Local Factorization of Functions

Technical Report

June, 2001 (Revised October, 2003)

Volker Tresp
Siemens Corporate Technology
Department of Information and Communications
81730 Munich, Germany
volker.tresp@siemens.com
and
Anton Schwaighofer
Graz University of Technology
Institute for Theoretical Computer Science, Austria
anton.schwaighofer@gmx.net

Abstract

This paper is concerned with the notion of a local factorization of a function where we are mostly interested in the special case that this function is a probability distribution. We introduce the notions of local independence and of the local Kullback-Leibler divergence. We introduce a specific approximate local factorization. The number of terms required in the approximation is linear in the number of input dimensions and the approximation does not require the calculation of higher derivatives (as in a Taylor expansion) and is not limited to approximations near the mode of a function. We provide examples where we believe the approximation might be useful as in the approximate calculation of certain integrals.

1 Introduction

This paper is concerned with the notion of a local factorization of a function where we are mostly interested in the special case that this function is a probability distribution.

Probability distributions which factorize locally display independence in a local neighborhood. We introduce the notions of local independence and of the local Kullback-Leibler divergence. We introduce a specific approximate local factorization, the *local factorizing approximation* (LFA). In contrast to other local approximations, such as the Laplace approximation, the LFA does not rely on an approximation in the neighborhood of the mode of the function. Also, in contrast to a Laplace approximation, the LFA is accurate in case that the function factorizes in some rotated coordinate system.

In the next section we introduce the notion of local independence, in Section 3 we define the local Kullback-Leibler divergence and in Section 4 we discuss the optimal local factorization. In Section 5 we will define the LFA and discuss its properties. Readers only interested in the algorithm should directly go to this section. In Section 6 we will outline potential applications. In Section 7 we supply conclusions.

2 Local Factorizability and Local Independence

If random variables are globally independent their probability distribution factorizes and vice versa. Here we study the case that variables might be locally independent and might factorize locally.

Let

$R_a^\delta = \{x : |x_i - a_i| < \delta_i, i = 1, \ldots, N\}$

be a rectangle, with $x, a, \delta \in \Re^N$, $\delta_i > 0$.

Definition 1 *The local distance between two functions $g(x)$ and $h(x)$ in a rectangle R_a^δ is defined as*

$\max_x |g(x) - h(x)| < \epsilon \quad \forall x \in R_a^\delta.$

Here, $x, a, \delta \in \Re^N$, $\delta_i > 0$.

Definition 2 *Two functions $g(x)$ and $h(x)$ are locally identical in a rectangle R_a^δ if their local distance is equal to zero.*

Definition 3 *Let $g(x)$ be a function and let*

$h(x) = \prod_{i=1}^{N} h_i(x_i), \quad \forall x \in R_a^\delta.$

If there are functions $h_i(x_i)$ such that $g(x)$ and $h(x)$ are locally identical in a rectangle R_a^δ, then $g(x)$ is locally factorizable in the rectangle R_a^δ.

Definition 4 *If $P(x)$ is a probability distribution which is locally factorizable in R_a^δ, we say that the variables $x_1, \ldots, x_N$ are locally independent in R_a^δ.*

Proposition 1 *A function which factorizes globally also factorizes locally.*

The converse is not true in general.

3 Local Kullback-Leibler (KL) Divergence

Here we introduce the local Kullback-Leibler (KL) divergence for functions which are strictly positive in a rectangle R_a^δ such that, after normalization, they can be treated as probability distributions.

Definition 5 *Let $P(x)$ and $Q(x)$ be two functions which are strictly positive within the rectangle R_a^δ. The local Kullback-Leibler (KL) divergence between $P(x)$ and $Q(x)$ with respect to the rectangle R_a^δ is defined as*

$$lKL_{R_a^\delta}(P(x)||Q(x)) = \frac{1}{Z^P_{R_a^\delta}} \int_{R_a^\delta} P(x) \log \left(\frac{P(x)}{Q(x)} \frac{Z^Q_{R_a^\delta}}{Z^P_{R_a^\delta}} \right) dx \tag{1}$$

where $Z^P_{R_a^\delta} = \int_{R_a^\delta} P(x)dx$ and $Z^Q_{R_a^\delta} = \int_{R_a^\delta} Q(x)dx$.

Essentially, we define the local KL-divergence with respect to the normalized local functions $P(x)/Z^P_{R_a^\delta}$ and $Q(x)/Z^Q_{R_a^\delta}$.

