Partially Observed Values

Tapani Raiko
Laboratory of Computer and Information Science
Helsinki University of Technology
FIN-02015 HUT, Espoo, Finland
E-mail: Tapani.Raiko@hut.fi
http://www.cis.hut.fi/projects/bayes/

***Abstract*— It is common to have both observed and missing values in data. This paper concentrates on the case where a value can be somewhere between those two ends, partially observed and partially missing. To achieve that, a method of using evidence nodes in a Bayesian network is studied. Different ways of handling inaccuracies are discussed in examples and the proposed approach is justified in the experiments with real image data. Also, a justification is given for the standard preprocessing step of adding a tiny amount of noise to the data, when a continuous-valued model is used for discrete-valued data.**

I. INTRODUCTION

Most of the data sets collected in real life are not perfect. They contain errors and missing values. There are also cases where some observations are left out on purpose, e.g. not all patients are sent to all laboratory tests. Also, some observations are more accurate or reliable than others. Usually there is some knowledge about these inaccuracies, but it is often ignored in machine learning. Fuzzy logic, on the other hand, is based on modeling inaccuracies.

Bayesian networks [1], [2] are very popular with the artificial intelligence and machine learning communities. They are graphical models [3] where nodes represent random variables and the lack of arcs represents conditional independence assumptions. A complex system is built by combining simpler parts. Traditional Bayesian networks use discrete variables but in this paper, the emphasis is on continuous variables. The experiments are run with Bayes blocks [4] that use variational Bayesian learning. They can handle missing values in a straightforward manner [5].

How to exploit the best features of the Bayesian and the fuzzy frameworks? Wald [6] proved that every admissible decision rule is a Bayes decision rule. Fuzzy logic is just a construction of heuristics, but on the other hand, fuzzy concepts are very intuitive. For instance, the distinction between the concepts *a cup* and *a bowl* is shown in [7] to be vague and context-dependent. Pearl [1] studies so called virtual evidence in Bayesian networks. It means that part of a situation is not carefully modelled but instead some evidence is summarized into virtual evidence. Virtual evidence corresponds essentially to fuzzy observations. This paper shows how virtual evidence can be used with a continuous valued model and what is it good for.

There are numerous approaches to handling missing values [8], [9] and some approaches work even in cases where the missingness of the value can depend on the actual value. But in these textbooks, a value is either observed or missing and there is no option in between. Heitjan and Rubin [10], [11] define coarse data which means that we might observe (no more and no less than) that a data value x belongs to some set, say $x \in [a, b)$. Examples include rounded and out-of-scale measurements. In this case, the value is not entirely missing, since we observe to which set it belongs to. Zhang and Honavar [12] use decision trees with partially specified data. They can specify discrete values at different levels of precision, e.g. the same shape can be described as a polygon in general or a square in specific. These hierarchies are a special case of coarse data.

Coarse data is already quite close to "fuzziness". The gap is closed completely by using a fuzzy membership function $U(x) \in [0, 1]$ as virtual evidence for x, instead of the regular set membership restriction. I will stay in the Bayesian framework and not use fuzzy logic. Section II describes two ways of introducing fuzzy membership functions into Bayesian networks. Section III briefly reviews the variational Bayesian framework for background. Two examples that illustrate different phenomena concerning partially observed values are given in Section IV. Experiments with independent factor analysis on image data are described in Section V. Subsequently, the matters are discussed and concluded.

II. VIRTUAL EVIDENCE FOR CONTINUOUS-VALUED VARIABLES

Figure 1 shows examples of membership functions $U(x)$, which can describe different types of observations: 1) An exact observation that a person is 183 cm tall. 2) A missing observation with no knowledge of the height of this particular person. 3) A coarse observation that the person is taller than 180 cm. 4) Finally, a fuzzy observation that a person is "tall". The common sense of peoples heights (no-one can be 3 meters tall etc.) corresponds to a model or prior experience. The question is, how to combine the knowledge given by the model to the knowledge given by the membership function.

