Mixtures of Gaussian process priors*

Jörg C. Lemm
Institut für Theoretische Physik I, Universität Münster
D–48149 Münster, Germany
E-mail: lemm@uni-muenster.de
http://pauli.uni-muenster.de/~lemm
Publication No.: MS-TP1-99-5

Abstract

Nonparametric Bayesian approaches based on Gaussian processes have recently become popular in the empirical learning community. They encompass many classical methods of statistics, like Radial Basis Functions or various splines, and are technically convenient because Gaussian integrals can be calculated analytically. Restricting to Gaussian processes, however, forbids for example the implemention of genuine nonconcave priors. Mixtures of Gaussian process priors, on the other hand, allow the flexible implementation of complex and situation specific, also nonconcave *a priori* information. This is essential for tasks with, compared to their complexity, a small number of available training data. The paper concentrates on the formalism for Gaussian regression problems where prior mixture models provide a generalisation of classical quadratic, typically smoothness related, regularisation approaches being more flexible without having a much larger computational complexity.

Contents

1 Introduction

The generalisation behaviour of statistical learning algorithms relies essentially on the correctness of the implemented *a priori* information. While Gaussian processes and the related regularisation approaches have, on one hand, the very important advantage of being able to formulate *a priori*

*This is an extended version of a contribution to the Ninth International Conference on Artificial Neural Networks (ICANN 99), 7–10 September 1999, Edinburgh, UK.

information explicitly in terms of the function of interest (mainly in the form of smoothness priors which have a long tradition in density estimation and regression problems [18, 17, 5]) they implement, on the other hand, only simple concave prior densities corresponding to quadratic errors. Especially complex tasks would require typically more general prior densities. Choosing mixtures of Gaussian process priors combines the advantage of an explicit formulation of priors with the possibility of constructing general non-concave prior densities.

While mixtures of Gaussian processes are technically a relatively straightforward extension of Gaussian processes, which turns out to be a computational advantage, practically they are much more flexible and are able to produce in principle, i.e., in the limit of infinite number of components, any arbitrary prior density.

As example, consider an image completion task, where an image have to be completed, given a subset of pixels ('training data'). Simply requiring smoothness of grey level values would obviously not be sufficient if we expect, say, the image of a face. In that case the prior density should reflect that a face has specific constituents (e.g., eyes, mouth, nose) and relations (e.g., typical distances between eyes) which may appear in various variations (scaled, translated, deformed, varying lightening conditions).

While ways how prior mixtures can be used in such situations have already been outlined in [6, 7, 8, 9, 10] this paper concentrates on the general formalism and technical aspects of mixture models and aims in showing their computational feasibility. Sections 2–4 provide the necessary formulae while Section 5 exemplifies the approach for an image completion task.

Finally, we remark that mixtures of Gaussian process priors do usually *not* result in a (finite) mixture of Gaussians [3] for the function of interest. Indeed, in density estimation, for example, arbitrary densities not restricted to a (finite) mixture of Gaussians can be produced by a mixture of Gaussian prior processes.

2 The Bayesian model

Let us consider the following random variables:

1. x, representing (a vector of) *independent, visible variables* ('measurement situations'),
2. y, being (a vector of) *dependent, visible variables* ('measurement results'), and
3. h, being the *hidden variables* ('possible states of Nature').

A Bayesian approach is based on two model inputs [1, 11, 4, 12]:

1. A *likelihood model* $p(y|x,h)$, describing the density of observing y given x and h. Regarded as function of h, for fixed y and x, the density $p(y|x,h)$ is also known as the (x–conditional) *likelihood* of h.
2. A *prior model* $p(h|D_0)$, specifying the *a priori* density of h given some *a priori* information denoted by D_0 (but before training data D_T have been taken into account).

Furthermore, to decompose a possibly complicated *prior density* into simpler components, we introduce *continuous hyperparameters* θ and *discrete hyperparameters* j (extending the set of hidden variables to $\tilde{h} = (h, \theta, j)$),

$$p(h|D_0) = \int d\theta \sum_j p(h,\theta,j|D_0). \tag{1}$$

In the following, the summation over j will be treated exactly, while the θ–integral will be approximated. A Bayesian approach aims in calculating the *predictive density* for outcomes y in *test* situations x

$$p(y|x,D) = \int dh\, p(y|x,h)\, p(h|D), \tag{2}$$

given data $D = \{D_T, D_0\}$ consisting of *a priori* data D_0 and i.i.d. training data $D_T = \{(x_i, y_i)|1 \leq i \leq n\}$. The vector of all x_i (y_i) will be denoted x_T (y_T). Fig.1 shows a graphical representation of the considered probabilistic model.

