Bayesian Smoothing and Information Geometry

Rudolf Kulhavý
Honeywell Laboratories and
Institute of Information Theory and Automation
Pod vodárenskou věží 4, Prague, Czech Republic
rudolf.kulhavy@honeywell.com

Abstract. Local, cased-based modeling offers a natural way of capturing the complex behavior of data. As such, it has been a subject of intensive research in computational statistics, machine learning and system identification. Also, it has been applied successfully to numerous problems in different fields. Yet, the very concept of smoothing continues to be perceived as somewhat heuristic. The purpose of this paper is to help understand better the connection of smoothing algorithms to Bayesian statistics and to present a natural geometry of local modeling.

1 Introduction

Local modeling is an intuitively appealing paradigm of learning from data. At its root is the observation (supported by the everyday life experience) that one does not need to build a global model in order to predict response in a particular case. In fact, one can even improve prediction when minimizing the local error only, using a simpler model, even though estimated from less data.

Local models capture easily a nonlinear behavior, cope with bias problems at boundaries and in regions of high curvature, handle naturally multiple-mode data, adapt to changes in the data behavior, need only a fraction of historical data to work, scale up well to huge datasets, can be estimated using closed-form algorithms, and, last but not least, are easy to understand and interpret.

Such attractive properties does not come for free. Rather than arriving at a single globally valid model, the user ends up with multiple *ad hoc* models of varying quality, depending largely on the amount of data available for particular cases. Local modeling requires all historical data available on demand, involves a database-intensive step of retrieving relevant data, needs data organized properly for quick retrieval, requires careful tuning of bandwidth parameters for optimum data fit, and may suffer of the curse of dimensionality. The steady increase in the computer and database performance is removing some of the earlier technical hurdles, yet large-scale applications of local modeling are still far from routine.

As many good ideas in science and engineering, local modeling has been long a recurrent concept, discovered or rediscovered more or less independently by different research communities. In *computational statistics*, local modeling has been studied within the frameworks of data smoothing, local fitting, locally weighted regression and classification, kernel-based

methods, and non-parametric estimation [6, 7, 11]. In *machine learning*, it has been known as local learning, case-based reasoning, example-based reasoning, memory-based learning, instance-based learning, or lazy learning [1, 4, 3]. In *system identification*, local modeling has been explicitly present in the concepts of just-in-time learning and on-demand modeling [10, 20].

In spite of much research attention and practical usage and solid understanding of the structure of smoothing algorithms [21], many conceptual questions remain. Is data smoothing a technique or method? Can the smoothing formulae be derived from a more general principle? What Bayesian interpretation can be given to the local modeling? Can one build a local modeling theory without referring to an underlying global model? What sort of geometry does the local modeling give rise to?

In the following, we try to shed some light on some of these questions. In particular, we demonstrate that the smothing algorithms can be derived by approximation of the Bayesian estimation, with a specific choice of prior distribution over a set of local, cased-based models. In addition, we show a natural geometry of local modeling, based on measuring information carried by the empirical density relative to the model-based conditional densities.

2 Problem Statement

Our focus will be on fitting a model to the historical data. This is a problem narrower in scope than data modeling in large, which includes the data preprocessing and model selection tasks.

Data Transforms. The data definition and model selection proceeds typically in three steps.

First, we recognize directly manipulated inputs u_k to the underlying system at time $k = 1, \ldots, N$ and distinguish them from the outputs y_k of the system at the same time. The outputs represent the response of the system to the past history of data $y^{k-1} = (y_{k-1}, \ldots, y_1)$ and $u^k = (u_k, \ldots, u_1)$. At this stage, we consider a dynamic system that can be described through the (stochastic) functional relationship

$$y_k = F(u^k, y^{k-1}), \quad k = 1, \ldots, N\,.$$

Second, we introduce a vector of auxiliary variables x_k, which capture the system dynamics. A popular time-series model defines x_k-entries through the time-lagged values $u_k, u_{k-1}, \ldots$ and $y_{k-1}, y_{k-2}, \ldots$. Other functions of the data history u^k, y^{k-1}, such as time aggregates or dynamically filtered values, can be used as well. The objective of the second phase is to turn the original dynamic system into a static (still stochastic) one

$$y_k = f(x_k), \quad k = 1, \ldots, N\,.$$

Third, the data vector x can be mapped onto a feature vector $\phi_k = \phi(x_k)$, possibly of much higher dimension. The purpose of this step is to come up with a simpler, parametric representation of the map $f(\cdot)$. A particular example of such a parametric model is the linear regression, for scalar response y_k,

$$y_k = \theta^T \phi_k + \varepsilon_k\,.$$

Here θ stands for the vector of regression coefficients and ε_k accounts for the unpredictable component of the model.

Data Set. We assume that the data is available in a table composed of x_k and y_k values for $k = 1, \ldots, N$

$$\left|\begin{array}{ccc|ccc} x_{1,1} & \ldots & x_{1,m} & y_{1,1} & \ldots & y_{1,n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & \ldots & x_{N,m} & y_{N,1} & \ldots & y_{N,n} \end{array}\right| .$$

The x-entries represent predictors or explanatory variables whereas y-entries stand for responses or target variables. In general, the predictors and responses can be continuous or discrete or mixed (e.g., when forecasting the energy load and price tariff as a function of time of day and day of week).

We use the term *case* when speaking of a specific value of the vector x.

Conventions. In order to cover the cases of continuous and discrete random variables with a single notation, we consider each random variable Z with values in $\mathcal{Z}$ and probability distribution P on a measurable space $(\mathcal{Z}, \mathbf{Z})$ described by the density (Radon-Nikodym derivative) of P with respect to a dominating (Lebesgue or counting or product) measure μ

$$p(z) = \frac{P(\mathrm{d}z)}{\mu(\mathrm{d}z)} .$$

In the sequel, we do not mention the measurable space and the probability distribution. When using the same symbol for the densities p and dominating measures μ, we always include the argument to identify them uniquely. The reader not familiar with the measure theory can simply replace $\int p(z)\, \mu(\mathrm{d}z)$ for continuous and discrete variables z with ordinary integration and summation, respectively.

Objective. The paper adopts a statistical perspective of learning. The data composed of responses observed under particular cases

$$y_1 | x_1, \ldots, y_N | x_N$$

is supposed to be a sample from the conditional density

$$p(y|x, \theta)$$

where θ stands for the unknown parameters.

The objective is to estimate the conditional density from the sample

$$y_1 | x_1, \ldots, y_N | x_N \to \hat{p}_N(y|x) .$$

3 Probability-Based Inference

The essence of the Bayesian approach to parameter estimation and response prediction is the symmetrical treatment of stochastic data and uncertain parameters. Both stochastic and uncertain quantities are dealt with as random variables.