Proposition 2 *$lKL_{R_a^\delta}(P(x)||Q(x))$ is equal to zero if and only if*

$$Q(x)/Z^Q_{R_a^\delta} = P(x)/Z^P_{R_a^\delta}, \quad \forall x \in R_a^\delta, \quad Z^Q_{R_a^\delta} > 0, Z^P_{R_a^\delta} > 0.$$

This says that after normalization, both distributions must be identical. The proposition is analogue to the corresponding global property of probability distributions.

Proposition 3 *If two functions $P(x)$ and $Q(x)$ are strictly positive and identical in the rectangle R_a^δ, then*

$$lKL_{R_a^\delta}(P(x)||Q(x)) = 0.$$

This proposition follows from the fact that in this case the log() in Equation 1 is equal to zero.

Even if the local KL-divergence is zero, the distance between $P(x)$ and $Q(x)$ might still be large, since they are just equal up to a constant factor. That's why we need a definition for mass equivalence:

Definition 6 *Given two functions $P(x)$ and $Q(x)$ that are strictly positive in the rectangle R_a^δ, then they are* locally mass equivalent*, if $Z^P_{R_a^\delta} = Z^Q_{R_a^\delta}$.*

Proposition 4 *Given two functions $P(x)$ and $Q(x)$ which are strictly positive in the rectangle R_a^δ. If they are locally mass equivalent and if*

$$lKL_{R_a^\delta}(P(x)||Q(x)) = 0$$

then they are identical in the rectangle R_a^δ.

This proposition follows directly from Proposition 2.

4 Optimal Local Factorization

We now consider the optimal local factorization with respect to the local KL-divergence.

Let $Q(x)$ be an approximating function which is strictly positive in R_a^δ and let $Q(x)$ factorize locally with respect to the rectangle R_a^δ in the form

$$Q(x) = \prod_{i=1}^{N} q_i(x_i) \quad \forall x \in R_a^\delta.$$

Proposition 5 *Under the constraints that*

$$\int_{R_a^\delta} q_i(x_i) dx_i = 1, \quad q_i(x) \geq 0, \quad \forall i,$$

the local KL-divergence between $P(x)$ and the approximation $Q(x)$ is minimum with respect to the rectangle R_a^δ if

$$q_i(x_i) = \frac{1}{Z_{R_a^\delta}^P} \int_{R_a^\delta} P(x)\, dx \setminus x_i \quad \forall x \in R_a^\delta, \quad k_i > 0.$$

Furthermore

$$\tilde{Q}(x) = Z_{R_a^\delta}^P \prod_{i=1}^{N} q_i(x_i) \quad \forall x \in R_a^\delta.$$

is locally mass equivalent

This proposition follows from the corresponding theorem for global distributions.

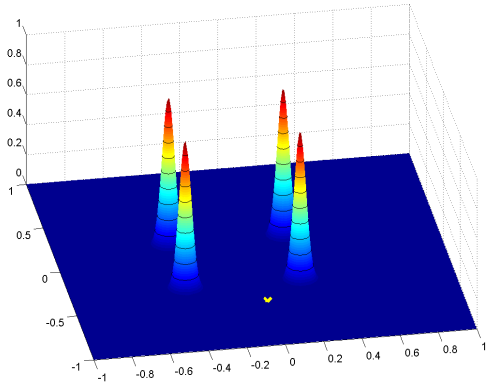
5 The Local Factorizing Approximation

As stated in Proposition 5, the minimum local KL-divergence can be achieved by a local marginalization but this requires the calculation of N-dimensional integrals. In general, closed-form solutions to those integrals will not exist and those integrals would have to be solved numerically.

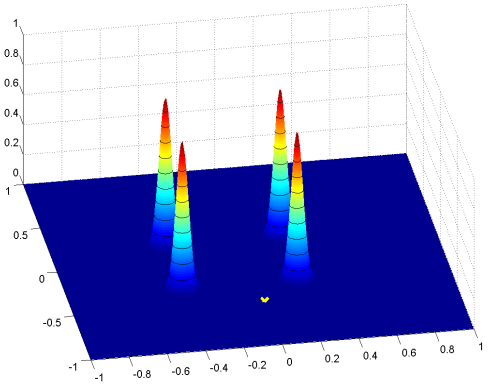
In this section we want to study the properties of a local factorization of a function which can easily be calculated.