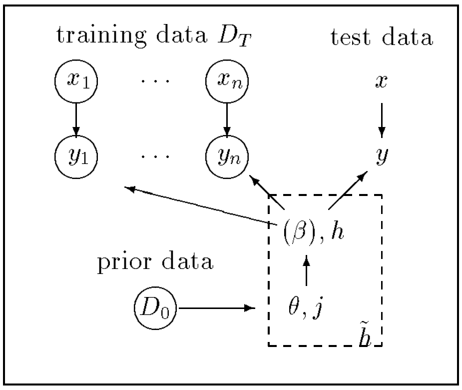

Figure 1: Graphical representation of the considered probabilistic model, factorising according to $p(x_T, y_T, x, y, h, \theta, j, (\beta)|D) = p(x_T)\ p(x)\ p(y_T|x_T, h, (\beta))\ p(y|x, h, (\beta))\ p(h|\theta, j, D_0, (\beta))\ p(\theta, j, (\beta)|D_0)$. (The variable β is introduced in Section 3.) Circles indicate visible variables.

In saddle point approximation (*maximum a posteriori approximation*) the h–integral becomes

$$p(y|x, D) \approx p(y|x, h^*), \tag{3}$$

$$h^* = \operatorname{argmax}_{h\in\mathcal{H}} p(h|D), \tag{4}$$

assuming $p(y|x, h)$ to be slowly varying at the stationary point. The *posterior density* is related to (x_T–conditional) likelihood and prior according to Bayes' theorem

$$p(h|D) = \frac{p(y_T|x_T, h)\, p(h|D_0)}{p(y_T|x_T, D_0)}, \tag{5}$$

where the h–independent denominator (evidence) can be skipped when maximising with respect to h. Treating the θ–integral within $p(h|D)$ also in saddle point approximation the posterior must be maximised with respect to h and θ simultaneously .

3 Gaussian regression

In general density estimation problems $p(y_i|x_i, h)$ is not restricted to a special form, provided it is non–negative and normalised [9, 10]. In this paper we concentrate on Gaussian regression where the single data likelihoods are assumed to be Gaussians

$$p(y_i|x_i, h) = \sqrt{\frac{\beta}{2\pi}}e^{-\frac{\beta}{2}(h(x_i)-y_i)^2}. \tag{6}$$

In that case the unknown regression function $h(x)$ represents the hidden variables and h–integration means functional integration $\int dh \to \int \prod_x dh(x)$.

As simple building blocks for mixture priors we choose Gaussian (process) prior components [2, 17, 14],

$$p(h|\beta, \theta, j, D_0) = \left(\frac{\beta}{2\pi}\right)^{\frac{d}{2}} (\det \mathbf{K}_j(\theta))^{\frac{1}{2}}$$
$$\times e^{-\frac{\beta}{2}\left(h - t_j(\theta),\ \mathbf{K}_j(\theta)(h - t_j(\theta))\right)} \tag{7}$$

the scalar product notation $(\cdot, \cdot)$ standing for x–integration. The mean $t_j(\theta)(x)$ will in the following also be called an (adaptive) *template function.* Covariances $\mathbf{K}_j^{-1}/\beta$ are real, symmetric, positive (semi–)definite (for positive semidefinite covariances the null space has to be projected out). The dimension d of the h–integral becomes infinite for an infinite number of x–values (e.g.

continuous x). The infinite factors appearing thus in numerator and denominator of (5) however cancel. Common smoothness priors have $t_j(\theta) = 0$ and as $\mathbf{K}_j$ a differential operator, e.g., the negative Laplacian.

Analogously to simulated annealing it will appear to be very useful to vary the 'inverse temperature' β simultaneously in (6) (for training but not necessarily for test data) and (7). Treating β not as a fixed variable, but including it explicitly as hidden variable, the formulae of Sect. 2 remain valid, provided the replacement $h \rightarrow (h, \beta)$ is made, e.g. $p(y_i|x_i, h) \rightarrow p(y_i|x_i, h, \beta)$ (see also Fig.1).

