Bayesian Field Theory
Nonparametric Approaches to Density Estimation

Jörg C. Lemm

Institut für Theoretische Physik I, Universität Münster
Wilhelm–Klemm–Str.9, D–48149 Münster, Germany
lemm@uni-muenster.de
http://pauli.uni-muenster.de/~lemm/

Abstract

Nonparametric approaches to density estimation are discussed from a Bayesian perspective. Being in general non–Gaussian the resulting models have to be solved by discretization. A numerical example shows that this can be computationally feasible for low–dimensional problems.

1 Basic definitions

A Bayesian approach to empirical learning problems requires two main inputs: 1. a (conditional) *likelihood model* $p(y|x,\phi)$ representing the probability of measuring y given (visible) condition x and (hidden) state of Nature ϕ. 2. a *prior model* $p_0(\phi)$ encoding the *a priori* information available on ϕ. We assume that the joint probability factorizes according to

$$p(x,y,\phi) = p(y|x,\phi)\, p(x)\, p(\phi). \tag{1}$$

Furthermore, we will consider n independent (training) data $D = \{(x_i, y_i)|1 \le i \le n\}$. Denoting by y_D (x_D) the vector of the y_i (x_i), we assume $p(y_D|x_D,\phi) = \prod_{i=1}^n p(y_i|x_i,\phi)$ and $p(x) = \prod_{i=1}^n p(x_i)$. Different classes of likelihood functions define different problem types. In the following we will be interested in general density estimation problems, the other problems listed in Tab. 1 being special cases thereof.

Within *nonparametric* approaches, the function values $P(x,y) = p(y|x,\phi)$ itself, specifying a state of Nature ϕ, are considered as the primary degrees of freedom, their only constraint being the positivity and normalization conditions for $p(y|x,\phi)$. In particular, we will study in this paper likelihoods P being functionals of *fields* $\phi(x,y)$, i.e., $P = P(\phi)$. Examples of such functionals $P(\phi)$

(conditional) likelihood $p(y\|x,\phi)$	problem type
of general form	density estimation
discrete y	classification
$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{[y-\phi(x)]^2}{2\sigma^2}}$ (Gaussian, σ fixed)	regression
$\sum_k \frac{1}{\sigma_k\sqrt{2\pi}} e^{-\frac{[y-\phi_k(x)]^2}{2\sigma_k^2}}$ (mixture of Gaussians)	clustering
$\mathrm{Trace}(\rho(\phi)\,\pi_{(y,x)})$	inverse quantum mechanics

Table 1: Special cases of density estimation. For classification and regression see [7, 6, 1, 2] and references therein, for clustering see [5], and for Bayesian approaches to inverse quantum mechanics, aiming for example in the reconstruction of potentials from observational data, see [3]. (ρ denotes the density operator of a quantum system, $\pi_{(y,x)}$ the projector to the eigenstate of observable x, represented by an hermitian operator, with eigenvalue y.)

are shown in Tab. 2. While the first and third row show the 'local' functionals $P(\phi)$ corresponding to $\phi(x,y) = P(x,y)$ and $\phi(x,y) = \ln P(x,y)$, the other functionals are nonlocal in y.

Depending on the choice of the functional $P(\phi)$ the normalization and positivity constraint for P take different forms for ϕ. For example, the normalization condition is automatically fulfilled in rows two and four of Tab. 2. In the last row, where $\phi(x,y) = \int_{-\infty}^{y} dy'\, P(x,y')$, the normalization constraint for P becomes the boundary condition $\phi(x,1) = 1$, and positivity of P means monotony of $\phi(x,\cdot)$.