5.1 Definition

We start with a definition of a local factorizing approximation.



(a) A factorizing density producing four Gaussians in joint space.



(b) LFA: in this case, the approximation is exact. The white maker indicates the position of a which is in a region with low probability.

Figure 1: Approximating a Factorizing Distribution

Definition 7 ***(Local Factorizing Approximation)*** *Given a function $g(x)$ with $x \in \Re^N$. The local factorizing approximation (LFA) approximates the function by a product of one-dimensional functions in the neighborhood of $x = a$ as*

$$LFA_a(g(x)) = \frac{1}{g(a)^{N-1}} \prod_{l=1}^{N} g(\tilde{a}_{-l}) \tag{2}$$

where $\tilde{a}_{-l}$ is equal to a except that the l-th component is replaced by x_l.

Note that $g(\tilde{a}_{-l})$ is a 1-D function of a line parallel to x_l and going through a.

Figures 1, 2 and 3 illustrate the approximation.

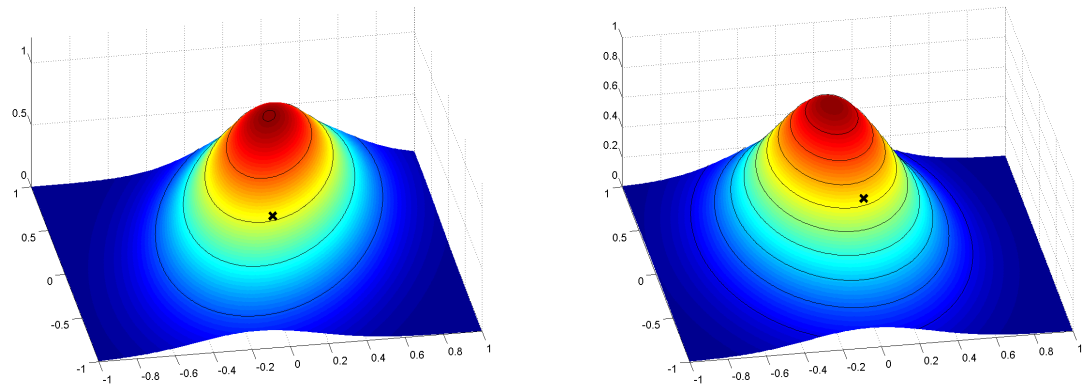
5.2 LFA for Factorizing Functions

A factorizing function can be written as

$$g(x) = \prod_{i=1}^{N} h_i(x_i).$$

If we substitute into the LFA

$$g(x) \approx \frac{1}{g(a)^{N-1}} \prod_{l=1}^{N} g(a) \frac{h_l(x_l)}{h_l(a_l)}$$



(a) A Gaussians with full covariance matrix.

(b) LFA. The maker indicates the position of a

Figure 2: Approximating a fistribution which does not factorize.

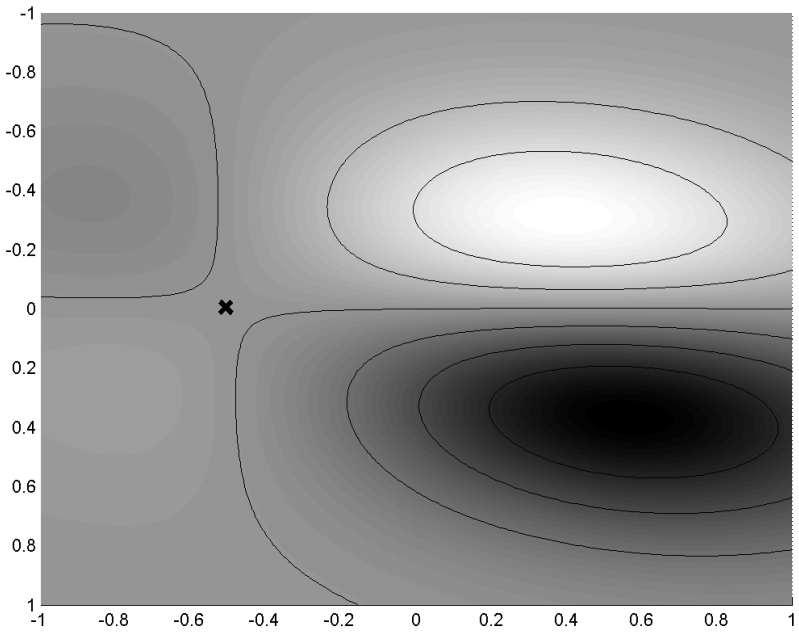


Figure 3: The difference between the true Gaussian and the LFA approximation from the previous figure. Note that near a and along axis-parallel lines through a, the approximation is exact.