Pearl's virtual evidence [1] can be implemented as follows. Let us consider a Bayesian network and a single value x in it. To make x partially observed, we add a binary node e called an evidence node [13], to it (see Figure 2). The evidence node e has x as the only parent and it has no children. The conditional probability function (cpf) $p(e = 1 \mid x) = U(x)$ is the fuzzy

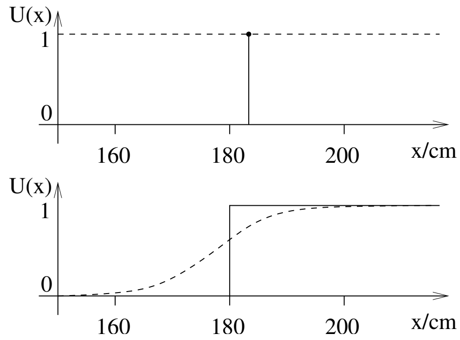

Fig. 1. Different types of observations of a person's height. Top, solid line: observed value, dashed line: missing value. Bottom, solid line: coarse observation, dashed line: fuzzy observation. All of these cases can be interpreted as partially observed values.

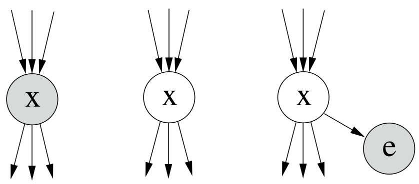

Fig. 2. The node x in a Bayesian network can be either observed (left), missing (middle) or partially observed (right). The node e is called an evidence node. A shaded node represents an observed variable and a white node represents a latent variable.

membership function $U(x)$. Now we leave x latent but observe $e = 1$. This provides evidence for x that corresponds exactly to $U(x)$ and therefore this can be called the *Evidence approach*.

The model $p(x \mid \mathcal{H}, \mathbf{X})$ for x given the model structure $\mathcal{H}$ and the rest of the data $\mathbf{X}$, is combined with the evidence given by $e = 1$. Together they form the posterior distribution

$$\begin{aligned} p(x \mid \mathcal{H}, \mathbf{X}, e = 1) &= \frac{p(x \mid \mathcal{H}, \mathbf{X}) p(e = 1 \mid x, \mathcal{H}, \mathbf{X})}{p(e = 1 \mid \mathcal{H}, \mathbf{X})} \\ &\propto p(x \mid \mathcal{H}, \mathbf{X}) p(e = 1 \mid x) \\ &= p(x \mid \mathcal{H}, \mathbf{X}) U(x). \end{aligned} \tag{1}$$

The partial observation $U(x)$ of x is thus further specified by the model. Note that the marginal likelihood $p(e = 1 \mid \mathcal{H}, \mathbf{X})$ is a constant w.r.t. x and can be thus ignored. The Evidence approach can be thought of as making a noisy observation e about x. The actual value x is then reconstructed during learning by combining prior experience $p(x \mid \mathcal{H}, \mathbf{X})$ with the evidence $U(x)$ from the noisy observation. One should be careful not to include any prior information in $U(x)$ since it would then be taken into account twice. Note also that if $U(x)$ is scaled by a constant, it still produces exactly the same evidence.

Morris et al. [14] define soft missing data by fixing a distribution over each data value: $p(x) \propto U(x)$. A Dirac delta function corresponds to a fully observed value, but unfortunately a very wide function does not approach a fully missing value as will be shown in Section IV-A. In this case, the model cannot further specify the partial observation $U(x)$, since the posterior distribution is fixed to $U(x)$. I call this the *Frozen approach*. It can be thought of as knowing that the true data is distributed in a specific way. This time, all prior information should be included in $U(x)$ but that might be difficult in practice.

Now, let us consider the continuous-valued case and a partial observation that x is probably greater than a constant c. For that, one can use the Evidence approach with a logistic membership function

$$U(x) = \frac{1}{1 + e^{-(x-c)/\alpha}}, \tag{2}$$

where α is a constant that sets the slope or fuzziness of the membership function. This can be implemented with a soft-max node [4] for e with x/α and c/α as parents. Using several different soft-max nodes combined with logical operations, one could build practically arbitrary membership functions. Note that the Frozen approach cannot handle unnormalisable membership functions such as the logistic function.

There are also other ways to produce a virtual evidence for x. One can use for instance the Gaussian evidence node [13]. A partial observation about x is that it is around x_0 with a variance σ^2. The cpf for a continuous-valued evidence node e is defined as $p(e \mid x) = N(e; x, \sigma^2)$. Observing $e = x_0$ changes the posterior distribution of x to

$$\begin{aligned} p(x \mid \mathcal{H}, \mathbf{X}, e = x_0) &= \frac{p(x \mid \mathcal{H}, \mathbf{X}) p(e = x_0 \mid x, \mathcal{H}, \mathbf{X})}{p(e = x_0)} \\ &\propto p(x \mid \mathcal{H}, \mathbf{X}) p(e = x_0 \mid x) \\ &= p(x \mid \mathcal{H}, \mathbf{X}) N(x; x_0, \sigma^2), \end{aligned} \tag{3}$$

corresponding to a Gaussian membership function $U(x) = N(x; x_0, \sigma^2)$. The last step of (3) becomes clear when noticing that the difference $e - x$ is normally distributed. The Frozen approach with a Gaussian distribution is handled simply by fixing $p(x) = N(x; x_0, \sigma^2)$.

