Tractable Inference for Probabilistic Data Models

Lehel Csató and Manfred Opper
Neural Computing Research Group,
School of Engineering and Applied Science,
Aston University, Birmingham B4 7ET, United Kingdom

Ole Winther
Digital Signal Processing, Informatics and Mathematical Modelling
Technical University of Denmark, B208, 2800 Lyngby, Denmark

September 17, 2002

Abstract

Based on ideas from statistical physics, we present an approximation technique for probabilistic data models with a large number of hidden variables. We give examples for two non–trivial applications.

02.50.-r,87.18.Sn

1 Introduction

Probabilistic data models explain the complexity of observed data by a set of hidden, unobserved causes which are modeled as random variables. Examples are: Bayes belief networks [23] (used as trainable expert systems), independent component analysis [11, 1] (abbreviated ICA, which detects independent sources in nonlinear signal processing), Gaussian process models [17] (modeling hidden spatial structures by random fields) and Boltzmann

machines [7] (the Ising version of the random fields). Based on the joint distribution of all variables one can assign plausible numerical values to the hidden causes from suitable conditional averages over the hidden variables. Unfortunately, except for a few simple cases (when the graph of dependencies between random variables has the structure of a tree, or, when the joint distribution is Gaussian) inference with probabilistic models usually requires approximations, when variables are non-Gaussian and/or when the number of variables is large.

In recent years, a variety of approximation techniques have been imported from the field of statistical physics. One of the simplest methods is the well known mean field (MF) approximation which approximates the joint distribution of variables by a factorizing one. To take the neglected correlations at least partly into account, a correction to the MF method such as the Bethe/Kikuchi approximation [29] and the TAP approach, see e.g.[12, 28, 16, 10, 25, 17, 9] and references in [15] have become popular.

The TAP method introduced by Thouless, Anderson & Palmer [4] for disordered materials has the appealing feature to become exact for certain statistical physics models with infinite ranged *random* interactions. We can view data models as disordered systems because the observed random data are parameters in the conditional distributions for the hidden variables. This should make the TAP approximation a good candidate for inference in probabilistic data modelling. Unfortunately, the method requires the exact knowledge of the distribution of the disorder, which for statistical physics models is usually assumed to be known, but for *real data* typically not. In order to make the method a general tool for practical applications, we have recently developed a version of the TAP approach [19, 18] which no longer requires the knowledge of the underlying distribution but adapts to the concrete observations. In this paper we present a simple simple derivation of our *adaptive* TAP (ADATAP) method and demonstrate its applications to a model for classification and an ICA model.

The paper is organized as follows. Section 2 gives two examples for rather different probabilistic models which both can be described by a class of probability distribution introduced in section 3. Our approximation method using the Gibbs free energy is explained in section 4 and applications are demonstrated in section 5. The paper concludes with a brief outlook.

2 Probabilistic Models: Two Examples

2.1 Gaussian Process Models for Classification

Gaussian process (GP) models have become popular in recent years as a nonparametric approaches for supervised learning [27, 6, 17]. Take e.g., a binary classification problem, where we would like to classify input features $x \in R^D$ (which might be the $D = 16 \times 16$ dimensional vectors of pixel values for digitized handwritten characters) into two classes $y = \pm 1$ (which could be the handwritten digit "3" against all other digits). A probabilistic model could assume that the observed class labels are generated as $y = \mathrm{sign}[f(x)+\xi]$ with an unknown function f, and where ξ is a zero mean noise process. The equation $f(x) = 0$ would give us the ideal decision surface for the separation of inputs with labels $y = 1$ and $y = -1$. Assuming Gaussian noise for ξ, the *likelihood* of observing a label y, based on *knowing* the function value $f(x)$ is given by

$$P(y|f(x)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} dt \exp\left[-\frac{t^2}{2}\right] \ . \tag{1}$$