$$= g(a) \prod_{l=1}^{N} \frac{h_l(x_l)}{h_l(a_l)} = \prod_{i=1}^{N} h_i(x_i) = g(x).$$

Proposition 6 *This result shows that for functions which factorize the LFA is exact.*

Note, that the factorization of a function is not unique. For, example, if we set

$$h_i(x) \rightarrow A h_i(x) \qquad h_j(x) \rightarrow \frac{1}{A} h_j(x)$$

for one $j \neq i$ we obtain another a valid factorization. On the other hand the LFA is unique.

5.3 Relationship to Taylor Expansion

Here we motivate that LFA is also a reasonable approximation if the function $g(x)$ does not factorize. Here we assume that $g(x) > 0$. Let $G(x) = \log g(x)$. For notational simplicity we assume that x is two-dimensional. Using a Taylor expansion one may write

$$G(x_1, x_2) = G(a_1, a_2) + \tag{3}$$

$$\sum_{j=1}^{\infty} \frac{1}{j!} \left[(x_1 - a_1) \frac{\partial}{\partial x_1'} + (x_2 - a_2) \frac{\partial}{\partial x_2'} \right]^j G(x_1', x_2') \Big|_{x_1'=a_1, x_2'=a_2} . \tag{4}$$

If we now *make the approximation that all interaction-terms are equal to zero*,

$$\frac{\partial^j}{\partial {x_1'}^j} \frac{\partial^i}{\partial {x_2'}^i} G(x_1', x_2') \Big|_{x_1'=a_1, x_2'=a_2} = 0 \quad \forall j, i > 0 \tag{5}$$

then we obtain

$$G(x_1, x_2) = G(a_1, a_2) + \tag{6}$$

$$\sum_{j=1}^{\infty} \frac{1}{j!} \left[(x_1 - a_1) \frac{\partial^j}{\partial (x_1')^j} + (x_2 - b_2) \frac{\partial^j}{\partial (x_2')^j} \right] G(x_1', x_2') \Big|_{x_1'=a_1, x_2'=a_2} \tag{7}$$

$$= G(x_1, a_1) + G(a_1, x_2) - G(a_1, a_2).$$

Now we set

$$g(x_1, x_2) = \exp(G(x_1, x_2)) \approx \frac{1}{g(a_1, a_2)} g(x_1, a_1) g(a_1, x_2) \tag{8}$$

which is the LFA for the 2-D case.

This result generalizes to the higher-dimensional case. The relationship to the Taylor expansion clarifies the approximation which is being made in the LFA.

5.4 Approximation along Axis Through a

Let's define $\tilde{a}_{-k}$ as before.

Proposition 7 *On points on $\tilde{a}_{-k}$, the LFA is exact,*

since

$$g(\tilde{a}_{-k}) \approx \frac{1}{g(a)^{N-1}} \prod_{l=1}^{N} g(\tilde{a}_{-l}) = \frac{1}{g(a)^{N-1}} g(a)^{N-1} g(\tilde{a}_{-k}) = g(\tilde{a}_{-k}).$$

5.5 Additional Favorable Properties of the LFA

The approximation is easy to calculate and only requires the evaluation of $g(x)$. The number of terms is linear in the dimension of x. The approximation is well-behaved, if $g(x)$ is well-behaved since we multiply functional values of $g(x)$.

Also we have,

Proposition 8 *For a function which factorizes globally the LFA is locally exact.*

5.6 The LFA for Positive Functions

We want to consider one more favorable property of the LFA:

Proposition 9 *Consider the situation in Proposition 5 and assume in addition that that $P(x)$ is a strictly positive* linear function *within R_a^δ. Then, the LFA minimizes the local KL-divergence and is also locally mass equivalent.*

Proof: Without loss of generality, let

$$P(x) = c_0 + \sum_{i=1}^{N} c_i(x_i - a_i) \quad \forall x \in R_a^\delta.$$

Then

$$q_i(x_i) = \frac{1}{2c_0\delta_i} \left(c_0 + c_i(x_i - a_i)\right)$$

and, with $Z_{R_a^\delta}^P = c_0 2^N \prod_{i=1}^N \delta_i$,

$$\tilde{Q}(x) = Z_{R_a^\delta}^P \prod_{i=1}^{N} q_i(x_i) = \frac{1}{c_0^{N-1}} \prod_{l=1}^{N} (c_0 + c_i(x_i - a_i)).$$