Joint Density. The starting point for derivation of the conditional density of the unknown parameters is to express the joint density $p(y^N, x^N, \theta)$ of data and parameters in terms of model assumptions. A recursive application of the chain rule makes it possible to decompose the joint density as follows

$$p(y^N, x^N, \theta) = \prod_{k=1}^{N} p(y_k|x_k, y^{k-1}, x^{k-1}, \theta)\, p(x_k|y^{k-1}, x^{k-1}, \theta)\, p(\theta)\,.$$

By the model assumption, the response y_k at any time k depends on the past data only through the current predictor x_k, i.e.,

$$p(y_k|x_k, y^{k-1}, x^{k-1}, \theta) = p(y_k|x_k, \theta)\,.$$

The joint density thus simplifies to

$$p(y^N, x^N, \theta) = \prod_{k=1}^{N} p(y_k|x_k, \theta)\, p(x_k|y^{k-1}, x^{k-1}, \theta)\, p(\theta)\,.$$

Furthermore, we assume that the predictor x_k at any time k is independent of θ given the past data

$$p(x_k|y^{k-1}, x^{k-1}, \theta) = p(x_k|y^{k-1}, x^{k-1})\,.$$

Under this assumption, introduced as *natural conditions of control* in [18], the joint density takes the form

$$p(y^N, x^N, \theta) = \prod_{k=1}^{N} p(y_k|x_k, \theta)\, p(x_k|y^{k-1}, x^{k-1})\, p(\theta)\,. \tag{1}$$

Posterior Density. Now, let us apply the chain rule in the other direction

$$p(y^N, x^N, \theta) = p(\theta|y^N, x^N)\, p(y^N, x^N)\,. \tag{2}$$

Combining the expressions (1) and 2), we obtain

$$p(\theta|y^N, x^N)\, p(y^N, x^N) = \prod_{k=1}^{N} p(y_k|x_k, \theta)\, p(x_k|y^{k-1}, x^{k-1})\, p(\theta)\,.$$

From this, the *posterior* density of the random variable Θ conditioned on y^N, x^N follows easily

$$p(\theta|y^N, x^N) \propto p(\theta) \prod_{k=1}^{N} p(y_k|x_k, \theta)\,.$$

Here $\propto$ stands for proportionality, i.e., equality up to a normalizing constant.

After introducing a short-cut notation

$$p_0(\theta) \stackrel{\triangle}{=} p(\theta),\;\; p_N(\theta) \stackrel{\triangle}{=} p(\theta|y^N, x^N),\;\; s_\theta(y|x) \stackrel{\triangle}{=} p(y|x, \theta)\,,$$

we obtain the posterior density formula

$$p_N(\theta) \propto p_0(\theta) \prod_{k=1}^{N} s_\theta(y_k|x_k)\,. \tag{3}$$

Likelihood Function. The conditional density of the observed data taken as a function of the unknown parameter for given data is known as a *likelihood function*

$$l_N(\theta) = \prod_{k=1}^{N} s_\theta(y_k|x_k)\,. \tag{4}$$

Using the likelihood function, the posterior density can be rewritten in a compact form

$$p_N(\theta) \propto p_0(\theta)\, l_N(\theta)\,.$$

Predictive Density. The unknown parameter can be eliminated (integrated out) using probability calculus rules

$$p(y|x, y^N, x^N) = \int p(y|x,\theta)\, p(\theta|y^N, x^N)\, \mu(\mathrm{d}\theta)\,.$$

Using a simpler notation

$$s_N(y|x) \stackrel{\triangle}{=} p(y|x, y^N, x^N)\,,$$

we can rewrite the predictive density as follows

$$s_N(y|x) = \int s_\theta(y|x)\, p_N(\theta)\, \mu(\mathrm{d}\theta)\,. \tag{5}$$

Example. Consider a linear normal regression model

$$Y_k = \theta^T \phi(X_k) + E_k, \quad E_k \sim N(0, \sigma^2)\,.$$

The vector of regression coefficients θ is the unknown parameter of the model. The variance σ^2 is considered known for simplicity.

The conditional density of Y given $X = x$ is

$$s_\theta(y|x) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y - \theta^T\phi(x))^2\right\}\,.$$

The likelihood function (4) for a sample y^N, x^N takes the form

$$\begin{aligned} l_N(\theta) &= \prod_{k=1}^{N} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y_k - \theta^T\phi(x_k))^2\right\} \\ &= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{N}(y_k - \theta^T\phi(x_k))^2\right\}, \end{aligned}$$

which can be rewritten as

$$l_N(\theta) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\, N\, V_N\right\} \exp\left\{-\frac{1}{2\sigma^2}\, N\, (\theta - \hat\theta_N)^T C_N\, (\theta - \hat\theta_N)\right\}$$

with the statistics

$$\begin{aligned} \hat\theta_N &= C_N^{-1}\, \mathbf{E}_N(\Phi Y)\,, & (6)\\ V_N &= \mathbf{E}_N(Y^2) - \mathbf{E}_N(Y\Phi^T)\, C_N^{-1}\, \mathbf{E}_N(\Phi Y)\,, & (7)\\ C_N &= \mathbf{E}_N(\Phi\Phi^T)\,. & (8) \end{aligned}$$

where $\Phi = \phi(X)$ and $\mathbf{E}_N(\cdot)$ denotes the empirical mean.

4 Information-Based Inference

We will show that the Bayesian inference implicitly measures the amount of information carried by the data relative to particular models parametrized by θ.

Empirical Density. For a given sample y^N, x^N, let us define the *joint empirical density* of (Y, X) as

$$r_N(y,x) = \frac{1}{N}\sum_{k=1}^{N} \delta(y-y_k, x-x_k)$$

where $\delta(y-y_k, x-x_k)$ stands for the density of a point-mass distribution at (y_k, x_k), with the properties

$$\delta(y-y_k, x-x_k) = 0 \text{ if } y \neq y_k \text{ or } x \neq x_k$$

and

$$\iint f(y,x)\,\delta(y-y_k, x-x_k)\,\mu(\mathrm{d}y)\,\mu(\mathrm{d}x) = f(y_k, x_k)$$

for all integrable f.

Let $\mathcal{X}_N \subset \mathcal{X}$ be the set of all distinct cases observed in the sample $x_1, \ldots, x_N$. We denote the empirical and model-based densities of Y for a particular case $x \in \mathcal{X}_N$ as

$$r_{N,x}(y) \stackrel{\triangle}{=} r_N(y|x), \quad s_{\theta,x}(y) \stackrel{\triangle}{=} s_\theta(y|x)\,.$$

Note that if the vector X includes continuous random variables, the probability of observing a perfectly identical case once again is theoretically zero and practically very low. The empirical density $r_{N,x}(y)$ is thus typically composed of a single δ-function. This does not affect the validity of the results presented in this and next sections, although it makes them perhaps less intuitive compared with discrete or discretized variables. The drastical lack of data just exhibits that when facing infinitely many (or just too many) cases, we cannot learn the response to one particular case without considering responses to other, related cases. More on this in Section 7.