2 Maximum A Posteriori Approximation

In this paper, our aim will be to reconstruct the field ϕ from observational data D in *Maximum A Posteriori Approximation* (MAP), i.e, by maximizing

$$p(\phi|D) \propto p(y_D|x_D,\phi)\, p_0(\phi), \tag{2}$$

with respect to ϕ. Implementing the normalization conditions on P (for all x) by a Lagrange term with multiplier function $\Lambda_X(x)$, introducing an additional Gaussian process prior term for ϕ (with zero mean and inverse covariance $\mathbf{K}$, real symmetric, positive definite), and considering the usual case where the positivity constraint is not active at the stationary point (or automatically satisfied by the choice of ϕ) so a corresponding Lagrange multiplier can be skipped, we have

$P(\phi)$	constraints	
$P(x,y) = \phi(x,y)$	norm	positivity
$P(x,y) = \frac{\phi(x,y)}{\int \phi(x,y')\,dy'}$	—	positivity
$P(x,y) = e^{\phi(x,y)}$	norm	—
$P(x,y) = \frac{e^{\phi(x,y)}}{\int e^{\phi(x,y')}\,dy'}$	—	—
$P(x,y) = \frac{d\phi(x,y)}{dy}$	boundary	monotony

Table 2: Constraints for some specific choices of ϕ

therefore to minimize the error functional,

$$E_\phi = -<\ln P(\phi)\,|\,N> + \frac{1}{2}<\phi\,|\,\mathbf{K}\phi> + <P(\phi)\,|\,\Lambda_X>. \tag{3}$$

Here N denotes the empirical density, i.e., $N(x,y) = \sum_{i=1}^n \delta(x-x_i)\delta(y-y_i)$, and the bra–ket notation $<\cdot\,|\,\cdot>$, denoting scalar products, stands for integration over x and y. (Thus, the error functional (3) is, up to a ϕ–independent constant and the constraint implementation by Lagrange multiplier terms, the negative logarithm of the posterior (2). Prior terms can be made more flexible by including hyperparameters. See [2] for a more detailed discussion.) Notice that a Gaussian prior term in the field ϕ is in general non–Gaussian in terms of the conditional likelihood P.

The MAP equation to be solved is easily found by setting the functional derivative of (3) with respect to $\phi(x,y)$ to zero, yielding,

$$0 = \mathbf{P}'(\phi)\mathbf{P}^{-1}(\phi)N - \mathbf{K}\phi - \mathbf{P}'(\phi)\Lambda_X, \tag{4}$$

where $\mathbf{P}'(x,y;x',y') = \frac{\delta P(x',y')}{\delta\phi(x,y)}$ and $\mathbf{P}(x,y;x',y') = \delta(x-x')\delta(y-y')P(x,y)$. For existing $\mathbf{P}\mathbf{P}'^{-1}$ $=(\mathbf{P}'\mathbf{P}^{-1})^{-1}$ this can be written

$$0 = N - \mathbf{P}\mathbf{P}'^{-1}\mathbf{K}\,\phi - \mathbf{P}\Lambda_X. \tag{5}$$

From the normalization condition $\mathbf{I}_X P = I$ [where $\mathbf{I}_X(x,y;x',y') = \delta(x-x')$ and $I(x,y) = 1$], it follows for $\Lambda_X(x,y) = \Lambda_X(x) \ne 0$ that $\mathbf{I}_X\mathbf{P}\Lambda_X = \Lambda_X$, and thus, multiplying Eq. (5) by $\mathbf{I}_X$,

$$\Lambda_X = \mathbf{I}_X\left(N - \mathbf{P}\mathbf{P}'^{-1}\mathbf{K}\,\phi\right) = N_X - \mathbf{I}_X\mathbf{P}\mathbf{P}'^{-1}\mathbf{K}\,\phi, \tag{6}$$

where $N_X(x,y) = N_X(x) = \sum_i \delta(x-x_i)$. Λ_X can now be eliminated by inserting Eq. (6) into Eq. (5)

$$0 = (\mathbf{I} - \mathbf{P}\mathbf{I}_X)\left(N - \mathbf{P}\mathbf{P}'^{-1}\mathbf{K}\,\phi\right). \tag{7}$$