Typically, inverse prior covariances can be related to *approximate symmetries*. For example, assume we expect the regression function to be approximately invariant under a permutation of its arguments $h(x) \approx h(\sigma(x))$ with σ denoting a permutation. Defining an operator $\mathbf{S}$ acting on h according to $\mathbf{S}h(x) = h(\sigma(x))$, we can define a prior process with inverse covariance

$$\mathbf{K} = (\mathbf{I} - \mathbf{S})^T (\mathbf{I} - \mathbf{S}), \tag{8}$$

with identity $\mathbf{I}$ and the superscript ^{T} denoting the transpose of an operator. The corresponding prior energy

$$E_0 = \frac{1}{2}(h,\, \mathbf{K}\, h) = \frac{1}{2}\Big((h - \mathbf{S})h,\, (h - \mathbf{S})h\Big), \tag{9}$$

is a measure of the deviation of h from an exact symmetry under $\mathbf{S}$. Similarly, we can consider a Lie group $\mathbf{S} = e^{\theta \mathbf{s}}$ with $\mathbf{s}$ being the generator of the infinitesimal symmetry transformation. In that case a covariance

$$\mathbf{K} = \frac{1}{\theta^2}(\mathbf{I} - \mathbf{S}_{\text{inf}})^T (\mathbf{I} - \mathbf{S}_{\text{inf}}) = \mathbf{s}^T \mathbf{s}, \tag{10}$$

with prior energy

$$E_0 = \frac{1}{2}(\mathbf{s}h,\, \mathbf{s}h), \tag{11}$$

can be used to implement approximate invariance under the infinitesimal symmetry transformation $\mathbf{S}_{\text{inf}} = \mathbf{I} + \theta \mathbf{s}$. For appropriate boundary conditions, a negative Laplacian $\mathbf{K}$ can thus be interpreted as enforcing approximate invariance under infinitesimal translations, i.e., for $\mathbf{s} = \partial/\partial x$.

4 Prior mixtures

4.1 General formalism

Decomposed into components the posterior density becomes

$$\begin{aligned} p(h, \beta|D) \;\; &\propto \;\; \int d\theta \sum_j^m p(y_T|x_T, h, \beta) \\ &\quad \times\, p(h|\beta, \theta, j, D_0)\, p(\beta, \theta, j|D_0). \end{aligned} \tag{12}$$

Writing probabilities in terms of energies, including parameter dependent normalisation factors and skipping parameter independent factors yields

$$\begin{aligned} p(y_T|x_T, h, \beta) \;\; &\propto \;\; e^{-\beta E_T + \frac{n}{2}\ln\beta} \\ p(h|\beta, \theta, j, D_0) \;\; &= \;\; e^{-\beta E_{0,j} + \frac{d}{2}\ln\beta} \\ &\qquad \times e^{\frac{1}{2}\ln\det \mathbf{K}_j(\theta)} \\ p(\beta, \theta, j|D_0) \;\; &\propto \;\; e^{-E_{\theta,\beta,j}}. \end{aligned} \tag{13}$$

This defines hyperprior energies $E_{\theta,\beta,j}$, prior energies $E_{0,j}$ ('quadratic concepts')

$$E_{0,j} = \frac{1}{2}\Big(h - t_j(\theta),\, \mathbf{K}_j(\theta)(h - t_j(\theta, j))\Big), \tag{14}$$

(the generalisation to a sum of quadratic terms $E_{0,j} = \sum_k E_{0,k,j}$ is straightforward) and training or likelihood energy (training error)

$$E_T = \frac{1}{2}\sum_i^n (h(x_i) - y_i)^2 \tag{15}$$

$$= \frac{1}{2}\left(\Big(h - t_T,\, \mathbf{K_T}(h - t_T)\Big) + \sum_i^n V_T(x_i)\right).$$

The second line is a 'bias–variance' decomposition where

$$t_T(x_i) = \sum_k^{n_{x_i}} \frac{y_k(x)}{n_{x_i}}, \tag{16}$$

is the mean of the n_{x_i} training data available for x_i, and

$$V_T(x_i) = \sum_k^{n_{x_i}} \frac{y_k^2(x)}{n_{x_i}} - t_T^2(x_i), \tag{17}$$

is the variance of y_i values at x_i. (V_i vanishes if every x_i appears only once.) The diagonal matrix $\mathbf{K}_T$ is restricted to the space of x for which training data are available and has matrix elements n_x.