Kerridge Inaccuracy. With the above notation, we can define the *Kerridge inaccuracy* [12] of conditional densities as

$$K(r_{N,x} : s_{\theta,x}) = \int r_{N,x}(y)\,\log\,\frac{1}{s_{\theta,x}(y)}\,\mu(\mathrm{d}y)\,.$$

The empirical expectation of the inaccuracy of conditional densities yields the *conditional Kerridge inaccuracy*

$$\begin{aligned} K(r_N : s_\theta) &= \mathbf{E}_N\, K(r_{N,x} : s_{\theta,x}) \\ &= \int r_N(x)\, K(r_{N,x} : s_{\theta,x})\,\mu(\mathrm{d}x) \\ &= \frac{1}{N}\sum_{x\in\mathcal{X}_N} N_x\, K(r_{N,x} : s_{\theta,x}) \\ &= \frac{1}{N}\sum_{k=1}^{N} K(r_{N,x_k} : s_{\theta,x_k})\,. \end{aligned} \tag{9}$$

Bayesian Inference Revisited. Using Kerridge inaccuracy and assuming $s(y|x) > 0$ for all (y, x), we can rewrite the likelihood function as

$$\begin{aligned}\prod_{k=1}^{N} s_\theta(y_k|x_k) &= \exp\Big(-\sum_{k=1}^{N} \log \frac{1}{s_\theta(y_k|x_k)}\Big) \\ &= \exp\Big(-N \iint r_N(y,x)\, \log \frac{1}{s_\theta(y|x)}\, \mu(\mathrm{d}y)\, \mu(\mathrm{d}x)\Big) \\ &= \exp\left(-N\, K(r_N : s_\theta)\right) \\ &= \exp\left(-N\, \mathbf{E}_N\, K(r_{N,X} : s_{\theta,X})\right)\,.\end{aligned}$$

With this expression, the posterior density becomes

$$p_N(\theta) \propto p_0(\theta)\, \exp\big(-N\, K(r_N : s_\theta)\big) \tag{10}$$

or, alternatively,

$$p_N(\theta) \propto p_0(\theta)\, \exp\big(-N\, \mathbf{E}_N\, K(r_{N,X} : s_{\theta,X})\big)\,. \tag{11}$$

Example. Let us assume the general regression model with a normally distributed additive noise

$$Y = f(X) + E, \quad E \sim N(0, \sigma^2)\,.$$

The sampling density for the model is

$$s_f(y|x) = (2\pi\sigma^2)^{\frac{1}{2}}\, \exp\left\{-\frac{1}{2\sigma^2}\,(y - f(x))^2\right\}.$$

The conditional inaccuracy relative the model is

$$K(r_N : s_\theta) = \frac{1}{2}\, \log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\, \frac{1}{N} \sum_{k=1}^{N} (y_k - f(x_k))^2\,.$$

Note it is a linear transform of the *empirical risk functional* in [22] and the *empirical error of* f in [9].

Prior Knowledge. We have seen that the likelihood function can be written as

$$l_N(\theta) = \exp\big(-N\, K(r_N : s_\theta)\big)\,.$$

Let us choose the prior density of Θ in the same form

$$p_0(\theta) = \exp\big(-\nu_0\, K(\rho_0 : s_\theta)\big) \tag{12}$$

where $\rho_0(y, x)$ denotes the prior density of (Y, X) and ν_0 stands for the "number of data" ρ_0 is built on, i.e., the degree of belief in ρ_0. The form (12) can be seen as a *generalized conjugate prior*. Indeed, the posterior density derived from such prior preserves its form

$$p_N(\theta) = \exp\big(-\nu_N\, K(\rho_N : s_\theta)\big)$$

while the statistics ν_N and ρ_N are updated as follows

$$\begin{aligned} \nu_N &= \nu_0 + N\,, \\ \rho_N(y,x) &= \frac{\nu_0}{\nu_0+N}\,\rho_0(y,x) + \frac{N}{\nu_0+N}\,r_N(y,x)\,. \end{aligned}$$

The prior can be built using *virtual* data provided by experts, generated by simulation models, created by controlled replication of existing data, etc. The virtual data can be combined from various sources, weighted according to the relevance and reliability of source, and turned into the prior density ρ_0 and the degree of belief ν_0.

Big Picture. The concepts introduced so far can be given the following interpretation. The model is a collection of densities of Y parametrized by the parameters θ *and* the case x. For each x observed in the data, one can define the empirical distribution of Y. The essence of modeling is in fitting of the empirical densities with model-based densities for each observed case (see Fig. 1). In Bayesian inference, the goodness of fit is expressed through the conditional (i.e., average) Kerridge innacuracy (9). The impact of data depends on the local density of x-points (cf. Fig. 2).

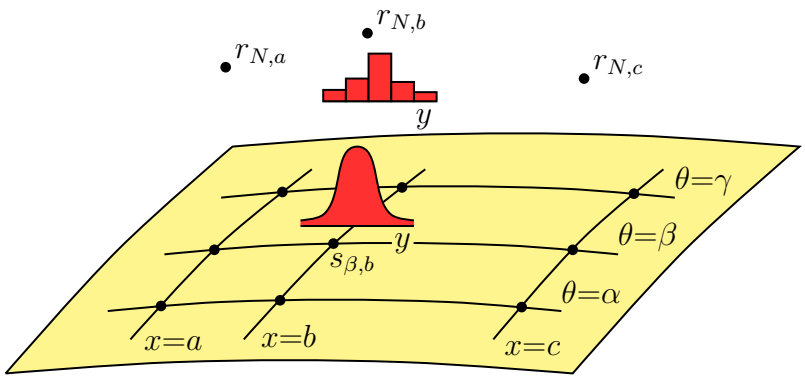

Figure 1: A statistical manifold of model- and case-based densities that approximate the empirical densities for the observed cases.

Note that the empirical distribution is discrete, concentrated on a finite number of points, but Kerridge inaccuracy is well defined even for continuous model distributions. Compare it with Kullback-Leibler divergence, which is infinite in this case.

To make the picture more intuitive for continuous cases x, one can consider a smoothened version of the conditional empirical density, e.g., by taking cross-section of a multivariate histogram of (Y, X).

The reader is referred to [13] for more discussion on various interpretations of Kerridge inaccuracy and its relationship to Kullback-Leibler divergence [17] and Shannon entropy [19].