where $u = yf(x)/\sigma_0$ and σ_0 measures the noise level. In a Bayesian probabilistic model also the unknown function f becomes a random variable. We will encode a vague prior knowledge about the variability of functions f with their arguments x by modelling them as a realizations of a Gaussian random field. For such *Gaussian processes* (with zero mean) the entire distribution $P_0[f]$ over function space is determined by its correlation function (or *kernel*) $K(x, x')$ which has to be supplied by the user of the algorithm. A popular choice of a general purpose kernel is the so-called radial basis function (RBF) kernel

$$K(x - x') = e^{-||x-x'||^2/l^2} \ , \tag{2}$$

where l is a lengthscale. When a dataset of N correctly classified input/label pairs $D = (x_1, y_1), \ldots, (x_N, y_N)$ is available for training, one can use Bayes' theorem of probability to convert the prior distribution P_0 together with the likelihood into a posterior, conditional distribution over functions

$$P[f|D] = \frac{1}{Z} P_0[f] \cdot \prod_{i=1}^{N} P(y_i|f_i) \ , \tag{3}$$

where $f_i \equiv f(x_i)$ and Z acts a a normalizer. With an increasing number of training data one expects that the posterior distribution (3) becomes more

and more concentrated around the function f which optimally classifies the data. Good predictions on novel test inputs x could then be based naturally on the the average $\langle f(x) \rangle$ over the distribution (3) which could be used to classify the new inputs x as $y = \text{sign}[\langle f(x) \rangle]$. Although it seems that we have to perform an explicit functional integration in order to obtain $\langle f(x) \rangle$ one can show [17] that it is possible to write the result as a weighted sum of correlation kernels centered at the training inputs

$$\langle f(x) \rangle = \sum_{i=1}^{N} \alpha_i K(x, x_i) \ , \tag{4}$$

with $\alpha_j = \langle \frac{\partial \ln P(y_j|f_j)}{\partial f_j} \rangle$. Hence, the computation of each α_j requires only the *marginal distribution* of f_j. Hence, we can restrict ourselves to the joint distribution of function values at the training inputs x_i, which is

$$p(f_1, \ldots, f_N|D) \propto \exp\left[-\frac{1}{2} \sum_{i,j} f_i (K^{-1})_{ij} f_j \right] \cdot \prod_{i=1}^{N} P(y_i|f_i) \tag{5}$$

and the matrix K is defined through the kernel via $K_{ij} = K(x_i, x_j)$.

At first glance this probabilistic, *nonparametric* approach seems like a fairly complicated technique to perform classifications. However, the advantage, compared to *parametric* techniques such as neural networks, lies in the fact that the effective complexity of the model is not fixed beforehand but will effectively adapt to the dataset. In practice, kernel methods are found to overfit only weakly and their performance can be optimized by adapting kernel *hyperparameters* (e.g. the lengthscale l in (2)). Furthermore, the applicability of *kernel machines* to various non-trivial problems has been increased by the development of new types of kernels which are especially designed for classifying complex types of objects such as texts or protein strings [8].

The *disadvantage* of the GP models comes from the fact that the necessary mathematical operations can not be performed exactly in an efficient way. Besides the problem of analytically intractable distributions (5), the high dimensionality of correlation matrices K_{ij} make computations inefficient, when the size N of the training data sets becomes large. The latter is also a problem for nonprobabilistic kernel methods such as *support vector machines* (SVMs) [2, 26]. Before discussing our solution to this problem, we will briefly introduce a second, quite different probabilistic model.

2.2 Probabilistic Independent Component Analysis (ICA)

ICA is a widely applicable approach [11, 1, 21] in nonlinear signal processing, which aims at decomposing signals obtained from different sensors into a set of statistically independent *sources*. This finds a variety of applications e.g. in the analysis of biomedical data, wheres on tries to separate an "interesting" part of the signal from other statistically independent contributions. In the simplest *probabilistic* formulation of ICA (for other approaches, which do not assume the full statistics of the sources, see [11, 1]), one assumes that the vector $\mathbf{X}_t$ of signals at time t is an instantaneous linear mixing of sources $\mathbf{S}_t$ corrupted by additive Gaussian noise Γ_{ij}. We can write