4.2 Maximum a posteriori approximation

In general density estimation the predictive density can only be calculated approximately, e.g. in maximum a posteriori approximation or by Monte Carlo methods. For Gaussian regression, however the predictive density of mixture models can be calculated exactly for given θ (and β). This provides us with the opportunity to compare the simultaneous maximum posterior approximation with respect to h and θ with an analytical h–integration followed by a maximum posterior approximation with respect to θ.

Maximising the posterior (with respect to h, θ, and possibly β) is equivalent to minimising the mixture energy (regularised error functional [13, 17, 15, 16])

$$E = -\ln \sum_j^m e^{-E_j + c_j}, \tag{18}$$

with component energies

$$E_j = \beta E_{h,j} + E_{\theta,\beta,j}, \quad E_{h,j} = E_T + E_{0,j}, \tag{19}$$

and

$$c_j(\theta, \beta) = \frac{1}{2}\ln\det \mathbf{K}_j(\theta) + \frac{d+n}{2}\ln\beta. \tag{20}$$

In a direct saddle point approximation with respect to h and θ stationarity equations are obtained by setting the (functional) derivatives with respect to h and θ to zero,

$$0 = \sum_j^m a_j \Big(\mathbf{K}_T(h - t_T) + \mathbf{K}_j(h - t_j)\Big), \tag{21}$$

$$0 = \sum_j^m a_j \left(\frac{\partial E_j}{\partial \theta} - \mathrm{Tr}\left(\mathbf{K}_j^{-1}\frac{\partial \mathbf{K}_j}{\partial \theta}\right)\right), \tag{22}$$

where the derivatives with respect to θ are matrices if θ is a vector,

$$a_j = p(j|h, \theta, D_0) \tag{23}$$

$$= \frac{e^{-\beta E_{0,j} - E_{\theta,\beta,j} + \frac{1}{2}\ln\det \mathbf{K}_j}}{\sum_k^m e^{-\beta E_{0,k} - E_{\theta,\beta,k} + \frac{1}{2}\ln\det \mathbf{K}_k}},$$

and

$$\frac{\partial E_j}{\partial \theta} = \frac{\partial E_{\theta,\beta,j}}{\partial \theta} + \beta \left(\frac{\partial t_j}{\partial \theta},\, \mathbf{K}_j (t_j - h) \right) + \frac{\beta}{2} \Big((h - t_j),\, \frac{\partial \mathbf{K}_j}{\partial \theta} (h - t_j) \Big). \tag{24}$$

Eq.(21) can be rewritten

$$h = \mathbf{K}_a^{-1} \left(\mathbf{K}_T t_T + \sum_l^m a_j \mathbf{K}_j t_j \right), \tag{25}$$

with

$$\mathbf{K}_a = \left(\mathbf{K}_T + \sum_j^m a_j \mathbf{K}_j \right). \tag{26}$$

Due to the presence of h–dependent factors a_j, Eq.(25) is still a nonlinear equation for $h(x)$. For the sake of simplicity we assumed a fixed β; it is no problem however to solve (21) and (22) simultaneously with an analogous stationarity equation for β.

4.3 Analytical solution

The optimal regression function under squared–error loss — for Gaussian regression identical to the log–loss of density estimation — is the predictive mean. For mixture model (12) one finds, say for fixed β,

$$\bar{y} = \int dy\, y\, p(y|x, D) = \sum_j \int d\theta\, b_j(\theta)\, \bar{t}_j(\theta), \tag{27}$$

with mixture coefficients

$$\begin{aligned} b_j(\theta) &= p(\theta, j|D) \\ &= \frac{p(\theta, j)\, p(y_T|x_T, D_0, \theta, j)}{\sum_j \int d\theta p(\theta, j)\, p(y_T|x_T, D_0, \theta, j)}. \end{aligned} \tag{28}$$