5 Single-Case Geometry

We start exploring the geometry of Bayesian inference by analyzing the single case $X = x$ first. We assume to have observed responses $(y_1, \ldots, y_{N_x})$ under this particular case. Note again (cf. discussion in the previous section) that the number of available samples generally

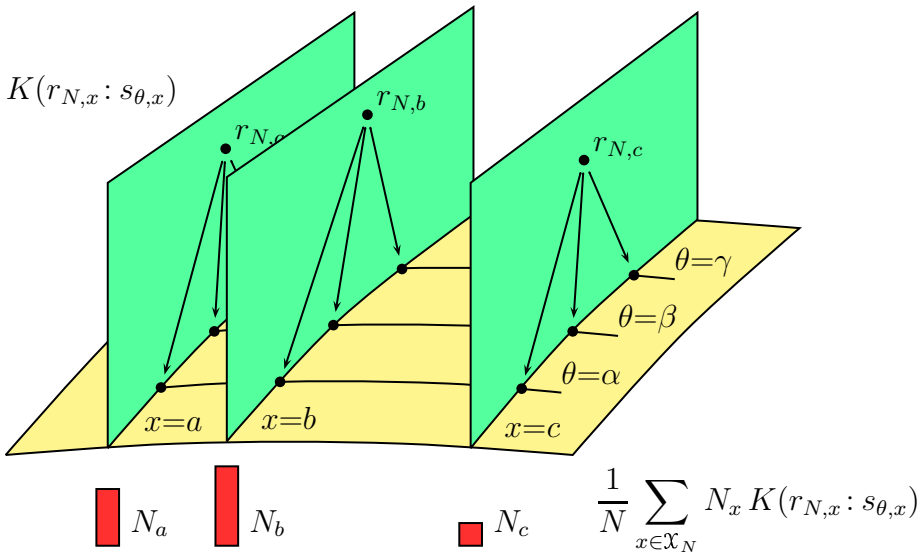

Figure 2: The overall goodness of fit is given by the empirical mean of Kerridge inaccuracy of case-based empirical densities relative to the corresponding model densities.

decreases with the increasing number of distinct cases x, but this fact does not affect the following analysis. Our objective is to fit the empirical density $r_{N,x}$ with a case-based sampling density $s_{\theta,x}$.

We will omit the subscript x for a while to simplify the notation.

Exponential Family. Consider an exponential family $\mathcal{S}_h$ composed of densities

$$s_\lambda(y) = s_0(y)\, \exp\left(\lambda^T h(y) - \psi(\lambda)\right)$$

where s_0 is a fixed density (the family origin), θ is a vector parameter, h is a vector canonical statistic and $\psi(\lambda)$ is logarithm of the normalizing divisor

$$\psi(\lambda) = \log \int s_0(y)\, \exp\left(\lambda^T h(y)\right)\, \mu(\mathrm{d}y)\,.$$

The parameter λ of the family depends on the model and a particular case, $\lambda = \lambda(\theta, x)$.

h-Projection. We define a h-projection $s_{\hat\lambda}(y)$ of $r_N(y)$ onto $\mathcal{S}_h$ by the equality

$$\int s_{\hat\lambda}(y)\, h(y)\, \mu(\mathrm{d}y) = \int r_N(y)\, h(y)\, \mu(\mathrm{d}y)\,.$$

Note this is a necessary condition for $\hat\lambda$ to minimize $K(r_N : s_\lambda)$

$$\begin{aligned} 0 &= \nabla_\lambda K(r_N : s_{\hat\lambda}) \\ &= \int r_N(y)\, h(y)\, \mu(\mathrm{d}y) - \int s_{\hat\lambda}(y)\, h(y)\, \mu(\mathrm{d}y)\,. \end{aligned}$$

The expectation of $h(Y)$ with respect to r_N amounts to its sample average

$$\int r_N(y)\, h(y)\, \mu(\mathrm{d}y) = \frac{1}{N}\sum_{k=1}^{N} h(y_k) \triangleq \bar h_N\,.$$

We introduce the set of all densities with the same h-projection as the empirical density has

$$\mathcal{R}_N = \left\{ \text{density } r(y) \colon \int r(y)\, h(y)\, \mu(\mathrm{d}y) = \bar{h}_N \right\} .$$

Pythagorean Relation. Let $\mathcal{S}_h$ be exponential and $s_{\hat{\lambda}}$ be a h-projection of r_N onto $\mathcal{S}_h$. Then, for every $s_\lambda \in \mathcal{S}_h$ and every $r \in \mathcal{R}_N$, it holds

$$K(r : s_\lambda) = K(r : s_{\hat{\lambda}}) + D(s_{\hat{\lambda}} \| s_\lambda) \tag{13}$$

where

$$D(s \| s') = \int s(y)\, \log \frac{s(y)}{s'(y)}\, \mu(\mathrm{d}y)$$

is Kullback-Leibler divergencee of $s(y)$ relative to $s'(y)$.

The identity (13) follows directly by definitions of Kerridge inaccuracy and Kullback-Leibler divergence

$$\begin{aligned} &\int \left[r(y) - s_{\hat{\lambda}}(y)\right] \left[\log\, s_{\hat{\lambda}}(y) - \log\, s_\lambda(y)\right] \mu(\mathrm{d}y) \\ &= \int \left[r(y) - s_{\hat{\lambda}}(y)\right] (\hat{\lambda} - \lambda)^T h(y)\, \mu(\mathrm{d}y) \\ &= (\hat{\lambda} - \lambda)^T \int \left[r(y) - s_{\hat{\lambda}}(y)\right]^T h(y)\, \mu(\mathrm{d}y) = 0 \,. \end{aligned} \tag{14}$$

The relation (13) can be viewed as a version of Pythagorean-like theorem. It allows us to decompose the Kerridge inaccuracy of the empirical density r_N relative to an exponential density s_λ with the canonical statistic $h(y)$ into sum of two terms – the Kerridge inaccuracy of r_N relative to the h-projection $s_{\hat{\lambda}}$ plus the Kullback-Leibler divergence of $s_{\hat{\lambda}}$ relative to s_λ (cf. Fig. 3).

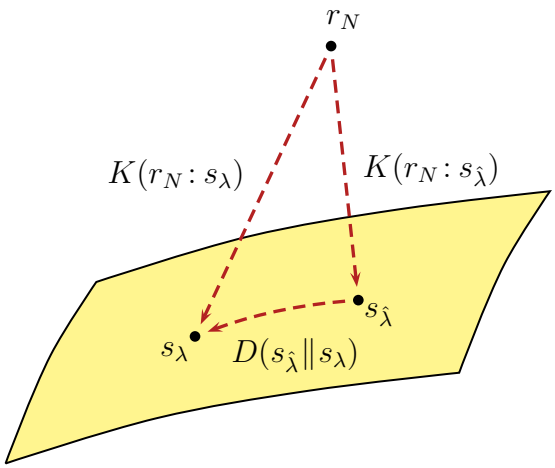

Figure 3: The Pythagorean relation for the h-projection $s_{\hat{\lambda}}$ of the empirical density r_N onto an exponential family $\mathcal{S}_h$.

Differential-Geometric View. The condition (14) suggests that the Pythagorean relationship can be rewritten as

$$\int \big[r(y) - s_{\hat\lambda}(y)\big]\,\big[\log\, s_{\hat\lambda}(y) - \log\, s_\lambda(y)\big]\,\mu(\mathrm{d}y) = 0\,.$$

This can be taken as a definition of *orthogonal projection of* r_N *onto* $\mathcal{S}_h$ or, from a dual viewpoint, *orthogonal projection of* s_λ *onto* $\mathcal{R}_N$.