$$\mathbf{X}_t = \mathbf{A}\mathbf{S}_t + \mathbf{\Gamma}_t \ , \tag{6}$$

where $\mathbf{A}$ is an unknown (but time independent) mixing matrix and the noise vector is assumed to be without temporal correlations having a time independent covariance matrix $\mathbf{\Sigma}$. The distribution (likelihood) of the signal vector for given parameters $\mathbf{A}$, $\mathbf{\Sigma}$ and the unknown sources $\mathbf{S}_t$ at time t therefore has form

$$P(\mathbf{X}_t|\mathbf{A},\mathbf{\Sigma},\mathbf{S}_t) = (\det 2\pi\mathbf{\Sigma})^{-\frac{1}{2}}\, e^{-\frac{1}{2}(\mathbf{X}_t-\mathbf{A}\mathbf{S}_t)^T\mathbf{\Sigma}^{-1}(\mathbf{X}_t-\mathbf{A}\mathbf{S}_t)} \ . \tag{7}$$

The total probability of the temporal signal is assumed to be factorizing in time, ie. $P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma},\mathbf{S}) = \prod_t P(\mathbf{X}_t|\mathbf{A},\mathbf{\Sigma},\mathbf{S}_t)$. The aim of independent component analysis is to recover the unknown quantities which are the mixing matrix $\mathbf{A}$, the noise covariance $\mathbf{\Sigma}$ and the unknown sources $\mathbf{S}$ from the observed data. The crucial assumption is that of *statistical independence* of the sources (the hidden variables) at each time t, i.e.

$$P(\mathbf{S}_t) = \prod_i P(S_{it}) \ . \tag{8}$$

A suitable functional form (which has to be non-Gaussian) for the source distribution $P(S)$ which incorporates e.g. known constraints (such as positivity or sub-Gaussian tails) must be chosen for each individual application. Alternatively, the source distribution can also be specified such that it can adapt to the data, see e.g. Ref. [20].

Again, we can get plausible values for the unobserved sources by averaging the random values S_{it} over the posterior distribution computed from the prior

(8) and the likelihood of the observations (7). However, we must also learn the the mixing matrix $\mathbf{A}$ and the noise covariance $\mathbf{\Sigma}$ in parallel. These can be estimated from the training data by the method of *maximum likelihood* (ML) [21], i.e. by maximizing the total probability of the observations

$$P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma}) = \int d\mathbf{S} P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma},\mathbf{S})\, P(\mathbf{S}) \tag{9}$$

under the statistical assumptions. For the estimator of the mixing matrix, the resulting set of nonlinear equations reads

$$\mathbf{A}_{\mathrm{ML}} = \sum_t \mathbf{X}_t \langle \mathbf{S}_t \rangle^T \left(\sum_{t'} \langle \mathbf{S}_{t'} \mathbf{S}_{t'}^T \rangle \right)^{-1} \tag{10}$$

$$\mathbf{\Sigma}_{\mathrm{ML}} = \frac{1}{N} \sum_t \langle (\mathbf{X}_t - \mathbf{A}\mathbf{S}_t)\, (\mathbf{X}_t - \mathbf{A}\mathbf{S}_t)^T \rangle \ . \tag{11}$$

The brackets again denote an average over the posterior distribution of the sources given the observations. Note, that also the right hand side depends on $\mathbf{A}_{\mathrm{ML}}$ via the posterior. This apparent complication can be solved using the EM–algorithm [5] where the moments and the parameters are updated alternately. It can be proved that the likelihood $P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma})$ increases (or stays constant) in each step of the EM–algorithm. Although the *prior* distribution (8) assumed independent sources, the posterior will obviously have *correlations* between different sources (but still independence for different times t), which again makes computations of averages non–trivial.