In contrast to regression with Gaussian process priors [7] (and similarly, classification for a specific choice of $P(\phi)$ [6]) the solution of Eq. (7) cannot be found in a finite dimensional space defined by the training data D. Thus, Eq. (7) has to be solved by discretization, analogously, for example, to solving field theories on a lattice [4]. Clearly, a direct discretization can only be useful for low dimensional x and y variables (say one– or two–dimensional, like time series or images). Due to increasing computing capabilities, however, many low–dimensional problems are now directly solvable by discretization [1]. For variables x or y living in higher dimensional spaces additional approximations are necessary [2].

3 A numerical example

The numerical feasibility of a solution of Eq. (7) has been tested for a density estimation problem with one–dimensional x and y variables and the choice $\phi(x,y) = \ln P(x,y)$ for which $\mathbf{P}' = \mathbf{P}$. Choosing a true P (shown in Fig. 1) 50 data points have been sampled (Fig. 2a). Including a Gaussian prior term for ϕ with zero mean and inverse covariance $-\frac{\partial^2}{\partial x^2} - \frac{1}{2}\frac{\partial^2}{\partial y^2}$ Eq. (7) has been solved on a mesh with 10 points in x–direction and 15 points in y–direction by standard iteration methods. (For technical details see [2].) The reconstructed conditional likelihood P is shown in Fig. 2b, true and reconstructed regression functions are compared in Fig. 3.

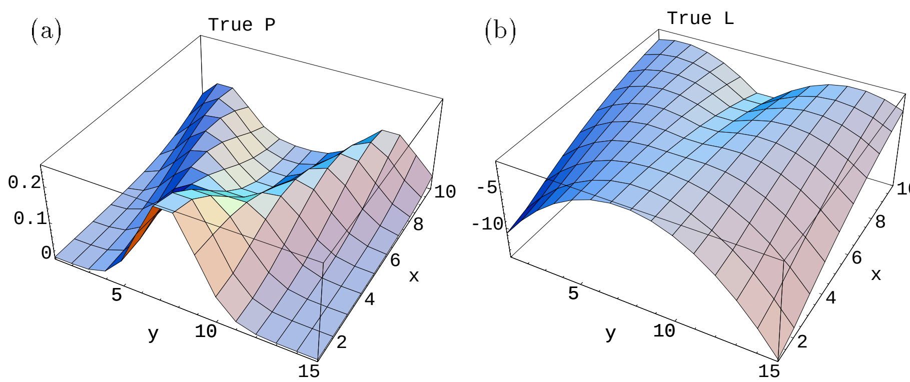

Figure 1: (a): True conditional likelihood $P_{\text{true}}(x,y) = p(y|x,\phi_{\text{true}})$. (b): Corresponding true field, here chosen as (conditional) log–likelihood $L = \ln P$, i.e., $\phi_{\text{true}}(x,y) = \ln P_{\text{true}}(x,y) = L_{\text{true}}(x,y)$.