The component means $\bar{t}_j$ and the likelihood of θ can be calculated analytically [17, 14]

$$\begin{aligned} \bar{t}_j &= (\mathbf{K}_T + \mathbf{K}_j)^{-1} (\mathbf{K}_T t_T + \mathbf{K}_j t_j) \\ &= t_j + \mathbf{K}_j^{-1} \widetilde{\mathbf{K}}_j (t_T - t_j), \end{aligned} \tag{29}$$

and

$$p(y_T|x_T, D_0, \theta, j) = e^{-\beta \widetilde{E}_{0,j} + \frac{1}{2} \ln \det (\frac{\beta}{2\pi} \widetilde{\mathbf{K}}_j)}, \tag{30}$$

where

$$\widetilde{E}_{0,j}(\theta) = \frac{1}{2} \big(t_T - t_j,\, \widetilde{\mathbf{K}}_j (t_T - t_j) \big), \tag{31}$$

$$\widetilde{\mathbf{K}}_j(\theta) = (\mathbf{K}_T^{-1} + \mathbf{K}_{j,TT}^{-1})^{-1}, \tag{32}$$

and $\mathbf{K}_{j,TT}^{-1}$ is the projection of the covariance $\mathbf{K}_j^{-1}$ into the $\tilde{n}$–dimensional space for which training data are available. ($\tilde{n} \le n$ is the number of data with distinct x–values.)

The stationarity equation for a maximum a posteriori approximation with respect to θ is at this stage found from (28,30)

$$0 = \sum_j b_j \left(\frac{\partial \widetilde{E}_j}{\partial \theta} - \mathrm{Tr} \left(\widetilde{\mathbf{K}}_j^{-1} \frac{\partial \widetilde{\mathbf{K}}_j}{\partial \theta} \right) \right), \tag{33}$$

where $\widetilde{E}_j = \beta \widetilde{E}_{0,j} + E_{\theta,\beta,j}$. Notice that Eq.(33) differs from Eq.(22) and requires only to deal with the $\tilde{n} \times \tilde{n}$–matrix $\widetilde{\mathbf{K}}$. The coefficient $b_j^* = b_j(\theta^*)$ for θ set to its maximum posterior value is of form (23) with the replacements $\mathbf{K}_j \to \widetilde{\mathbf{K}}_j$, $E_j \to \widetilde{E}_j$.

4.4 High and low temperature limits

Low and high temperature limits are extremely useful because in both cases the stationarity Eq.(21) becomes linear, corresponding thus to classical quadratic regularisation approaches.

In the *high temperature limit* $\beta \to 0$ the exponential factors a_j become h–independent

$$a_j \stackrel{\beta\to 0}{\longrightarrow} a_j^0 = \frac{e^{-E_{\theta,\beta,j}+\frac{1}{2}\ln\det\mathbf{K}_j}}{\sum_k^m e^{-E_{\theta,\beta,k}+\frac{1}{2}\ln\det\mathbf{K}_k}}, \tag{34}$$

(for $b_j^* \to b_j^{0,*}$ replace $\mathbf{K}_j$ by $\widetilde{\mathbf{K}}_j$). The solution $h = \bar{t}$ is a (generalised) 'complete template average'

$$\bar{t} = \mathbf{K}_{a^0}^{-1}\left(\mathbf{K_T} t_T + \sum_l^m a_j^0\,\mathbf{K}_j t_j\right), \tag{35}$$

with

$$\mathbf{K}_{a^0} = \mathbf{K}_T + \sum_j a_j^0\,\mathbf{K}_j. \tag{36}$$

This high temperature solution corresponds to the minimum of the quadratic functional $E_T + \sum_j^m a_j^0 E_{h,j}$,

In the *low temperature limit* $\beta \to \infty$ only the maximal component contributes, i.e.,

$$a_j \stackrel{\beta\to\infty}{\longrightarrow} a_j^\infty = \begin{cases} 1 & : \; j = \mathrm{argmin}_j E_{h,j} \\ 0 & : \; j \neq \mathrm{argmin}_j E_{h,j} \end{cases}, \tag{37}$$

(for b_j^* replace $E_{h,j}$ by $\widetilde{E}_j$) assuming $E_{\beta,\theta,j} = E_\beta + E_{\theta,j}$ or $E_{\beta,\theta,j} = E_\beta + E_j + \beta E_\theta$. Hence, low temperature solutions $h = \bar{t}_j$, are all (generalised) 'component averages' $\bar{t}_j$ provided they fulfil the stability condition

$$E_{h,j}(h = \bar{t}_j) < E_{h,j'}(h = \bar{t}_j), \quad \forall j' \neq j, \tag{38}$$

or, after performing a (generalised) 'bias–variance' decomposition, $2V_j < B_{j'}(j,j) + 2V_{j'}$, with $m \times m$ matrices

$$B_j(k,l) = \Big(\bar{t}_k - \bar{t}_j,\, (\mathbf{K}_D + \mathbf{K}_j)\;(\bar{t}_l - \bar{t}_j)\Big) \tag{39}$$

and (generalised) 'template variances'

$$\begin{aligned} V_j \;=\;& \frac{1}{2}\Bigg(\Big(t_T,\, \mathbf{K}_T\, t_T\Big) + \Big(t_j,\, \mathbf{K}_j\, t_j\Big) \\ -\;& \Big(\bar{t}_j, (\mathbf{K}_T + \mathbf{K}_j)\,\bar{t}_j\Big)\Bigg) = \widetilde{E}_{0,j}. \end{aligned} \tag{40}$$