III. Variational Bayesian Learning

Variational Bayesian learning techniques are based on approximating the true posterior probability density of the unknown variables of the model by a function with a restricted form. Currently the most common technique is ensemble learning [15], [16], [17], [18] where the Kullback-Leibler divergence measures the misfit between the approximation and the true posterior.

In ensemble learning, the posterior approximation $q(\boldsymbol{\theta})$ of the unknown variables $\boldsymbol{\theta}$ is required to have a suitably factorial form

$$q(\boldsymbol{\theta}) = \prod_i q(\boldsymbol{\theta}_i), \tag{4}$$

where $\boldsymbol{\theta}_i$ denotes a subset of the unknown variables. The misfit between the true posterior $p(\boldsymbol{\theta} \mid \mathbf{X})$ and its approximation $q(\boldsymbol{\theta})$ is measured by the Kullback-Leibler divergence. An additional term $-\ln p(\mathbf{X})$ is included to avoid calculation of the model evidence term $p(\mathbf{X}) = \int p(\mathbf{X}, \boldsymbol{\theta}) d\boldsymbol{\theta}$. The cost function then has the form [19], [15]

$$\begin{aligned}\mathcal{C} &= D(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta}|\mathbf{X})) - \ln p(\mathbf{X}) \\ &= \langle \ln q(\boldsymbol{\theta})\rangle - \langle \ln p(\mathbf{X}, \boldsymbol{\theta})\rangle , \end{aligned} \tag{5}$$

where $\langle \cdot \rangle$ denotes expectation over the distribution $q(\boldsymbol{\theta})$. Note that since $D(q \parallel p) \geq 0$, it follows that the cost function provides a lower bound $p(\mathbf{X}) \geq \exp(-\mathcal{C})$ for the model evidence $p(\mathbf{X})$.

For each update of the posterior approximation $q(\theta_i)$, the variable θ_i requires the prior distribution $p(\theta_i \mid \text{parents})$ given by its parents and the likelihood $p(\text{children} \mid \theta_i, \text{co-parents})$ obtained from its children. The relevant part of the Kullback-Leibler divergence to be minimised is, up to a constant independent of $q(\theta_i)$

$$\mathcal{C}(q(\theta_i)) = \left\langle \ln \frac{q(\theta_i)}{p(\theta_i \mid \text{parents}) p(\text{children} \mid \theta_i, \text{co-parents})} \right\rangle . \tag{6}$$

To make it concrete, let us look at a Gaussian variable node [4] which is a basic building block for a number of models.

A Gaussian variable s has two inputs m and v and a cpf $p(s|m, v) = N(s; m, \exp(-v))$. The variance is parametrised this way because then the mean and expected exponential of v suffice for computing the cost function. It can be shown that when s, m and v are mutually independent a posteriori, i.e. $q(s, m, v) = q(s)q(m)q(v)$, $\mathcal{C}_p(q_s(s)) = -\langle \ln p(s|m, v)\rangle$ yields

$$\begin{aligned}\mathcal{C}_p(q(s)) = \frac{1}{2}\Big\{ \langle \exp v\rangle \Big[(\langle s\rangle - \langle m\rangle)^2 + \mathrm{Var}\,\{m\} + \\ + \mathrm{Var}\,\{s\} \Big] - \langle v\rangle + \ln 2\pi \Big\} .\end{aligned} \tag{7}$$

For observed variables this is the only term in the cost function but for latent variables there is also a term $\mathcal{C}_q$ resulting from $\langle \ln q(s)\rangle$. The posterior approximation $q(s)$ is defined to be Gaussian with mean $\overline{s}$ and variance $\widetilde{s}$: $q(s) = N(s; \overline{s}, \widetilde{s})$. This yields

$$\mathcal{C}_q(q(s)) = -\frac{1}{2} \ln 2\pi e \widetilde{s} \tag{8}$$

which is the negative entropy of a Gaussian variable with variance $\widetilde{s}$. The parameters $\overline{s}$ and $\widetilde{s}$ are optimised during learning.