3 A canonical Model

It is not hard to show that the two previous examples of probabilistic models (and, in fact, many others) require the computation of averages over posterior distributions of hidden variables which are of the type

$$P(\mathbf{S}) = \frac{\rho(\mathbf{S})}{Z} \exp\left[\frac{1}{2} \sum_{i,j} S_i J_{ij} S_j \right] \ . \tag{12}$$

The set of couplings J_{ij}'s encodes pairwise dependencies between the random variables $\mathbf{S} = (S_1, \ldots, S_N)$. The factorizing term $\rho(\mathbf{S}) = \prod_j \rho_j(S_j)$ (called

likelihood in the following) usually contains local observations at a site i, but can also incorporate additional local prior information about the variables S_i. E.g., by proper choices of the ρ_j's we can include both *discrete and continuous* random variables in the same model (12). The normalizing partition function Z is often (within a constant) equal to the probability that the model gives to the observed variables, which can be used as a yardstick for comparing different models or optimizing their hyperparameters. (12) includes the GP classifier model, if we define $\mathbf{S}$ to be the restriction of the random field to the training inputs ie, $S_i = f(\mathbf{x}_i)$. The prior correlations between the variables leads to the interactions $\mathbf{J} = -\mathbf{K}^{-1}$ where $K_{ij} = K(x_i, x_j)$. The ICA model is recovered by identifying $\mathbf{S}$ with the vector of sources and setting the coupling matrix to $\mathbf{J} = \mathbf{A^T \Sigma^{-1} A}$ and $\rho(\mathbf{S}_t) = P(\mathbf{S}_t) \exp \mathbf{X^T \Sigma^{-1} A S}_t$.

In the rest of the paper we will develop a simple and computationally efficient method for approximating marginal moments and correlation functions for the distribution (12) which enables us to deal with a variety of probabilistic models on real data.

4 The Gibbs Free Energy

We will derive our approximation scheme based on a *Gibbs Free Energy* G. It is an entropic quantity which allows us to compute moments of the distribution P, eq. (12) as well as the log of the normalization, $-\ln Z$ within the same approach. G is defined by a constrained minimization of a relative entropy measure

$$D(Q||P) \equiv \int d\mathbf{S}\; Q(\mathbf{S}) \ln \frac{Q(\mathbf{S})}{P(\mathbf{S})} \tag{13}$$

between a distribution Q and the posterior distribution P, where a set of relevant marginal moments are fixed. To be precise, we define

$$G(\mathbf{m}, \mathbf{M}) = \min_Q \left\{ D(Q||P) \mid \langle \mathbf{S} \rangle_Q = \mathbf{m}, \langle \mathbf{S}^2 \rangle_Q = \mathbf{M} \right\} - \ln Z \;, \tag{14}$$

where the brackets denote expectations with respect to the variational distribution Q. $\langle \mathbf{S}^2 \rangle_Q$ is shorthand for a vector with elements $\langle S_i^2 \rangle_Q$. Minimizing G with respect to all arguments obviously leads to $\min_{\mathbf{m},\mathbf{M}} G(\mathbf{m}, \mathbf{M}) = -\ln Z$, where total the minimizer is just $Q = P$. Hence, the moments of the distri-

bution P are obtained as

$$\langle \mathbf{S} \rangle, \langle \mathbf{S}^2 \rangle = \underset{\mathbf{m},\mathbf{M}}{\operatorname{argmin}} G(\mathbf{m}) \ . \tag{15}$$

Also correlation functions can be obtained from G as derivatives

$$\frac{\partial^2 G}{\partial m_i \partial m_j} = (\boldsymbol{\chi}^{-1})_{ij} \ , \tag{16}$$

where $\chi_{ij} = \langle S_i S_j \rangle - \langle S_i \rangle \langle S_j \rangle$ and the derivatives are taken at the minimum of G. An explicit expression for G is obtained by solving the constrained minimization problem (14) with the help of Lagrange multipliers. The minimizng distribution is found to be of the form