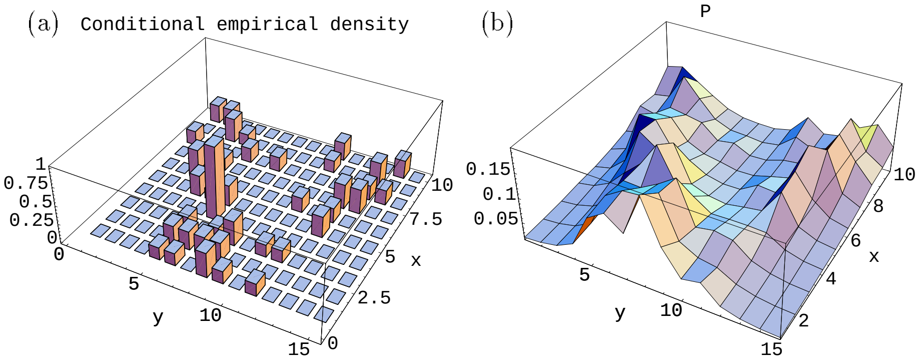

Figure 2: (a): Conditional empirical density $P_{\rm emp}$, defined as $P_{\rm emp}(x,y) = \sum_i \delta(x-x_i) \sum_i \delta(y-y_i) \sum_i / \sum_i \delta(x-x_i)$, as obtained from 50 data points sampled from the true likelihood $P_{\rm true}$ of Fig. 1a under uniform $p(x)$. (b): Conditional likelihood P reconstructed from the conditional empirical density shown on the l.h.s. using, besides the normalization constraint, an additional Gaussian prior term on ϕ (see text).

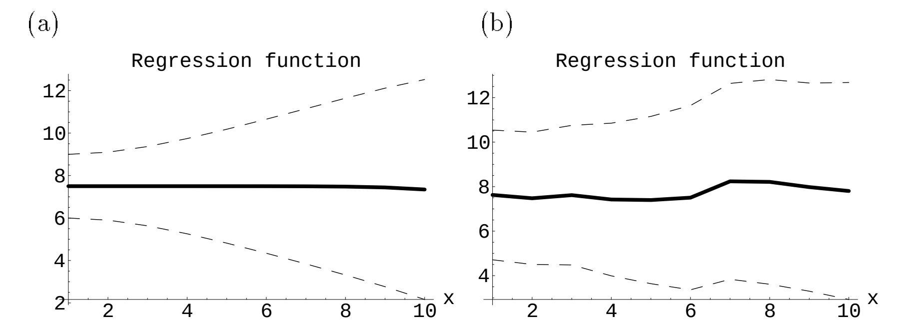

Figure 3: (a): Regression function for $P_{\rm true}$ of Fig. 1a. (b): Regression function for the reconstructed P of Fig. 2b. (The regression function for a field ϕ is defined as $h(x) = \int dy \, y \, p(y|x,\phi)$. The dashed lines indicate the range of one standard deviation above and below the regression function.)

4 Conclusions

Many nonparametric real world learning problems are generically non–Gaussian. Examples include general density estimation or the reconstruction of forces for quantum systems from observational data. Increasing computational resources allow to solve at least some of such low–dimensional non–Gaussian problems directly by discretization.

Acknowledgements

The author wants to thank J. Uhlig and A. Weiguny for stimulating discussions.

References

[1] Lemm, J.C. (1999) Mixtures of Gaussian Process Priors. In *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN99)*, IEEE Conference Publication No. 470. London: Institution of Electrical Engineers.

[2] Lemm, J.C. (1999) *Bayesian Field Theory. Nonparametric approaches to density estimation, regression, classification, and inverse quantum problems.* Technical Report, Münster University, MS-TP1-99-1. `arXiv: physics/9911077`.

[3] Lemm, J.C., Uhlig J.,& Weiguny, A. (1999) *A Bayesian Approach to Inverse Quantum Statistics.* Technical Report, MS-TP1-99-6, Münster University, (submitted to *Phys. Rev. Lett.*) `arXiv:cond-mat/9907013`.

[4] Montvay, I. & Münster, G. (1994) *Quantum Fields on a Lattice.* Cambridge: Cambridge University Press.

[5] Rose, K., Gurewitz, E., & Fox, G.C. (1990) Statistical mechanics and phase transitions in clustering. *Phys. Rev. Lett.* **65**, 945–948.

[6] Williams, C.K.I. & Barber, D. (1998) Bayesian Classification With Gaussian Processes *IEEE Trans. on Pattern Analysis and Machine Intelligence.* **20**(12), 1342–1351.

[7] Williams, C.K.I. & Rasmussen, C.E. (1996) Gaussian Processes for Regression. in *Advances in Neural Information Processing Systems* **8**, D.S. Touretzky et al (eds.), 515–520, Cambridge, MA: MIT Press.