In contrast to the orthogonality of vectors in the Euclidean space, the appearance of logarithm makes the above condition asymmetric in $r(y)$ and $s(y)$. Consequently, in the space of probability distributions there is no straightforward analogy of the "natural" inner product as we know it from the Euclidean space.

We can, however, rewrite the condition as follows

$$\int s_{\hat\lambda}(y)\,\frac{\partial}{\partial\mu}\log\Big[\mu\, r_N(y) + (1-\mu)\, s_{\hat\lambda}(y)\Big]\Big|_{\mu=0} \cdot \frac{\partial}{\partial\lambda}\,\log\Big[s_0(y)\,\exp\big(\lambda\, h(y) - \psi(\lambda)\big)\Big]\Big|_{\lambda=\hat\lambda}\,\mathrm{d}y = 0\,.$$

This definition gives rise to a specific kind of Riemannian geometry on a differentiable manifold of probability distributions. The underlying metric tensor is closely related to the *Fisher information matrix*. In contrast to the classical Riemannian-geometric picture, two dual affine connections need to be considered at the same time in order to explain the asymmetry of the geometry. In these connections, *exponential and mixture families* of probability distributions act as analogy of hyperplanes in the Euclidean case.

Elaboration of this view is beyond the scope of the paper. The interested reader is referred for details to [5, 8, 2, 13, 23].

Dual Optimization Tasks. It follows directly from the Pythagorean relation (13) that the projection $s_{\hat\lambda}$ is a solution to two dual optimization tasks (cf. Fig. 4).

Maximum Likelihood Estimate: For every $r \in \mathcal{R}_N$

$$K(r\colon s_{\hat\lambda}) = \min_\lambda\, K(r\colon s_\lambda)\,.$$

Note that replacing of r_N with ρ_N yields the *maximum a posteriori probability* estimate.

Maximum Entropy Estimate: For every $s_\lambda \in \mathcal{S}_h$,

$$D(s_{\hat\lambda}\|\, s_\lambda) = \min_{r\in\mathcal{R}_N} D(r\|\, s_\lambda)\,.$$

Note that minimizing Kullback-Leibler divergence is – up to a term relative to s_λ – equivalent to maximizing Shannon entropy. This explains the *maximum entropy* label given to this task.

Inverse Problem. At the beginning of this section, we have said that the parameter λ of the exponential family $\mathcal{S}_h$ depends on the model θ and a particular case x, i.e., $\lambda = \lambda(\theta, x)$. Given the projection $\hat\lambda_N$, we can define the estimate $\hat\theta_{N,x}$ of θ as a solution to the equation

$$\lambda(\theta, x) = \hat\lambda_N$$

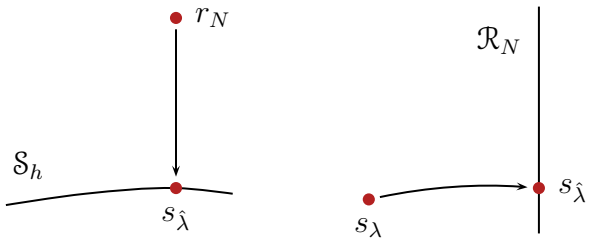

Figure 4: The h-projection $s_{\hat{\lambda}}$ of the empirical density r_N onto an exponential family $\mathcal{S}_h$ is a solution to the maximum likelihood and maximum (relative) entropy tasks.

for a given x. Typically, $\hat{\theta}_{N,x}$ is not unique. For any solution $\hat{\theta}_{N,x}$, the Pythagorean relation for conditional densities $s_{\theta,x}(y)$ reads

$$K(r_{N,x} : s_{\theta,x}) = K(r_{N,x} : s_{\hat{\theta}_{N,x},x}) + D(s_{\hat{\theta}_{N,x},x} \| s_{\theta,x}) \,. \tag{15}$$

6 Average-Case Geometry

We will consider now all observed cases x jointly. Our objective is to fit a single model s_θ to the data so that the average inaccuracy is minimized.

Exponential Family. Consider an *exponential family* $\mathcal{S}_h$ composed of densities

$$s_\theta(y|x) = s_0(y|x) \, \exp\left(\theta^T h(y,x) - \psi(\theta,x)\right)$$

where s_0 is a fixed density, θ is a vector parameter, h is a vector canonical statistic and ψ is logarithm of the normalizing divisor

$$\psi(\theta,x) = \log \int s_0(y|x) \, \exp\left(\theta^T h(y,x)\right) \, \mu(\mathrm{d}y) \,.$$

h-Projection. We define a h-projection of $r_N(y,x)$ onto $\mathcal{S}_h$ by the equality

$$\int r_N(x) \int s_{\hat{\theta}}(y|x) \, h(y,x) \, \mu(\mathrm{d}y) \, \mu(\mathrm{d}x) = \iint r_N(y,x) \, h(y,x) \, \mu(\mathrm{d}y) \, \mu(\mathrm{d}x) \,.$$

This is a necessary condition for $\hat{\theta}$ to minimize $K(r_N : s_\theta)$

$$0 = \nabla_\theta \, K(r_N : s_{\hat{\theta}}) \,.$$

Pythagorean Relation. Let $\mathcal{S}_h$ be exponential and $s_{\hat{\theta}}$ be a h-projection of r_N onto $\mathcal{S}_h$. Then, for every $s_\theta \in \mathcal{S}_h$, the following Pythagorean relation holds

$$K(r_N : s_\theta) = K(r_N : s_{\hat{\theta}}) + \mathbf{E}_N \, D(s_{\hat{\theta}} \| s_\theta) \,.$$

The relation can be alternatively written as

$$\mathbf{E}_N \, K(r_{N,X} : s_{\theta,X}) = \mathbf{E}_N \, K(r_{N,X} : s_{\hat{\theta},X}) + \mathbf{E}_N \, D(s_{\hat{\theta},X} \| s_{\theta,X}) \,.$$

Compare this *global estimation* formula with the *local estimation* formula (15)

$$K(r_{N,x}\colon s_{\theta,x}) = K(r_{N,x}\colon s_{\hat{\theta},x}) + D(s_{\hat{\theta},x}\|\, s_{\theta,x})\,.$$

Clearly, the former is just the expected version of the latter.

The global model found through the sample average of Kerridge inaccuracy is clearly a tradeoff. The use of all available data sets N to its maximum, thus reducing the total uncertainty of estimation. On the other hand, unless a single model with constant θ explains well the data behavior for all cases x, the global error expressed through the case-averaged inaccuracy typically increases.

Example. The conditional inaccuracy for the linear normal regression model takes the form

$$K(r_N\colon s_\theta) = \frac{1}{2}\,\log\,(2\pi\sigma^2) + \frac{1}{2\sigma^2}V_N + \frac{1}{2\sigma^2}\big(\theta-\hat{\theta}_N\big)^T C_N\big(\theta-\hat{\theta}_N\big)$$

with the statistics $\hat{\theta}_N$, V_N and C_N introduced earlier through (6)–(8).

