent any distribution as long as the number of Gaussians in the mixture is big enough. The PDF of a vector $\mathbf{v}$ can be modeled as:

$$p(\mathbf{v}) = \sum_{i=1}^{n} w_i \cdot \mathcal{N}(\mathbf{v}, \mu_i, \Sigma_i) \tag{1}$$

where w_i, μ_i and Σ_i are respectively the weight, the mean and the covariance matrix of the i^{th} of the n Gaussians of the model. All w_i are positive and sum to 1.

In the present study, we are interested in modeling the distribution $p(y|\mathbf{x})$ of a variable y given its position $\mathbf{x}$. The simplest way to do so is to start from the definition:

$$p(y|\mathbf{x}) = \frac{p(y, \mathbf{x})}{p(\mathbf{x})}. \tag{2}$$

We could model these two distributions separately. We have chosen instead to model the numerator $p(y, \mathbf{x})$ using a diagonal[1] GMM. It is then possible to write:

[1] i.e. a GMM where the covariance matrix of each Gaussian is diagonal

$$p(y, \mathbf{x}) = \sum_{i=1}^{n} w_i \mathcal{N}(y, \mu_{yi}, \sigma_{yi}) \mathcal{N}(\mathbf{x}, \mu_{\mathbf{x}i}, \Sigma_{\mathbf{x}i}) \tag{3}$$

and the denominator $p(\mathbf{x})$ can be obtained by simply removing the y contribution to $p(y, \mathbf{x})$. The expression of $p(y|\mathbf{x})$ becomes

$$p(y|\mathbf{x}) = \frac{\sum_{i=1}^{n} w_i \mathcal{N}(y, \mu_{yi}, \sigma_{yi}) \mathcal{N}(\mathbf{x}, \mu_{\mathbf{x}i}, \Sigma_{\mathbf{x}i})}{\sum_{i=1}^{n} w_i \mathcal{N}(\mathbf{x}, \mu_{\mathbf{x}i}, \Sigma_{\mathbf{x}i})} \tag{4}$$

which is equivalent to:

$$p(y|\mathbf{x}) = \sum_{i=1}^{n} W_i(\mathbf{x}) \mathcal{N}(y, \mu_{yi}, \sigma_{yi}) \tag{5}$$

with:

$$W_i(\mathbf{x}) = \frac{w_i \mathcal{N}(\mathbf{x}, \mu_{\mathbf{x}i}, \Sigma_{\mathbf{x}i})}{p(\mathbf{x})}. \tag{6}$$

It is easy to show that $\sum_{i=1}^{n} W_i(\mathbf{x}) = 1$. The new expression of $p(y|\mathbf{x})$ is then a Mixture of Gaussians on y whose weights are conditioned by $\mathbf{x}$. It is thus possible to model a dependency between y and $\mathbf{x}$ as long as the GMM contains more than one Gaussian.

2.2 Sequential Gaussian Simulations

The idea of stochastic simulations is to develop a spatial Monte Carlo generator that will be able to generate many, and in some sense equally probable, realizations of a random function (in general, described by a joint probability density function).

Simulations differ from regression models as reconstruction of the histogram and of the spatial variability of original data takes precedence over local accuracy.

In the present study, SGS were applied. This method consists of generating values corresponding to a large number of given spatial locations, using a modelization of the spatial correlation (called variogram model in Geostatistics) of a normally distributed known data set. The variogram model is used to compute the weights of a linear regression method called Kriging [1] (related to Gaussian Processes [4]). It is a weighted sum of a subset of data, allowing not only to estimate the value of new datum but also to compute the variance of this estimation. Each simulated value is then generated from a normal distribution whose mean and variance are computed by Kriging of the neighboring (original *and* previously simulated) data points, based on the global variogram model.

3 Methodology

In the experiments presented in this paper, the data set is segmented into two parts. The first part is the training set, defined as $\mathcal{Z} = (\mathbf{x}_i, z_i), \quad \forall i = 1, \ldots, N$, where $\mathbf{x}$ is the input vector (which represent the co-ordinates of the sample on a map), and z is the scalar output (studied value). The second part is the test set, defined as $\mathcal{Y} = (\mathbf{u}_i, y_i), \quad \forall i = 1, \ldots, M$, where $\mathbf{u}$ is the input vector, and the output y is hidden to the models.