IV. Phenomena with Partially Observed Values

This Section gives two examples that illustrate interesting phenomena that might occur with partially observed values. Both examples concern Gaussian membership functions. In the first case, the variances are large and a comparison is done to the fully missing value. The second case shows how adding even the tiniest amount of inaccuracy to the data can make a difference by getting rid of degenerate solutions.

A. Wide Membership Functions

Figure 3 depicts an example of two-dimensional (x, y) data for factor analysis. Factor analysis is a version of principal component analysis (PCA) with a noise model. Some of the values $x(t)$ are only partially observed. Their distributions are

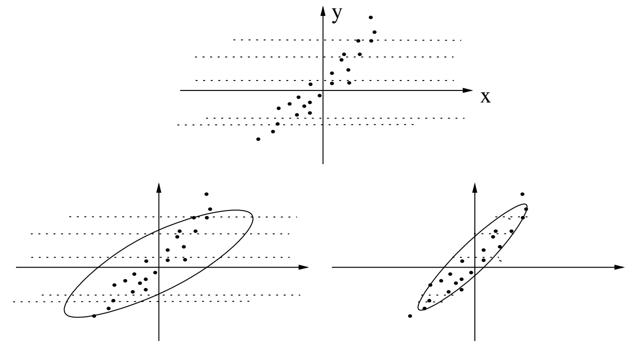

Fig. 3. Some x-values of the data are observed only partially. They are marked with dotted lines representing their confidence intervals. Top: A toy data set for a factor analysis problem. Bottom left: In the Frozen approach, the model needs to adjust to cover the distributions. Bottom right: In the Evidence approach, the partially observed values are reconstructed based on the model.

Gaussians with fairly large variances that are assumed to be known.

The Frozen approach (see Section I) assumes that the data is really distributed according to the membership function $U(x(t)) = N(x(t); \overline{x}(t), \widetilde{x}(t))$. Therefore, the model has to cover the whole distributions. In the Evidence approach, on the other hand, the posterior distribution (Eq. 1) of the partially observed values can be adjusted based on the model. Figure 3 shows the (hypothetical) situation after learning. The Frozen approach is disturbed by the partially observed values, whereas the Evidence approach reconstructs them based on the rest of the data.

When the variance $\widetilde{x}(t)$ of a Gaussian membership function goes to infinity, $U(x(t))$ is constant in any finite set. In the Evidence approach, the constant evidence corresponds to a (fully) missing value. To see what happens in the Frozen approach, one can write down the sample variance of the x-component over the data set

$$\mathrm{Var}\{x\} = \frac{1}{T-1} \sum_{t=1}^{T} \left[(\overline{x}(t) - \mathrm{E}\{x\})^2 + \widetilde{x}(t) \right] . \tag{9}$$

The model has to adjust to account for the variance in the data. When any $\widetilde{x}(t) \to \infty$, also the whole sample variance $\mathrm{Var}\{x\} \to \infty$. That is, the learning will lead to a degenerate solution in which the model for x is unreasonably wide.

B. Narrow Membership Functions

Let us think about an example of a one-dimensional mixture-of-Gaussians model for data. In case there are T data samples $x(1), \ldots, x(T)$ exactly at the same point, a Gaussian cluster with a mean $m = x(1) = \cdots = x(T)$ might specialise in those samples with a tiny variance σ^2. Ignoring the rest of the clusters and data samples, the essential likelihood factor is proportional to T/σ. When the cluster gets narrower, $\sigma \to 0$, the posterior density $p(m, \sigma \mid \mathcal{H}, \mathbf{X}) \to \infty$. That is, the solution is degenerate but it gets an infinitely good score. Note that the problem occurs even in case $T = 1$, that is, when nothing is assumed about the data.

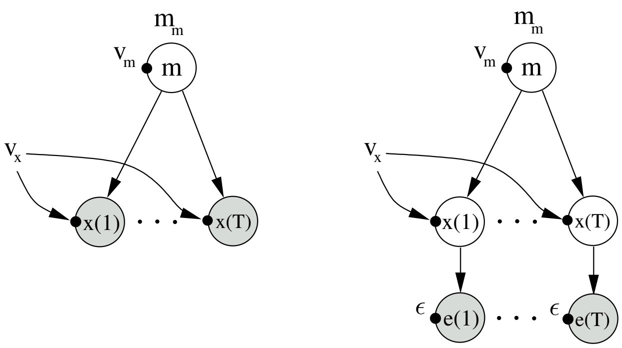

Fig. 4. A model structure representing a single Gaussian cluster with mean m containing data samples $x(1), \ldots, x(T)$. On the left, the data points are fully observed and on the right, only partially observed. The dark dot at the side of a node represents the variance input.