The posterior expectation $\mathcal{E}_N(\cdot)$ of the conditional inaccuracy follows after some algebraic manipulations

$$\mathcal{E}_N\, K(r_N\colon s_\theta) = \frac{1}{2}\,\log\,(2\pi\sigma^2) + \frac{1}{2}\,\frac{V_N}{\sigma^2} + \frac{1}{2}\,\frac{\dim\theta}{N}$$

assuming that the prior is flat and the posterior covariance is positive definite.

The formula combines all ingredients of the modeling task – the sum of residuals squared V_N, the model variance σ^2, the model complexity $\dim\theta$, and the sample size N. It suggests that the model performance can be tuned up by balancing the coherence of data ("use *only* relevant data"), sample size ("use *all* relevant data") and model complexity ("strive for the *simplest* model").

7 Similar-Case Modeling

In Sections 5 and 6, we have considered two extreme approaches to modeling – building of a strictly local model for the case-specific data and fitting of a global model to all the data. In this section, we show how one can smoothly move between the extremes.

Local Models. Assume a set of local models $\mathcal{M} = \{\theta_x\colon x \in \bar{\mathcal{X}}\}$ with x coming from a *finite* set $\bar{\mathcal{X}}$ such that $\mathcal{X}_N \subset \bar{\mathcal{X}} \subset \mathcal{X}$.

The posterior density over the model set $\mathcal{M}$ follows by the standard probability calculus rules

$$p_N(\{\theta_x\}) \propto p_0(\{\theta_x\}) \prod_{k=1}^{N} s_{\theta_{x_k}}(y_k|x_k)\,.$$

Using the Kronecker delta

$$\delta_{a,b} = \begin{cases} 1 & \text{if } a=b \\ 0 & \text{if } a\neq b \end{cases}\,,$$

the local likelihood $l_{N,x}(\theta_x)$ at $X = x$ and the corresponding number of data N_x can be defined as

$$\begin{aligned} l_{N,x}(\theta_x) &= \prod_{k=1}^{N} s_{\theta_x}(y_k|x_k)^{\delta_{x,x_k}} , \\ N_x &= \sum_{k=1}^{N} \delta_{x,x_k} . \end{aligned}$$

The *joint posterior density* of all models can be rewritten in terms of the local likelihoods as

$$p_N(\{\theta_x\}) \propto p_0(\{\theta_x\}) \prod_x l_{N,x}(\theta_x) .$$

The *marginal posterior density* for the local model at $X = \xi$ is obtained from the joint posterior density by integrating out all θ_x for $x \neq \xi$

$$p_N(\theta_\xi) \propto \int \dots \int p_0(\{\theta_x\}) \prod_x l_{N,x}(\theta_x) \prod_{x \neq \xi} \mu(\mathrm{d}\theta_x) . \tag{16}$$

Multiple-Model Prior. It is the choice of the prior density $p_0(\{\theta_x\})$ in (16) that determines to which level information accumulated about $\theta_x,\ x \neq \xi$ is used to reduce the uncertainty of θ_ξ. Let us consider three basic options:

1. All the data is fitted with a single model.
2. The data is fitted separately for each condition x.
3. The selected data, for x close to a given ξ, is fitted with a local model.

Single Global Model. If we choose the prior density concentrated on a constant

$$p_0(\{\theta_x\}) \propto p_0(\theta) \prod_x \delta(\theta_x - \theta) ,$$

the posterior density combines all the data

$$\begin{aligned} p_N(\theta) &\propto p_0(\theta) \prod_x l_{N,x}(\theta)^{N_x} \\ &\propto p_0(\theta)\, l_N(\theta) . \end{aligned}$$

Multiple Strictly Local Models. If we choose the prior density in the product form

$$p_0(\{\theta_x\}) \propto \prod_x p_0(\theta_x) ,$$

the posterior density uses only the local data

$$p_N(\theta_\xi) \propto p_0(\theta_\xi)\, l_{N,\xi}(\theta_\xi) .$$

If there is no data for $X = \xi$, the posterior coincides with the prior

$$N_\xi = 0 \ \Rightarrow \ p_N(\theta_\xi) = p_0(\theta_\xi) ,$$

i.e., we do not learn any way from the other data available.

Statistically Dependent Models. Let us choose the prior density as a mixture

$$p_0(\{\theta_x\}) \propto \sum_{\xi} \pi(\xi)\, p_0(\{\theta_x\}|\xi)$$

where for each ξ the parameters $\{\theta_x : x \neq \xi\}$ are *conditionally independent* given θ_ξ

$$p_0(\{\theta_x\}|\xi) = p_0(\theta_\xi) \prod_{x\neq\xi} p_0(\theta_x|\theta_\xi)\,.$$

The posterior density of θ_ξ is then

$$p_N(\theta_\xi) \propto p_0(\theta_\xi)\, l_{N,\xi}(\theta_\xi) \prod_{x\neq\xi} \int p_0(\theta_x|\theta_\xi)\, l_{N,x}(\theta_x)\, \mu(\mathrm{d}\theta_x)\,. \tag{17}$$

Let us analyze the last option in more detail.

Cross-Model Likelihood. In the formula (17), information available about θ_x for $x \neq \xi$ affects estimation of θ_ξ through the cross-model likelihood factor

$$l_{N,x}(\theta_\xi) = \int p_0(\theta_x|\theta_\xi)\, l_{N,x}(\theta_x)\, \mu(\mathrm{d}\theta_x)\,. \tag{18}$$

Consider two extreme instances of cross-model dependence:

A. θ_x is identical (coincides) with θ_ξ,

$$p_0^{\mathbf{A}}(\theta_x|\theta_\xi) = \delta(\theta_x - \theta_\xi) \;\Rightarrow\; l_{N,x}^{\mathbf{A}}(\theta_\xi) = l_{N,x}(\theta_\xi)\,.$$

B. θ_x is independent of θ_ξ,

$$p_0^{\mathbf{B}}(\theta_x|\theta_\xi) = p_0(\theta_x) \;\Rightarrow\; l_{N,x}^{\mathbf{B}}(\theta_\xi) = \mathrm{const}.$$

Now, rather than calculating (18) directly, we can approximate it by smoothing between the extremes $\mathbf{A}$ and $\mathbf{B}$

$$l_{N,x}^{w}(\theta_\xi) = c' \left[l_{N,x}^{\mathbf{A}}(\theta_\xi)\right]^{w(x,\xi)} \left[l_{N,x}^{\mathbf{B}}(\theta_\xi)\right]^{1-w(x,\xi)}$$

with respective weights

$$w(x,\xi) \;\text{ and }\; 1 - w(x,\xi)$$

satisfying

$$0 \leq w(x,\xi) \leq w(\xi,\xi) = 1\,.$$

Since $l_{N,x}^{\mathbf{B}}(\theta_\xi)$ is a constant independent of θ_ξ, the cross-model likelihood is approximated through

$$l_{N,x}^{w}(\theta_\xi) = c \left[l_{N,x}(\theta_\xi)\right]^{w(x,\xi)}\,. \tag{19}$$