$$Q^*(\mathbf{S}) = Z^{-1}(\boldsymbol{\gamma}, \boldsymbol{\lambda}) \rho(\mathbf{S}) \exp[\sum_i \gamma_i S_i - \frac{1}{2} \sum_{i,j} S_i (\delta_{ij} \lambda_i - J_{ij}) S_j)] \ , \tag{17}$$

where the λ_i's and γ_i's are Lagrange parameters which must be chosen to fulfill the constraints, and $Z(\boldsymbol{\gamma}, \boldsymbol{\lambda})$ is a normalizing partition function. Using (17) one can show that

$$G(\mathbf{m}, \mathbf{M}) = \max_{\boldsymbol{\lambda}, \boldsymbol{\gamma}} \left\{ - \ln Z(\boldsymbol{\gamma}, \boldsymbol{\lambda}) + \mathbf{m}^T \boldsymbol{\gamma} - \frac{1}{2} \mathbf{M}^T \boldsymbol{\lambda} \right\} \ . \tag{18}$$

Unfortunately, (17) is as complicated as the original distribution (12). To approximate the Gibbs free energy, we split G into two terms

$$G = G^0 + \Delta G, \tag{19}$$

where G^0 is the Gibbs free energy for the distribution (12), but where all couplings between the random variables are set to zero, ie. where $J_{ij} = 0$. The computation of the corresponding partition function is easy and the Gibbs free energy G^0 for such a "free" model is obtained from (18) by solving a convex optimization problem. Previous versions of the TAP approximation have been obtained by truncating a power series expansion of ΔG with respect to the interactions J_{ij} at second order [24, 25].

In contrast, our ADATAP approximation (which was motivated by the treatment of Parisi and Potters [22] of an Ising model with random orthogonal coupling matrix) will include terms of arbitrary order in the interactions. It

will be defined in such a way that it becomes exact when (12) is a Gaussian distribution, ie. when the likelihoods are of the form

$$\rho_i^g(S) = \exp[a_i S - \frac{b_i}{2} S^2] \tag{20}$$

for $i = 1, \ldots, N$. The interaction part for likelihoods (20) is $\Delta G^g \equiv G_{\rho^g} - G^0_{\rho^g}$, where the subscripts on the right hand side denote the explicit dependence on the likelihood ρ^g. However, using (17) and (18) it can be shown, that the resulting ΔG^g (and the optimizing Gaussian distributions (17)) come out *independent of* the actual Gaussian likelihood ρ_i^g (20) chosen to compute G_{ρ^g}. It is only a function of the moments $\mathbf{m}$ and $\mathbf{M}$ and equals

$$\Delta G^g(\mathbf{m}, \mathbf{M}) = \max_{\mathbf{\Lambda}} \left\{ \frac{1}{2} \ln \det(\mathbf{\Lambda} - \mathbf{J}) - \frac{1}{2} \mathbf{m}^T \mathbf{J} \mathbf{m} - \frac{1}{2} \sum_i \Lambda_i \chi_{ii} \right\} + \frac{1}{2} \sum_i \ln \chi_{ii} + \frac{N}{2} \,, \tag{21}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with entries Λ_i and $\chi_{ii} \equiv M_i - m_i^2$. The *universal form* (21) will serve as an approximation to ΔG for *arbitrary likelihoods*. Hence, our ADATAP approximation for the Gibbs free energy is simply

$$G \approx G^0 + \Delta G^g \ . \tag{22}$$

Hence, in our approximation, the problem of computing certain averages with the distribution (12) has been reduced to an optimization problem. Furthermore, the method also computes implicitly a (Gaussian) approximation (via (17) and ΔG^g) to the *full joint distribution* (12) which can be useful in a variety of applications.

There are various ways of finding (at least local) minima of G. We have developed a *message passing algorithm* (based on an earlier idea of T. Minka [14]) that has quadratic convergence close to a minima and is found to perform efficiently in practice.