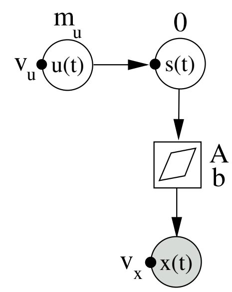

Fig. 5. The model structure used for the experiments. Each node corresponds to a matrix of variables. Variance sources u are used for making the sources s super-Gaussian. The square node represents an affine transformation with a weight matrix $\mathbf{A}$ and a bias vector $\mathbf{b}$. Hierarchical priors are hidden for clarity.

The problem is not that serious when variational Bayesian learning is used instead. Figure 4 depicts the model structure. The cluster mean has a cpf $N(m; m_m, \exp(-v_m))$ and a posterior $q(m) = N(m; \overline{m}, \widetilde{m})$. The cpfs for the data variables $x(t)$ are $N(x(t); m, \exp(-v_x))$. The essential terms of the cost function from Equations (7) and (8) are

$$\begin{aligned} \mathcal{C}(q(x,m)) &= \frac{T}{2}\left(\langle \exp v_x \rangle \widetilde{m} - \langle v_x \rangle\right) \\ &\quad + \frac{1}{2}\left(\langle \exp v_m \rangle \widetilde{m} - \ln \widetilde{m}\right). \end{aligned} \tag{10}$$

Solving the $\widetilde{m}$ to minimize $\mathcal{C}(q(x,m))$ gives

$$\widetilde{m} = \frac{1}{T \langle \exp v_x \rangle + \langle \exp v_m \rangle} \tag{11}$$

which is substituted back into (10) to give

$$\begin{aligned} &\mathcal{C}(q(x,m)) = \\ &\quad \frac{1}{2}\left[1 + \ln\left(T \langle \exp v_x \rangle + \langle \exp v_m \rangle\right) - T \langle v_x \rangle\right]. \end{aligned} \tag{12}$$

In case $T > 1$, when v_x goes to infinity (corresponding to $\sigma^2 \to 0$), the cost goes to negative infinity. This means that a similar degenerate solution, that is rated infinitely good, exists in case $T > 1$.

Let us then assume that the data samples are not exactly observed. Instead, they have a Gaussian membership function with a variance $\epsilon^2 > 0$. The likelihood term at x does not change which means that maximum a posteriori learning is still prone to the same problem. Variational Bayesian learning, on the other hand, gets rid of the problem even in cases $T > 1$. Figure 4 depicts the model structure with evidence nodes. The posterior of $x(t)$ is $q(x(t)) = N(x(t); \overline{x}(t), \widetilde{x}(t))$ and the cpf of an evidence node $e(t)$ is $p(e(t) \mid x(t)) = N(e(t); x(t), \epsilon^2)$. Variances $\widetilde{m}$ and $\widetilde{x}(t)$ can be solved like in (11) and the resulting cost is similar to (12) with an additional term $(T/2)\ln(\langle \exp v_x \rangle + \epsilon^{-2})$. Now the cost approaches positive infinity when $v_x \to \infty$ and thus the degenerate solution no longer exists.

An interpretation of the situation follows. When using a point estimate for the cluster mean m, the cluster can be made infinitely narrow with no cost. In variational Bayesian learning, describing the cluster mean m with a great accuracy shows up in the cost. In case there is just one data sample $x(1)$ in the cluster, the advantage in cost is similar to the cost that went into describing m well. When $T > 1$, the advantage is T-fold and thus the degenerate solution seems infinitely good. The “happy surprise” that the data points $x(1), \ldots, x(T)$ collide is as great at all levels of accuracy. But when an explicit inaccuracy of ϵ is introduced, the surprise of data points colliding is limited to the level of accuracy ϵ. An information theoretic point of view [20] to the situation is enlightening.