That means single component averages $\bar{t}_j$ (which minimise $E_{h,j}$ and thus $-\beta E_j + c_j$) become solutions at zero temperature $1/\beta$ in case their (generalised) variance V_j measuring the discrepancy between data and prior term is small enough.

4.5 Equal covariances

Especially interesting are j–independent $\mathbf{K}_j(\theta) = \mathbf{K}_0(\theta)$ with θ–independent determinants so $\det\mathbf{K}_j$ or $\det\widetilde{\mathbf{K}}_j$, respectively, do not have to be calculated.

Notice that this still allows completely arbitrary parameterisations of $t_j(\theta)$. Thus, the template function can for example be a parameterised model, e.g., a neural network or decision tree, and maximising the posterior with respect to θ corresponds to training that model. In such cases the prior term forces the maximum posterior solution h to be similar (as defined by $\mathbf{K_0}$) to this trained parameterised reference model.

The condition of invariant $\det\mathbf{K}_0(\theta)$ does not exclude adaption of covariances. For example, transformations for real, symmetric positive definite $\mathbf{K}_0(\theta)$ leaving determinant and eigenvalues (but not eigenvectors) invariant are of the form $\mathbf{K}(\theta_0) \to \mathbf{K}(\theta) = \mathbf{O}(\theta)\mathbf{K}\mathbf{O}^{-1}(\theta)$ with real,

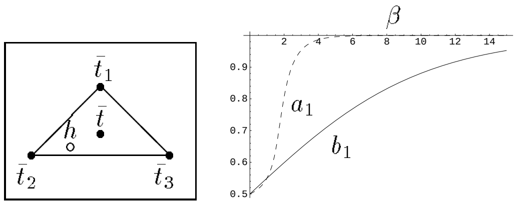

Figure 2: Left: Example of a solution space for $m = 3$. Shown are three low temperature solutions $\bar{t}_j$, high temperature solution $\bar{t}$, and a possible solution h at finite β. Right: Exact b_1 vs. (dominant) a_1 (dashed) for $m = 2$, $b = 2$, $\widetilde{E}_1 = 0.405$, $\widetilde{E}_2 = 0.605$.

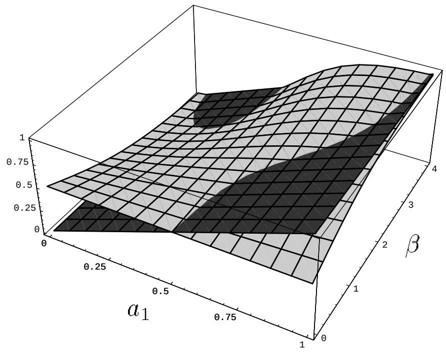

Figure 3: Shown are the plots of $f_1(a_1) = a_1$ and $f_2(a_1) = \frac{1}{2}(\tanh\Delta + 1)$ within the inverse temperature range $0 \le \beta \le 4$ (for $b = 2$, $\widetilde{E}_2 - \widetilde{E}_1 = 0.1\beta$). Notice the appearance of a second stable solution at low temperatures.

orthogonal $\mathbf{O}^{-1} = \mathbf{O}^T$. This allows for example to adapt the sensible directions of multidimensional Gaussians. A second kind of transformations changing eigenvalues but not eigenvectors and determinant is of the form $\mathbf{K}(\theta_0) = \mathbf{O}\mathbf{D}(\theta_0)\mathbf{O}^T \to \mathbf{K}(\theta) = \mathbf{O}\mathbf{D}(\theta)\mathbf{O}^T$ if the product of eigenvalues of the real, diagonal $\mathbf{D}(\theta_0)$ and $\mathbf{D}(\theta)$ are equal.