The training set is used to tune the model's parameters and hyper-parameters (cf. section 3.1 and 3.2 for details). The tuned models are then used to predict the corresponding local PDF of the outputs of the test set. The relative performance of the models is evaluated on this last prediction.

3.1 GMM Experimental Protocol

GMM are trained using the Expectation-Maximization (EM) algorithm [2]. However, many hyper-parameters, i.e. user-defined parameters, need to be defined, such as the number of Gaussians in the mixture, the lower bounds for the variances of each dimension of the Gaussians, and the Dirichlet prior on the weights of the Gaussians. Moreover, in order to initialize the EM procedure, a few K-Means iterations are first performed. K-Means itself is randomly initialized. For each set of hyper-parameters studied, a k-fold cross-validation method is used to obtain out-of-sample estimation of the Mean Absolute Error (MAE) between the real value y and the mean of the local PDF given by the GMM. The model yielding the best estimation is kept and retrained on the whole training set.

It is important to note that the random initialization of K-Means introduces some variability in the performance of a given hyper-parameter set. However, we consider that the variability was not significantly perturbating the choice of the optimal hyper-parameter set.

For the present experiments, we used a 5-fold cross-validation procedure on 1000 training points. The optimal hyper-parameter set found contains 1000 Gaussians. This is quite a surprising result as GMMs are known to be good for generalization. A hypothesis is that this large number of Gaussians is caused by the complexity of the data set (cf. section 4.1).

3.2 SGS Experimental Protocol

SGS can only be used on normally distributed data. As a consequence, if this is not the case for original data, a Normal Score transformation is needed. This transformation consists of the function $\text{NS} : F_{\mathcal{Z}} \rightarrow \mathcal{N}(0,1)$, where $F_{\mathcal{Z}}(z)$ is the cumulative distribution function of z in $\mathcal{Z}$. Then, one can model the spatial correlation of the transformed data set (i.e. the variogram) and proceed to the simulations.

One simulation procedure consists of first defining a random path visiting each location $\mathbf{u}$ of $\mathcal{Y}$ once. Then, values $\hat{y}_{\text{NS}}$ are obtained as described in section 2.2 and back-transformed using $\text{NS}^{-1} : \mathcal{N}(0,1) \rightarrow F_{\mathcal{Z}}$.

The model of spatial correlation was developed taking into account anisotropy of the data and variability at different scales. 100 realizations were generated with SGS and post processed to extract an estimation of the local PDF at each point.

3.3 Model Comparison Method

Comparing local PDF models is very difficult when there is only one realization of the studied phenomenon. In order to solve this problem, we used large data sets (thousands of samples) from which we only kept a small portion for training (a typical training set in Geostatistics contains a few hundreds samples). Remaining data consitute the test set.

A k-nearest-neighbors (KNN) procedure can then be used on the test set to estimate the local PDF's central moments. For this paper, we used 20 neighbors[2] for each one of the 5000 test points. The same estimation can be done by the local PDF models, and thus, an "absolute" error criterion consists of comparing the Mean Absolute Errors made by each method while estimating these central moments.

The second comparison method consists of comparing the distributions quantiles obtained by each method at some points. The amount of data available in the test set is too small to permit a good absolute comparison to a real data distribution, but it is interesting to visualize the way both methods are constructing their local CDF. For both models, we used a 100-point random generation method to compute the quantiles.

[2]Please note that this value is arbitrary. What only maters is to have enough points to model the central moments and not too much to stay close to the studied point

4 Experiments

4.1 Data description

The training and testing data sets were randomly extracted from the digital elevation model (DEM) of Switzerland. This DEM is a grid of more than 95000 altitudes values, measured every 1 km. The training set contains 1000 points while the test set was reduced to 5000 points in order to speed up the computations.

The interesting aspect of this data set is that it is anisotropic (general tendency from South-West to North-East) and spatially non-stationary (altitudes are low and flat in the upper-left half, high and sharp in the other half). This is a typical case where Geostatistical methods are very difficult to use.

Note that for numerical stability reasons, the co-ordinate values and the altitudes have been linearly transformed for GMM computations.