5 Applications

5.1 Sparse approximation for Gaussian process classifiers

A straightforward application of the ADATAP approximation for computing predictions with the GP classifier model (5) is possible (for details [17]),

but the operations involving large matrices makes the method in its standard version impractical for datasets of several thousand training examples. Hence, a further approximation introducing *sparsity* is necessary. The idea is to replace the distribution P (more precisely, its tractable Gaussian approximation Q^* which is implicitly computed with ΔG^g) by another one having a likelihood which depends only on a smaller subset of variables called "basis vectors" (BV) of size $n << N$ [3]. In order to minimize the loss of information caused by sparsity, the new distribution $\hat{Q}$ with a sparse likelihood is optmized by minimizing the relative entropy $D(\hat{Q}||Q^*)$. For $\hat{Q}$ and Q^* Gaussian, this can be done in closed form. The sparse approximation is implemented in the sequential message passing algorithm and it is decided at each step, whether the new variable is included in the BV set or not. Since the algorithm sweeps several times through the set of variables, one can not discard the non-BVs completely. Some non-BVs variables may become BVs in a later sweep. Hence, the maximal size of the matrices involved in the algorithm is $n \times N$ (rather than $n \times n$), which is still better than the original size $N \times N$.

We have run the sparse ADATAP algorithm on the USPS dataset[1] of gray-scale handwritten digit images of size 16×16. For the kernel we choose an RBF kernel $K(x, x') = a_K \exp(-\|x - x'\|^2/(m\sigma_K^2))$ where m is the dimension of the inputs (256 in this case), and a_K and σ_K are parameters. In the simulations we used $N = 7000$ random training examples. The task was to classify the digits into fours and non-fours. Figure 1 shows the percentage of errors of the classifier on 2007 test patterns for different sizes n of the BV set. The result shows a saturation of errors with increasing BV set suggesting that the sparse approximation extracts sufficient information from the data. We have also compared multiple sweeps of the algorithm with the result of a single sweep (averaged over different permutations examples in the sequence) which show that fluctuations caused by different orders of presentations are diminished for multiple sweeps.

5.2 Independent Component Analysis

Independent component analysis is applied in a wide variety of data analysis and blind separation tasks e.g. for images, sound, text and telecommuni-

[1] Available from `http://www.kernel-machines.org/data/`

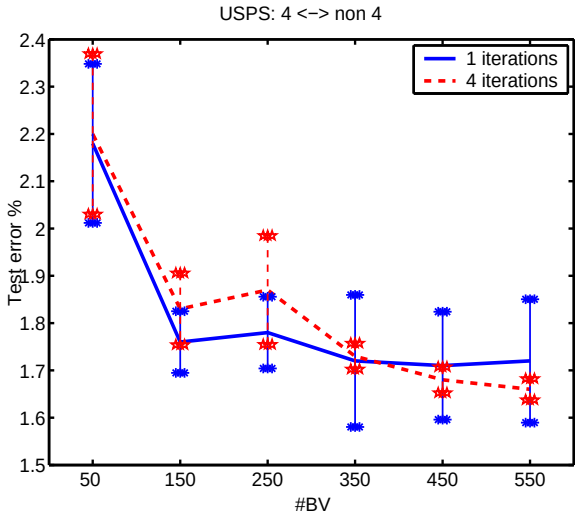

Figure 1: Test errors for classification with different BV sizes (x-axis) and multiple sweeps through the data.

cation problems [11, 1]. Here we will shortly present an application of the message passing implementation of ADATAP algorithm to feature extraction in hand-written digits. For more detail about this problem and other examples that illustrate the flexibility and range of applications of ICA, see [21]. We compare the ICA result to a standard feature extraction/visualization technique namely principal component analysis (PCA).

We assume positive components of $\mathbf{A}$ (enforced by Lagrange multipliers) and a positive exponential prior on the sources

$$P(S_{it}) = \Theta(S_{it}) \exp(-S_{it}) \tag{23}$$

As in [21] we used 500 handwritten '3's which are assumed to be generated by 25 hidden images. The motivation for enforcing positivity is that such strong constraints (i.e. the images are generated by positive additions) will force the solution to become sparse, i.e. with many zeros in $\mathbf{A}$ and $\langle \mathbf{S} \rangle$. This will give us the statistically independent different stroke styles as seen in figure 2. This can be compared to 25 principal components with largest eigenvalues that exhibit the typical "shadow effects" that occur when both negative and positive values are possible. The sparse basis set found by the ICA algorithm can be seen as a statistically more reasonable representation of the components of images than the one found by PCA since it models

more closely the true generative process of handwriting. Projection on this basis can be a powerful preprocessing step for hand-written digit classifiers.