V. Experiments

A model structure that implements Independent factor analysis (IFA) [16], is depicted in Figure 5 and used for the experiments. The data vectors $\mathbf{x}(t)$ are assumed to be generated from unknown sources $\mathbf{s}(t)$ through an unknown linear mapping with noise

$$p(\mathbf{x}(t) \mid \cdot) = N(\mathbf{x}(t); \mathbf{A}\mathbf{s}(t) + \mathbf{b}, \mathrm{diag}(\exp(-\mathbf{v}_x))), \tag{13}$$

where $\mathrm{diag}(\exp(-\mathbf{v}_x)))$ is a diagonal covariance matrix with values $\exp$ applied componentwise to the vector $-\mathbf{v}_x$, on the diagonal. The sources $\mathbf{s}(t)$ have a zero-mean super-Gaussian distribution generated as a Gaussian with a varying variance:

$$p(\mathbf{s}(t) \mid \mathbf{u}(t)) = N(\mathbf{s}(t); \mathbf{0}, \mathrm{diag}(\exp(-\mathbf{u}(t)))). \tag{14}$$

The variables $\mathbf{A}$, $\mathbf{b}$, and $\mathbf{u}(t)$ have hierarchical priors [9]. The prior of $\mathbf{A}$ is sparse (mixture of a Gaussian and a delta function at zero) and the other priors are Gaussians.

The model is initialised randomly and learned using variational Bayesian learning. The learning scheme is designed to minimise the cost function $\mathcal{C}$ in Equation (5) by iterative updates, by addition and pruning of weights, and by line search. More details can be found in [21].

The first experiment is a comparison of different ways to reconstruct corrupted values, when exact knowledge of the corruption is available. The second experiment shows a situation where the learning diverges towards a degenerate solution. Solution to avoid the problem is given.

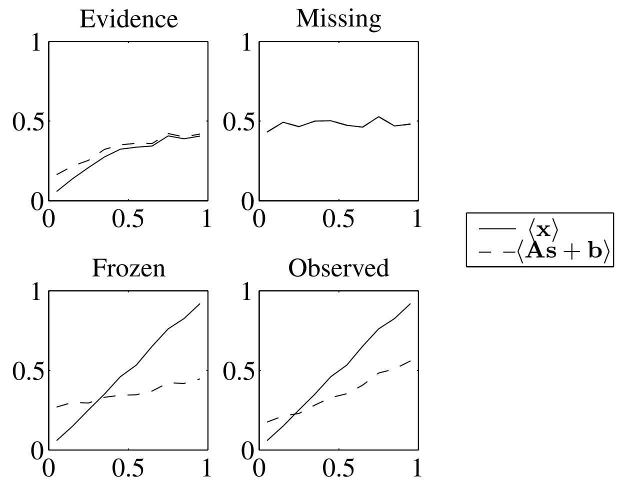

Fig. 6. Reconstruction error as a function of the amount of corruption (std).

A. Reconstruction

The first data set consists of 13 different gray-scale natural images. 1000 samples of 10-by-10-pixel patches are chosen randomly. The patches are normalised to zero mean and unit variance. Each pixel has a 10% chance of being corrupted by a Gaussian noise with a standard deviation (std) that is evenly distributed from 0 to 1. The amount of corruption is assumed to be known. That is, in addition to the data $\mathbf{x}(t)$, the stds $\mathbf{v}_e(t)$ are known. The ICA model initialised with 100 sources is learned for 1000 sweeps through the data in four different settings:

- Evidence: Evidence approach as defined in Section II. The corrupted data values ($v_{e,i}(t) > 0$) are marked missing and Gaussian evidence nodes (Eq. 3) are attached to them: $p(e_i(t) \mid x_i(t)) = N(e_i(t); x_i(t), v_{e,i}(t))$.
- Missing: Corrupted values are discarded and treated as missing values.
- Frozen: Frozen approach as defined in Section II. A Gaussian distribution with the given mean and std is fixed over each corrupted data value.
- Observed: The knowledge about corruption is discarded and values are treated as observed values.

The following table shows the root mean square errors for the reconstruction of corrupted values in different settings:

	Evidence	Missing	Frozen	Observed
$\langle \mathbf{x} \rangle$	0.31	0.48	0.57	0.57
$\langle \mathbf{As}+\mathbf{b} \rangle$	0.34	0.48	0.36	0.38

Both the expectation over the posterior distribution of data $\langle \mathbf{x}(t) \rangle$ and the conditional probability $\langle p(\mathbf{x}(t) \mid \mathbf{A}, \mathbf{s}(t), \mathbf{b}) \rangle = \langle \mathbf{As}(t) + \mathbf{b} \rangle$ are presented, because the Frozen and the Observed settings have the corrupted data directly as $\langle \mathbf{x}(t) \rangle$ (the result 0.57 is the corruption level). The same results are separated into 10 different levels of corruption and shown as curves in Figure 6. Note that the optimal constant prediction 0 gives the reconstruction error 1 since the data is normalised.