Example. Consider the linear normal regression model with the local likelihood in the form

$$l_{N,x}(\theta_x) = c_1 \, \exp\left\{ -\frac{1}{2} \, (\theta_x - \hat{\theta}_x)^T P_x^{-1} (\theta_x - \hat{\theta}_x) \right\} .$$

Let us define the dependence of θ_x on θ_ξ explicitly via the stochastic equation

$$\theta_x = \theta_\xi + v, \;\; v \sim N(0, Q) \, .$$

The cross-model likelihood (18) results after some algebraic manipulations

$$\int p_0(\theta_x | \theta_\xi) \, l_{N,x}(\theta_x) \, \mu(\mathrm{d}\theta_x) = c_2 \, \exp\left\{ -\frac{1}{2} \, (\theta_\xi - \hat{\theta}_x)^T (P_x + Q)^{-1} (\theta_\xi - \hat{\theta}_x) \right\} .$$

Compare the result with the approximate expression (19)

$$l^w_{N,x}(\theta_\xi) = c_3 \, [l_{N,x}(\theta_\xi)]^w = c_4 \, \exp\left\{ -\frac{1}{2} \, (\theta_\xi - \hat{\theta}_x)^T w \, P_x^{-1} (\theta_\xi - \hat{\theta}_x) \right\} .$$

Both Q and w depend here on (x, ξ).

The resulting formulae for the update of the covariance matrix of θ_ξ are related similarly as the Kalman filter-like *linear forgetting*

$$P_x + Q, \quad Q > 0$$

and *exponential forgetting*

$$\frac{1}{w} \, P_x, \quad 0 \leq w \leq 1 \, .$$

(cf. [16, 15]).

Continuous and Discrete Predictors. When dealing with continuous predictor variables, a popular approach is to define the weights $w(x, \xi)$ via a suitable *kernel function*

$$W_h(x, \xi)$$

the shape of which can be fine-tuned by bandwidth parameters h. The weight on the model at x relative to ξ thus depends on the Euclidean distance of x from ξ. To put it other way, one makes us of the topology of the predictor space to infer on the similarity of respective models.

This approach does not work for discrete (categorical) predictors. Consider, e.g., days of week. Lacking any natural embedding of the respective values into a Euclidean space, the weight such as

$$w(\mathrm{Tuesday}, \mathrm{Friday})$$

can be derived only from the "strength" of statistical dependence of the respective models expressed through the density

$$p_0(\theta_{\mathrm{Tuesday}} | \theta_{\mathrm{Friday}}) \, .$$

Posterior Density. After substituting the approximate expression (19) of the cross-model likelihoods (18) in the posterior density (17), we obtain a locally weighted formula

$$\begin{aligned} p_N(\theta_\xi) &\propto p_0(\theta_\xi)\, l_{N,\xi}(\theta_\xi) \prod_{x\neq\xi} [l_{N,x}(\theta_\xi)]^{w(x,\xi)} \\ &\propto p_0(\theta_\xi) \prod_{x\in\mathcal{X}_N} [l_{N,x}(\theta_\xi)]^{w(x,\xi)} \\ &\propto p_0(\theta_\xi) \prod_{k=1}^{N} \big[s_{\theta_\xi}(y_k|x_k)\big]^{w(x_k,\xi)} \ .) \end{aligned} \tag{20}$$

8 Locally Weighted Geometry

The locally weighted Bayesian estimation developed in the previous section can be given an intuitive geometric interpretation again.

Empirical Density. For a given sample $y^N,\ x^N$ and a fixed "query" point ξ, we define the *effective number of data* and the *marginal empirical density* of X as

$$\begin{aligned} N_\xi^w &= \sum_{k=1}^{N} w(x_k,\xi)\,, \\ r_{N_\xi^w}(x) &= \frac{1}{N_\xi^w} \sum_{k=1}^{N} w(x_k,\xi)\,\delta(x-x_k)\,. \end{aligned}$$

The *joint empirical density* of (Y,X) combines the original conditional density $r_N(y|x)$ and the weighted marginal density $r_{N_\xi^w}(x)$

$$r_{N_\xi^w}(y,x) = r_N(y|x)\, r_{N_\xi^w}(x)\,.$$

Kerridge Inaccuracy. The empirical expectation of the inaccuracy of conditional densities yields the *conditional Kerridge inaccuracy*

$$\begin{aligned} K(r_{N_\xi^w} : s_\theta) &= E_{N_\xi^w}\, K(r_{N,x} : s_{\theta,x}) \\ &= \int r_{N_\xi^w}(x)\, K(r_{N,x} : s_{\theta,x})\, \mu(\mathrm{d}x) \\ &= \frac{1}{N_\xi^w} \sum_{x\in\mathcal{X}_N} w(x,\xi)\, N_x\, K(r_{N,x} : s_{\theta,x}) \\ &= \frac{1}{N_\xi^w} \sum_{k=1}^{N} w(x_k,\xi)\, K(r_{N,x_k} : s_{\theta,x_k})\,. \end{aligned}$$

Likelihood Function. The resulting likelihood function can be rewritten as

$$\begin{aligned}\prod_{k=1}^{N} s_\theta(y_k|x_k)^{w(x_k,\xi)} &= \exp\Big(-\sum_{k=1}^{N} w(x_k,\xi)\,\log\frac{1}{s_\theta(y_k|x_k)}\Big)\\ &= \exp\Big(-N_\xi^w \iint r_{N_\xi^w}(y,x)\,\log\frac{1}{s_\theta(y|x)}\,\mu(\mathrm{d}y)\,\mu(\mathrm{d}x)\Big)\\ &= \exp\big(-N_\xi^w\,K(r_{N_\xi^w}:s_\theta)\big)\\ &= \exp\big(-N_\xi^w\,E_{N_\xi^w}\,K(r_{N,X}:s_{\theta,X})\big)\end{aligned}$$

Posterior Density. The posterior density (20) can thus be expressed in the following compact form

$$p_N(\theta) \propto p_0(\theta)\,\exp\Big(-N_\xi^w\,K(r_{N_\xi^w}:s_\theta)\Big)\ .$$

Where appropriate, the prior density $p_0(\theta)$ can be chosen in the generalized conjugate form (12).

Big Picture. In locally weighted estimation, we deliberately modify the empirical density of the observed cases so as to use information learnt about θ_ξ *only* at cases x "relevant" or "similar" to a given case ξ. The weighting in the empirical density results in the same weighting of conditional Kerridge inaccuracy. The practical effect of relevance weighting is in making the model $s_{\hat\theta_N^w}$ fit better the local data at cases x "close to" ξ (cf. Fig. 5).