6 Conclusion and Outlook

We have demonstrated how approximation techniques from statistical physics can help to solve problems in data modelling. We expect that our ADATAP approximation will become a practical tool for inference with a variety of probabilistic data models. In fact, we are presently developing program packages both for ICA, Gaussian processes and general model of the type (12) that are made available online[2].

An important future direction of research will be the development of systematic improvements of the approximation. The computation of corrections will not only be of interest from a theoretical point of view but could also provide a user of the method with a measure of how well the final result can be trusted. It will also be of special importance to understand the practical relevance of possible multiple minima of the approximate Gibbs free energy, cooresponding to multiple solutions to TAP equations, which in statistical physics are well known for models of spinglass type [13]. In such a case, one might expect that more complex types of approximations become necessary.

Acknowledgments

The work was supported by EPSRC grant no. GR/M81601.

References

[1] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[2] C. Cortes and V. N. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[3] L. Csató and M. Opper. Sparse on-line gaussian processes. *Neural Computation*, 14:641–668, 2002.

[2]Available from `http://isp.imm.dtu.dk/staff/winther/`.

[4] D. J. Thouless, P. W. Anderson, and R. G. Palmer. Solution of a 'solvable model of a spin glass'. *Phil. Mag.*, 35:593, 1977.

[5] N. M. Dempster, A. P. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39:185–197, 1977.

[6] M. N. Gibbs. *Bayesian Gaussian Processes for Regression and Classification.* PhD thesis, Cambridge University, 1997.

[7] G. E. Hinton and T. J. Sejnowski. In *IEEE Conference on Computer Vision and Pattern Recognition (Washington, 1983)*, page 448. IEEE Press, New York, 1983.

[8] T. Jaakkola, M. Diekhans, and D. Haussler. Using the fisher kernel method to detect remote protein homologies. In *The Seventh International Conference on Intelligent Systems for Molecular Biology*, 1999.

[9] Y. Kabashima and D. Saad. Belief propagation vs. tap for decoding corrupted messages. *Euro. Phys. Lett.*, 44:668, 1998.

[10] H. J. Kappen and F. B. Rodríguez. Efficient learning in boltzmann machines using linear response theory. *Neural Computation*, 10:1137, 1998.

[11] T.-W. Lee. *Independent Component Analysis.* Kluwer Academic Publishers, Boston, 1998.

[12] M. Mézard. The space of interactions in neural networks: Gardner's computation with the cavity method. *J. Phys. A (Math. Gen.)*, 22:2181, 1989.

[13] M. Mézard, G. Parisi, and M. A. Virasoro. *Spin Glass Theory and Beyond*, volume 9 of *Lecture Notes in Physics.* World Scientific, 1987.

[14] T. P. Minka. *Expectation propagation for approximate Bayesian inference.* PhD thesis, Dep. of Electrical Eng. and Comp. Sci.; MIT, 2000.

[15] M. Opper and eds D. Saad. *Advanced Mean Field Methods, Theory and Practice.* MIT Press, 2001.

[16] M. Opper and O. Winther. A mean field approach to bayes learning in feed-forward neural networks. *Phys. Rev. Lett.*, 76:1964, 1996.

[17] M. Opper and O. Winther. Gaussian processes for classification: Mean field algorithms. *Neural Computation*, 12:2655, 2000.

[18] M. Opper and O. Winther. Adaptive and self-averaging thouless-anderson-palmer mean field theory for probabilistic modeling. *Phys. Rev. E*, 64:056131, 2001.