The following observations can be made:

- Evidence: As expected, the Evidence approach was the best way of reconstructing corrupted values at all corruption levels. Small corruption leads to accurate reconstructions and as the corruption level increases, the Evidence setting approaches the Missing setting.
- Missing: The data posterior $\langle \mathbf{x} \rangle$ is the same as $\langle \mathbf{As}+\mathbf{b} \rangle$. The reconstructions are independent of the corruption level since all the corrupted values were discarded. The discarded information was so important that the reconstructions were the worst.
- Frozen: Reconstructions are the second best overall. One would still need to justify when and why to use the reconstructions given by $\langle \mathbf{As}+\mathbf{b} \rangle$ and not by $\langle \mathbf{x} \rangle$. If the corruption level increases further, the reconstructions become worse than those of the Missing setting.
- Observed: Ignoring the corruption mechanism gives the second worst results. Reconstruction accuracy depends much on the corruption level.

B. Problem with noiseless data

The second data set consists of 13 different diagram-like images. They have discrete gray-scale values from 0 to 255 even though mostly they are black and white. Setting 1 has no added noise, whereas in Setting 2, a tiny amount of Gaussian noise with a standard deviation of 0.1 is added to the images. After that, figures are normalised to zero-mean and unit variance. 1000 samples of 6 by 6 image patches are chosen randomly. The same ICA-model is used, this time initialised with an over-complete basis of 50 sources.

Figure 7 shows the learning curves for the first 100 sweeps through the data. In the beginning, the two settings behave similarly, but after 45 sweeps they start to differ. After 100 sweeps, the modelled variance of the data is of the order 10^{-28} in Setting 1 and the learning is becoming unstable for numerical reasons. The same phenomenon as explained in Section IV-B is applying. The learning is diverging towards a degenerate solution that is rated infinitely good. Setting 2 is stable, even though the original difference in the two settings was very small.

The problem of a degenerate solution is often encountered when variances are modelled. As explained in Section IV-B, the problem is not as serious when using variational Bayesian learning as when using point estimates, but it still exists. The solution is to add a tiny amount of noise to the data. Whether it is done by explicitly sampling noise using a random number generator or adding the noise implicitly using either the Evidence or the Frozen approach, makes no real difference in results. Explicit sampling is usually the simplest and computationally lightest so it has become the standard.

VI. Discussion

Some real-world applications for partially observed values could be brought from the fuzzy logic community to machine learning community. Perhaps the most promising option is to find some clinical data which would contain information about

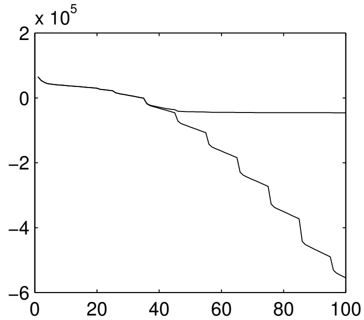

Fig. 7. The behavior of the cost function $\mathcal{C}$ during learning. The diverging lower curve corresponds to no added noise and the upper curve to a tiny amount of added noise. The regular fluctuation is expected, it reflects our learning process where every tenth iteration is done in a different manner.

the inaccuracies. Morris [14] studied speech recognition with soft missing data.

Often, it is known that the data set contains errors, but it is not known which values are erroneous. This could be modelled as evidence of evidence. The first evidence node would be left latent and its posterior distribution would tell the probability of the corresponding value to be correct or not. The second evidence node would be observed and it would give a membership function for the first evidence, and through that, some likelihood factor for the actual data value, too. It would be easier to find data for this kind of a model, since it does not require explicit knowledge of individual errors. Applications for outlier detection [22] are already well known.

Variational Bayesian learning is prone to local minima so tricks to avoid them during learning are useful. The Gaussian evidence node was first used in [13] to keep parts of the network fixed to initial values until the other parts have adapted appropriately. The width of the Gaussian evidence was increased after each iteration until the whole node was removed. The persistence of the initialisation could be thus controlled accurately.