Thus,

$$Q(x) = LFA_a\left(P(x)\right).$$

6 Applications

6.1 Normalized RBF

In some applications, one needs to calculate normalized radial basis functions (rbf) of the form

$$\mathrm{rbf}_i(x) = \frac{\exp(-\frac{1}{2\sigma^2}(x - {c^{(i)}}^2))}{\sum_{j=1}^{M} \exp(-\frac{1}{2\sigma^2}(x - {c^{(j)}}^2))}.$$

If we apply the LFA to the denominator around $a = c_i$, we obtain

$$\mathrm{rbf}_i(x) \approx \exp\left(-\frac{1}{2\sigma^2}(x - c^{(i)})^2\right) \frac{n(c_i)^{N-1}}{\prod_{l=1}^{N} n(\tilde{c}_{-l}^{(i)})} = n(c_i)^{N-1} \prod_{l=1}^{N} \frac{\exp(-\frac{1}{2\sigma^2}(x_l - {c_l^{(i)}}^2))}{n(\tilde{c}_{-l}^{(i)})}$$

where

$$n(x) = \sum_{j=1}^{M} \exp(-\frac{1}{2\sigma^2}(x - {c^{(j)}}^2)).$$

We obtain a product of one-dimensional functions which are easy to evaluate.

6.2 Improving the Laplace Approximation

Consider an integral of the form

$$P(D) = \int P(D|w)P(w)dw$$

where D denotes observed data and where w are model parameters. Such integrals need to be evaluated for calculating the evidence of a model (Heckerman, 1999). The Laplace approximation gives

$$P(D) \approx P(D|w_{MAP})P(w_{MAP})(2\pi)^{-d/2}|A|^{-1/2}$$

with

$$w_{MAP} = \arg\max_w(P(D|w)P(w)) \quad A = -\frac{\partial^2}{\partial w^2} \log P(D|w)P(w)\bigg|_{w=w_{MAP}}.$$

Let's now consider the coordinate system, in which A is diagonal. Here, the Laplace approximation corresponds to a Taylor expansion in $\log P(D|w)P(w)$ where only quadratic non-interacting terms are considered. We can now perform the LFA in this new coordinate system which maintains all higher non-interacting terms in the Taylor expansion.

The LFA applied in the new coordinate system leads to

$$P(D) \approx \frac{1}{P(w_{MAP})^{N-1}} \prod_{l=1}^{N} \int g(\tilde{w}_{MAP,-l}) dw_l.$$

The LFA is exact, if $P(D|w)P(w)$ factorizes in the rotated coordinate system. The 1-D integrals can be solved numerically.

As a example, consider that

$$P(D|w)P(w) = B\,\mathcal{N}(w; w_{MAP}, \Sigma) + \frac{C}{V} \mathrm{b}(w - w_{MAP})$$

where $\mathcal{N}(w; w_{MAP}, \Sigma)$ is a normal density of w, evaluated at w_{MAP} with covariance Σ. Furthermore, $b(x)$ is equal to one if $|x| < 1$ and zero elsewhere, B and C are constants and $V = \int \mathrm{b}(w - w_{MAP}) dw$.

Then, $\int P(D|w)P(w)dw = B + C$ which is the result the LFA would also give if we apply the LFA in the coordinate system in which Σ is diagonal. The Laplace approximation would provide the incorrect result

$$B + \frac{C}{V}(2\pi)^{d/2} |\Sigma|^{-1/2}.$$

7 Conclusions

We have introduced the notions of a local factorization, a local independence and of the local Kullback-Leibler divergence. We have introduced the local factorizing approximation (LFA) and have highlighted a number of potential applications. In addition to the evidence, the LFA can also be used to calculate an approximation to the entropy of a distribution and might also find applications in variational approximations (see, for example, Jaakkola, 2000).

References

Heckerman, D. (1999). A Tutorial on Learning with Bayesian Networks. In Learning in Graphical Models, M. Jordan, ed.. MIT Press, Cambridge, MA.

Jaakkola, T. (2000). Tutorial on variational approximation methods. In Advanced mean field methods: theory and practice. MIT Press, 2000.