[19] M. Opper and O. Winther. Tractable approximations for probabilistic models: The adaptive thouless-anderson-palmer mean field approach. *Phys. Rev. Lett.*, 86:3695–3699, 2001.

[20] P. A. d. F. R. Højen-Sørensen, O. Winther, and L. K. Hansen. Analysis of functional neuroimages using ica adaptive binary sources. *to appear in Neurocomputing*, 2002.

[21] P. A. d. F. R. Højen-Sørensen, O. Winther, and L. K. Hansen. Mean field approaches to independent component analysis. *Neural Computation*, 14:3695–3699, 2002.

[22] G. Parisi and M. Potters. Mean-field equations for spin models with orthogonal interaction matrices. *J. Phys. A (Math. Gen.)*, 28:5267, 1995.

[23] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, 1988.

[24] T. Plefka. Convergence condition of the tap equations for the infinite-ranged ising spin glass model. *J. Phys. A*, 15:1971, 1982.

[25] T. Tanaka. Mean-field theory of boltzmann machine learning. *Phys. Rev. E*, 58:2302, 1998.

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

[27] C. K. I. Williams and C. E. Rasmussen. Gaussian proceses for regression. In *Advances in Neural Information Processing Systems*, number 8, pages 514–520, 1996.

[28] K. Y. M. Wong. Microscopic equations and stability conditions in optimal neural networks. *Europhys. Lett.*, 30:245, 1995.

[29] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In T.G. Dietterich T.K. Leen and V. Tresp, editors, *Advances in Neural Information Processing Systems*, number 13, pages 689–695. MIT Press, 2001.

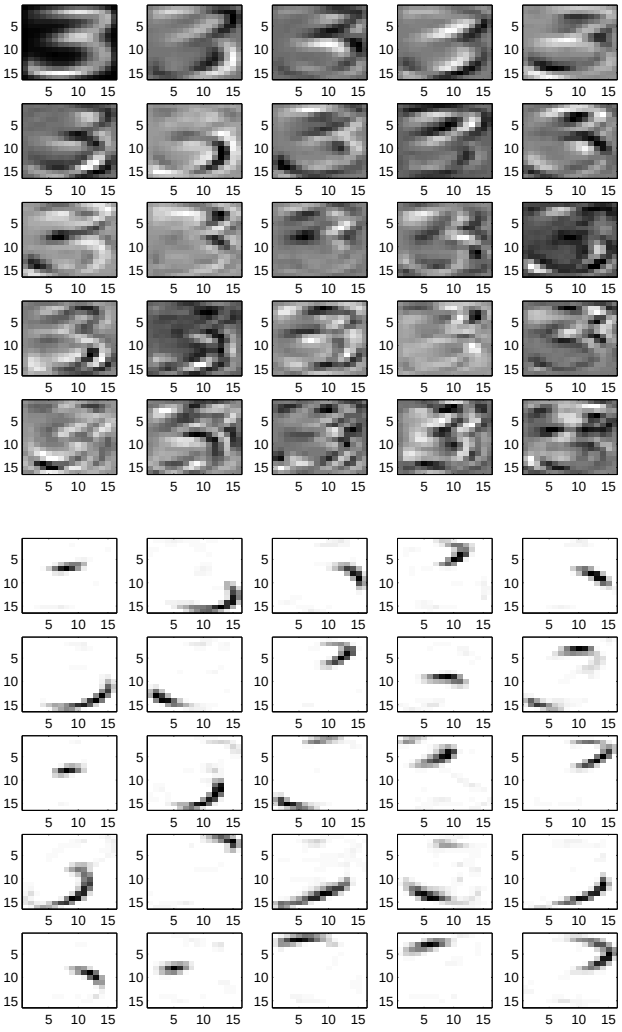

Figure 2: Feature extraction for hand-written digits: The top plot show the first 25 principal components ordered according to eigen values. The buttom plot shows the 25 mean images (sources) for ICA with positive mixing matrix $\mathbf{A}$ and exponential (positive) source prior.