VII. Conclusion

Partially observed values fill the gap between observed and missing values in data. A distinction is made between fixing a distribution over a data value (the Frozen approach) and getting evidence about the data value through a noisy observation (the Evidence approach). Only the Evidence approach has a missing value as a limit case. It can be implemented by adding an extra node to a Bayesian network for each partially observed value.

Experiments with natural image data and an IFA model with variational Bayesian learning show that making use of the knowledge about inaccuracies pays off. Also, a problem with applying continuous-valued models to discrete data is solved by using variational Bayesian learning combined with a tiny amount of additional noise to the data.

Acknowledgment

The author would like to thank Markus Harva, Antti Honkela, Alexander Ilin, Juha Karhunen, Erkki Oja, Jan-Hendrik Schleimer, and Harri Valpola for useful discussions. This work was supported by the Finnish Centre of Excellence Programme (2000-2005) under the project New Information Processing Principles.

References

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

[2] F. Jensen, *Bayesian Networks and Decision Graphs*. New York: Springer, 2001.

[3] K. Murphy, "An introduction to graphical models," Intel Research Technical Report, Tech. Rep., 2001.

[4] H. Valpola, T. Raiko, and J. Karhunen, "Building blocks for hierarchical latent variable models," in *Proc. 3rd Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego, USA, 2001, pp. 710–715.

[5] T. Raiko, H. Valpola, T. Östman, and J. Karhunen, "Missing values in hierarchical nonlinear factor analysis," in *Proc. of the Int. Conf. on Artificial Neural Networks and Neural Information Processing, ICANN/ICONIP 2003*, Istanbul, Turkey, 2003, pp. 185–189.

[6] A. Wald, *Statistical Decision Functions*. New York: John Wiley & Sons, 1950.

[7] W. Labov, "The boundaries of words and their meaning," in *New ways of analyzing variation of English*, J. Fishman, Ed. Georgetown Press, 1973, pp. 340–373.

[8] R. Little and D.B.Rubin, *Statistical Analysis with Missing Data*. J. Wiley & Sons, 1987.

[9] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis*. Boca Raton, Florida: Chapman & Hall/CRC Press, 1995.

[10] D. F. Heitjan and D. B. Rubin, "Inference from coarse data via multiple imputation with application to age heaping," *Journal of the American Statistical Association*, pp. 304–314, 1990.

[11] ——, "Ignorability and coarse data," *The Annals of Statistics*, pp. 2244–2253, 1991.

[12] J. Zhang and V. Honavar, "Learning from attribute value taxonomies and partially specified instances," in *Proc. of the 20th International Conference on Machine Learning (ICML-2003)*, 2003, pp. 880–887.

[13] H. Valpola, T. Östman, and J. Karhunen, "Nonlinear independent factor analysis by hierarchical models," in *Proc. 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, 2003, pp. 257–262.

[14] A. Morris, J. Barker, and H. Bourlard, "From missing data to maybe useful data: soft data modelling for noise robust ASR," IDIAP, IDIAP-RR 06, 2001.

[15] H. Lappalainen and J. Miskin, "Ensemble learning," in *Advances in Independent Component Analysis*, M. Girolami, Ed. Berlin: Springer-Verlag, 2000, pp. 75–92.

[16] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, no. 4, pp. 803–851, 1999.

[17] J. Miskin and D. J. C. MacKay, "Ensemble learning for blind source separation," in *Independent Component Analysis: Principles and Practice*, S. Roberts and R. Everson, Eds. Cambridge University Press, 2001, pp. 209–233.

[18] H. Valpola, E. Oja, A. Ilin, A. Honkela, and J. Karhunen, "Nonlinear blind source separation by variational Bayesian learning," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E86-A, no. 3, pp. 532–541, 2003.

[19] D. Barber and C. Bishop, "Ensemble learning in Bayesian neural networks," in *Neural Networks and Machine Learning*, C. Bishop, Ed. Berlin: Springer, 1998, pp. 215–237.

[20] A. Honkela and H. Valpola, "Variational learning and bits-back coding: an information-theoretic view to Bayesian learning," *IEEE Trans. on Neural Networks*, 2004, to appear.

[21] H. Valpola, M. Harva, and J. Karhunen, "Hierarchical models of variance sources," in *Proc. 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, 2003, pp. 83–88.

[22] V. Barnett and T. Lewis, *Outliers in Statistical Data*. New York: John Wiley and Sons, 1994.