Eqs.(29,35) show that the high temperature solution becomes a linear combination of the (potential) low temperature solutions

$$\bar{t} = \sum_j^m a_j^0 \bar{t}_j = \sum_j^m b_j^{0,*} \bar{t}_j. \tag{41}$$

Similarly, Eq.(21) simplifies to

$$h = \sum_j^m a_j \bar{t}_j = \bar{t} + \sum_j^m (a_j - a_j^0)\, \bar{t}_j, \tag{42}$$

and Eq.(23) to

$$a_j = \frac{e^{-\frac{\beta}{2} a B_j a - \widetilde{E}_j}}{\sum_k e^{-\frac{\beta}{2} a B_k a - \widetilde{E}_k}} = \frac{b_j\, e^{-\frac{\beta}{2} a B_j a}}{\sum_k b_k\, e^{-\frac{\beta}{2} a B_k a}}, \tag{43}$$

introducing vector a with components a_j, $m \times m$ matrices B_j defined in (39). Eq.(42) is still a nonlinear equation for h, it shows however that the solutions must be convex combinations of the h–independent $\bar{t}_j$ (see Fig. 2). Thus, it is sufficient to solve Eq.(43) for m mixture coefficients a_j instead of Eq.(21) for the function h.

For two prior components, i.e., $m = 2$, Eq.(42) becomes

$$h = \frac{\bar{t}_1 + \bar{t}_2}{2} + (\tanh\Delta)\, \frac{\bar{t}_1 - \bar{t}_2}{2}, \tag{44}$$

with

$$\Delta = \frac{E_2 - E_1}{2} = \frac{\beta}{4}\, b(2a_1 - 1) + \frac{\widetilde{E}_2 - \widetilde{E}_1}{2}, \tag{45}$$

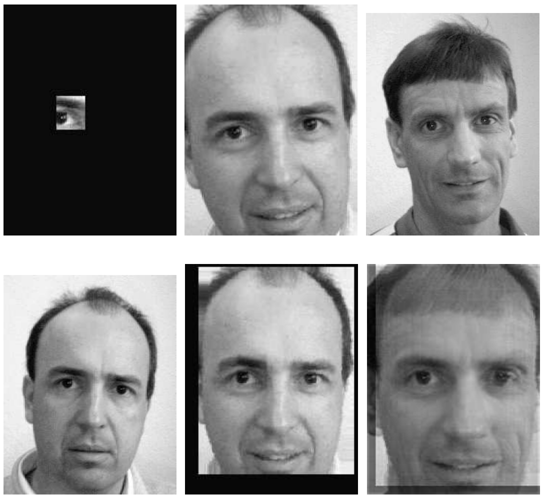

Figure 4: Top row, from left to right: Data points sampled with Gaussian noise, two template functions t_1, t_2. Bottom row, from left to right: Original, reconstructed solutions (regression function h, 180×240 pixels) at low and at high temperature.

because the matrices B_j are in this case zero except $B_1(2,2) = B_2(1,1) = b$. For $E_{\theta,\beta,j}$ uniform in j we have $(\bar{t}_1 + \bar{t}_2)/2 = \bar{t}$ so that $a_j^0 = 0.5$. The stationarity Eq.(43), being analogous to the celebrated mean field equation of a ferromagnet, can be solved graphically (see Fig.3 and Fig.2 for a comparison with b_j), the solution is given by the point where

$$a_1 = \frac{1}{2}\left(\tanh\Delta + 1\right). \tag{46}$$

5 A numerical example

As numerical example we study a two component mixture model for image completion. Assume we expect an only partially known image (corresponding to pixel-wise training data drawn with Gaussian noise from the original image) to be similar to one of the two template images shown in Fig.4. Next, we include hyperparameters parameterising deformations of templates. In particular, we have chosen translations (θ_1, θ_2) a scaling factor θ_3, and a rotation angle (around template center) θ_4.

Interestingly, it turned out that due to the large number of data ($\tilde{n} \approx 1000$) it was easier to solve Eq.(21) for the full discretized image than to invert (32) in the space of training data. A prior operator $\mathbf{K}_0$ has been implemented as a 3×3 negative Laplacian filter. (Notice that using a Laplacian kernel, or another smoothness measure, instead of a straight template matching using simply the squared error between image and template, leads to a smooth interpolation between data and templates.) Completed images h for different β have been found by iterating according to

$$h^{k+1} = h^k + \eta\mathbf{A}^{-1}\Big[\mathbf{K}_T(t_T - h^k) + \mathbf{K}_0\Big(\sum_j a_j^k t_j - h^k\Big)\Big], \tag{47}$$

performed alternating with θ–minimisation. A Gaussian learning matrix $\mathbf{A}^{-1}$ (implemented by a 5×5 binomial filter) proved to be successful. Typically, the relaxation factor η has been set to 0.05.