4.2 Moments prediction results

SwissDEM data set	Sequential Gaussian Simulations	Conditional GMM
MAE on 1^{st} moment	159	146
MAE on 2^{nd} moment	170	101
MAE on 3^{rd} moment	323	249
MAE on 4^{th} moment	253	142

Table 1: Mean Absolute Error (MAE) of SGS and GMM on the prediction of the first four local central moments of the test set. The reference moments were computed using KNN on test data. Values are expressed in meters.

Table 1 confirms that even if the 1^{st} moment predictions from both methods give similar results, the conditional GMM is significantly better than SGS in predicting the higher moments. This was expected given the complexity of the data set for Geostatistical methods.

4.3 Quantiles evaluation comparison

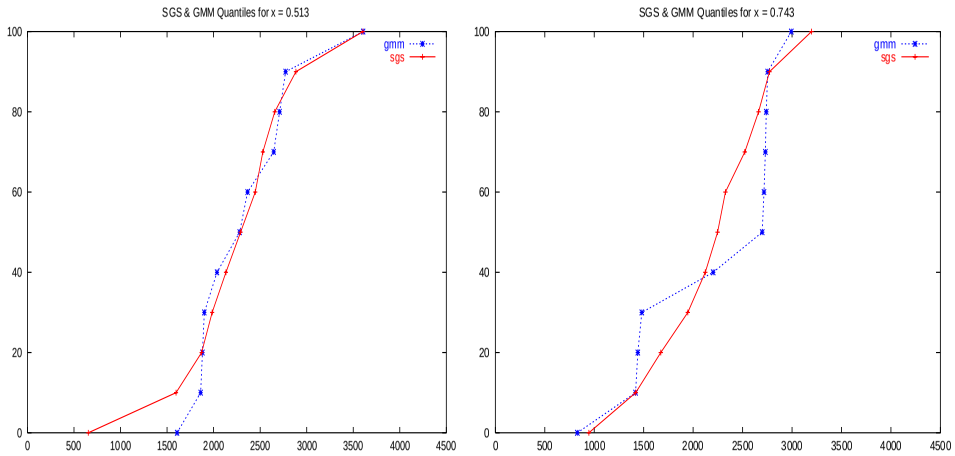

Figure 1: Comparison between altitude quantiles from SGS and conditional GMM for two different points inside the test set. SGS quantiles curves are in plain red. GMM's are in dashed blue

In figure 1, the differences between SGS and conditional GMM for the reconstruction of the local PDF are clearer. While quantile curves from the SGS are uni-modal and quite similar for both locations, those from the conditional GMM are very different from one another. In addition, the second curve of GMM is clearly a multi-modal local PDF, something which is impossible to model with SGS.

5 Conclusion

Conditional GMM proved their efficiency in solving local PDF estimation on a complex data set. When comparing with SGS, they appear to be more flexible (i.e. they need less theoretical constraints on data). They also need less expert knowledge to be tuned. Another advantage is that they give an analytical solution to local PDF estimation, which can be easily reused for risk mapping. However, the method should be improved with a better initialisation of the k-means procedure. It is also necessary to test the efficiency of GMM when the training inputs are not independently and identically distributed, which is often the case in Geostatistics. It might also be possible to improve the training procedure by optimizing directly a local model instead of conditioning a global model.

6 Acknowledgements

This work was supported by Swiss National Science Foundation (CARTANN project: FN 20-63859.00).

The GMM experiments were conducted using the Torch Machine Learning Library[3]. The SGS experiments were performed using the Geostat Office software.

Thanks to Johnny Mariethoz for his help during the long debugging phase.

References

[1] Chauvet, P.: Processing Data with a Spatial Support: Geostatistics and its Methods. Cahiers de Géostatistique 4. ENSMP Paris (1993)

[2] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum-Likelihood from Incomplete Data via the EM Algorithm. Journal of Royal Statistical Society B (1977)

[3] Lantuéjoul, Ch.: Geostatistical Simulations, Models and Algorithms. Springer, Berlin (2002)

[4] Williams, C.K.I., Rasmussen, C.E.: Gaussian Processes for Regression. in Tourersky, Mozer, Hasselmo (eds.): Advances in Neural Information Processing Systems 8, MIT Press (1996)

[3] `http://www.torch.ch`