The weight put on the case x reflect the level of statistical dependence of θ_x on θ_ξ. "If I knew θ_ξ, how much would that affect my knowledge of θ_x at $x \neq \xi$?" The perfect dependence (identity) implies weight 1 while independence means weight 0.

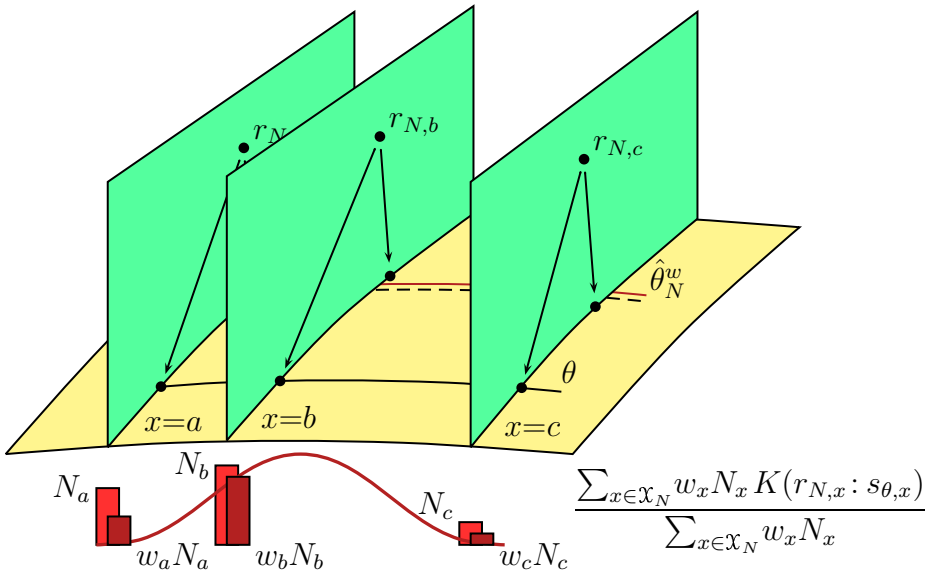

Figure 5: The goodness of fit in locally weighted Bayesian inference is given by the weighted empirical mean of Kerridge inaccuracy of case-based empirical densities relative to the corresponding model densities.

9 Concluding Remarks

The paper has analyzed local modeling using the information geometry of Bayesian inference for conditional probabilities.

Bayesian Smoothing

- We have shown that nonparametric regression can be derived as a special case of Bayesian inference over a set of local, case-based models with a properly chosen prior density linking the models.
- The result can be approximated as a weighted Bayesian inference with weights on the local likelihoods being proportional to the "strength" of statistical cross-model dependence.
- The practical advantage of the chosen approximation is that one needs to retrieve from the history only data that are assigned positive weights.

Why Information Geometry?

- We can view statistical inference as approximation of the empirical density rather than estimation of a hypothetical "true" density.
- Kerridge inaccuracy provides us with a generalized empirical error, which changes consistently with the underlying model family.
- We can elicit prior knowledge via the virtual data and capture the knowledge in the prior density of data using rigorous statistical methods.
- We can fine-tune the model by analyzing the orthogonal projection "trace" of conditional empirical (possibly smoothened) distributions onto the model manifold.
- The resulting "big picture" provides a natural departure point for design of approximations to the optimal but intractable solutions.

Acknowledgments

The author's research has been supported in part by the Grant Agency of the Czech Republic through Grant 102/01/0021. The support is gratefully acknowledged.

References

[1] D.W. Aha, D. Kibler and M.K. Albert, Instance-based learning algorithms, Machine Learning **6** (1991) 37–66.

[2] S. Amari, Differential-Geometrical Methods in Statistics, Vol. 28 of Lecture Notes in Statistics, Springer-Verlag, Berlin (1985).

[3] C.G. Atkeson, S.A. Schaal and A.W. Moore, Locally weighted learning, AI Review **11** (1997) 11–73.

[4] L. Bottou and V. Vapnik, Local learning algorithms, Neural Computation **4** (1992) 888–900.

[5] N.N. Chentsov, Statistical Decision Rules and Optimal Inference (in Russian), Nauka, Moscow (1972). English translation in Translations of Mathematical Monographs 53, Amer. Math. Soc., Providence, RI (1982).

[6] W.S. Cleveland, Robust locally weighted regression and smoothing scatterplots, J. Amer. Statist. Assoc. **74** (1979) 829–836.

[7] W.S. Cleveland and S.J. Devlin. Locally-weighted regression: an approach to regression analysis by local fitting, J. Amer. Statist. Assoc. **83** (1988) 596–610.

[8] I. Csiszár, *I*-divergence geometry of probability distributions and minimization problems, Ann. Probab. **3** (1975) 146–158.

[9] F. Cucker and S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. **39** (2002) 1–49.

[10] G. Cybenko, Just-in-time learning and estimation, in S. Bittanti and G. Picci (eds.), Identification, Adaptation, Learning, 423–434, NATO ASI Series, Springer-Verlag (1996).

[11] W. Härdle, Applied Non-parametric Regression, Cambridge University Press (1990).

[12] D.F. Kerridge, Inaccuracy and inference, J. Roy. Statist. Soc. Ser. B **23** (1961) 284–294.

[13] R. Kulhavý, Recursive Nonlinear Estimation: A Geometric Approach. Vol. 216 of Lecture Notes in Control and Information Science, Springer-Verlag, London (1996).

[14] R. Kulhavý and P. Ivanova, Memory-based prediction in control and optimisation, in Proc. 14th World Congress of IFAC, Beijing, PRC (1999) Vol. H, pp. 289–294.

[15] R. Kulhavý and F.J. Kraus, On duality of regularized exponential and linear forgetting, Automatica **32** (1996) 1403–1415.

[16] R. Kulhavý and M.B. Zarrop, On a general concept of forgetting, Int. J. Control **58** (1993) 905–924.

[17] S. Kullback and R.A. Leibler, On information and sufficiency, Ann. Math. Statist. **22** (1951) 79–86.

[18] V. Peterka, Bayesian approach to system identification, in P. Eykhoff (Ed.), Trends and Progress in System Identification. Pergamon Press, Elsmford, NY (1981), pp. 239—304

[19] C.E. Shannon, A mathematical theory of communication, Bell System Tech. J. **26** (1948) 379–423, 623–656.

[20] A. Stenman, Model on demand: algorithms, analysis and applications, Ph.D. Thesis No. 571, Dept. of EE, Linköping University (1999).

[21] D.M. Titterington, Common structure of smoothing techniques in statistics. Intern. Stat. Review **53** (1985) 141–170.

[22] V.N. Vapnik, The Nature of Statistical Learning, Springer, New York (1995).

[23] H. Zhu and R. Rohwer, Information geometry, Bayesian inference, ideal estimates and error decomposition, Technical Report No. 98-06-045, Santa Fe Institute (1998).