Being a mixture model with $m = 2$ the situation is that of Fig.3. Typical solutions for large and small β are shown in Fig.4.

6 Conclusions

Prior mixture models are capable to build complex prior densities from simple, e.g., Gaussian components. Going beyond classical quadratic regularisation approaches, they still can use the nice analytical features of Gaussians, and allow to control the degree of the resulting non-convexity explicitly. Combined with parameterised component mean functions and covariances they seem to provide a powerful tool.

Acknowledgements The author was supported by a Postdoctoral Fellowship (Le 1014/1–1) from the Deutsche Forschungsgemeinschaft and a NSF/CISE Postdoctoral Fellowship at the Massachusetts Institute of Technology. Part of the work was done during the seminar 'Statistical Physics of Neural Networks' at the Max–Planck–Institut für Physik komplexer Systeme, Dresden. The author also wants to thank Federico Girosi, Tomaso Poggio, Jörg Uhlig, and Achim Weiguny for discussions.

References

[1] Berger, J.O.: *Statistical Decision Theory and Bayesian Analysis.* New York: Springer Verlag, 1980.

[2] Doob, J.L.: *Stochastic Processes.* New York: Wiley, 1953 (New edition 1990).

[3] Everitt, B.S. & Hand, D.J.: Finite Mixture Distributions. Chapman & Hall, 1981.

[4] Gelman A., Carlin, J.B., Stern, H.S., & Rubin, D.B.: *Bayesian Data Analysis.* New York: Chapman & Hall, 1995.

[5] Girosi, F., Jones, M., & Poggio, T.: Regularization Theory and Neural Networks Architectures. *Neural Computation* **7** (2), 219–269, 1995.

[6] Lemm, J.C.: *Prior Information and Generalized Questions.* A.I.Memo No. 1598, C.B.C.L. Paper No. 141, Massachusetts Institute of Technology, 1996. (available at http://pauli.uni–muenster.de/~lemm)

[7] Lemm, J.C.: *How to Implement A Priori Information: A Statistical Mechanics Approach.* Technical Report MS-TP1-98-12, Universität Münster, 1998. (**cond-mat/9808039**, also available at http://pauli.uni–muenster.de/~lemm.)

[8] Lemm, J.C.: Quadratic Concepts. In Niklasson, L, Boden, M, Ziemke, T.(eds.): *Proceedings of the 8th International Conference on Artificial Neural Networks (ICANN 98)*, Skövde, Sweden, September 2-4, 1998, Springer Verlag, 1998.

[9] Lemm, J.C.: *Bayesian Field Theory.* Technical Report MS-TP1-99-1, Universität Münster, 1999. (available at http://pauli.uni–muenster.de/~lemm.)

[10] Lemm, J.C., Uhlig, J., & Weiguny, A.: *A Bayesian Approach to Inverse Quantum Statistics.* Technical Report MS-TP1-99-6, Universität Münster, 1999. (**cond-mat/9907013**, also available at http://pauli.uni–muenster.de/~lemm.)

[11] Robert, C.P.: *The Bayesian Choice.* New York: Springer Verlag, 1994.

[12] Sivia, D.S.: *Data Analysis: A Bayesian Tutorial.* Oxford: Oxford University Press, 1996.

[13] Tikhonov A.N. & Arsenin V.: *Solution of Ill–posed Problems.* New York: Wiley, 1977.

[14] Williams, C.K.I. & Rasmussen, C.E.: Gaussian processes for regression. In *Proc. NIPS8*, MIT Press, 1996.

[15] Vapnik, V.N.: *Estimation of dependencies based on empirical data.* New York: Springer Verlag, 1982.

[16] Vapnik, V.N.: *Statistical Learning Theory.* New York: Wiley, 1998.

[17] Wahba, G.: *Spline Models for Observational Data.* Philadelphia: SIAM, 1990.

[18] Whittaker, E.T., On a new method of graduation. *Proc. Edinborough Math. Assoc.*, 78, 81-89, 1923.