Variational Bayesian methods

Release 0.1-dev

Jaakko Luttinen

April 16, 2015

Contents

Contents:

CHAPTER 1

Introduction

Todo

Testing TODO.

CHAPTER 2

Bayesian inference

The Bayesian framework provides a principled way to model and analyze data. The framework uses probabilities to represent the knowledge of the modelled process and the unknown quantities. Thus, simple rules of probability theory can be used for inference. The same basic rules are used regardless of the complexity or the application field of the problem. The beauty of the Bayesian framework is that it can be derived from simple axioms as a unique way of doing rational reasoning.

Bayesian modelling has several advantages over ad hoc approaches: 1) The probabilities account for the uncertainty in the results. 2) Missing values are usually not a problem because the whole framework is about incomplete knowledge. 3) Model comparison can be done in a principled way. 4) Overfitting is prevented by combining many models and taking into account their complexities. 5) Modelling assumptions and priors are expressed explicitly and can be modified. 6) Existing models can be straightforwardly modified, extended or used as building blocks for more complex models.

This chapter gives a brief introduction to Bayesian modelling. Section *Probability Theory* summarizes the foundation by explaining how the probability theory can be interpreted as a unique system of consistent rational reasoning under uncertainty. Section *inference* shows how this theory is applied in Bayesian modelling.

2.1 Probability Theory

The Bayesian framework is based on probability theory by interpreting probabilities as plausibility assignments. This differs from the frequentist approach, which interprets probabilities as frequencies in repeated experiments. The Bayesian framework uses probability theory as an extension of logic to handle uncertain propositions which do not need to be related to random events or repeated experiments. Thus, the rules of the probability theory can be applied to a wide range of problems involving incomplete knowledge instead of random events or repeated experiments.

The Bayesian interpretation of probability can be derived from desired qualitative properties for rational reasoning [12][5]. The idea is that propositions have subjective plausibilities and the rules for handling the plausibility assignments should have rational properties. The desired properties of rational reasoning can be roughly summarized as follows:

- *Comparability:* degrees of plausibility can be compared and are represented by real numbers.
- *Continuity:* an infinitesimally greater plausibility corresponds to an infinitesimally greater number.
- *Logicality:* rules are consistent with Aristotelian logic.
- *Rationality:* rules have qualitative correspondence with weak syllogisms.[#weak-syllogisms]_
- *Consistency:* every possible way of reasoning must lead to the same result and equivalent plausibilities are represented by equal numbers.
- *Neutrality:* all relevant evidence is taken into account without ignoring any information.

The list is a slightly rephrased and simplified version of the list presented by [12].

From the properties of rational reasoning, one can derive a unique set of quantitative rules. Omitting the long and rigorous derivations [12], the resulting rules are the well-known product rule

$$p(A, B) = p(A|B)p(B) = p(B|A)p(A) \tag{2.1}$$

and the sum rule

$$p(A) + p(\overline{A}) = 1,$$

where A and B are propositions, and $\overline{A}$ is the complement of A. The probabilities $p(\cdot)$ represent the state of knowledge, where certainty is represented by 1 and impossibility by 0. Therefore, applying probability theory to inference problems means that one uses common sense consistently.

From the product rule (2.1), it follows that

$$p(A|B) = \frac{p(A, B)}{p(B)} = \frac{p(B|A)p(A)}{p(B)}, \tag{2.2}$$

which is the Bayes' theorem. This can be seen as a formula for updating the beliefs about A after given new evidence B. Thus, the properties of rational reasoning determine how we should rationally change the beliefs we have when given new evidence. However, note that the rules do not determine which beliefs are a priori rational. kk

2.2 Inference

In Bayesian modelling, the probability theory provides tools for constructing generative models for data and obtaining knowledge about the models given some data (see, e.g., [4][2][16]). This information can be used to get insight into the data and to make predictions. A generative model $\mathcal{M}$ consists of a likelihood function $p(Y|Z, \mathcal{M})$ explaining the data Y with parameters Z and a prior function $p(Z|\mathcal{M})$ providing the prior knowledge about the model parameters. The goal is to find the posterior distribution of the model parameters:

$$p(Z|Y, \mathcal{M}) = \frac{p(Y|Z, \mathcal{M})p(Z|\mathcal{M})}{p(Y|\mathcal{M})},$$

which can be used, for instance, to make predictions. The denominator $p(Y|\mathcal{M})$ is called the marginal likelihood, defined as

$$p(Y|\mathcal{M}) = \int p(Y|Z, \mathcal{M})p(Z|\mathcal{M})dZ,$$

which is the probability (density) of the observations when the model $\mathcal{M}$ is assumed to be true. Typically, the conditioning on the model is not explicitly shown if there is no risk of misunderstanding. Thus, we discard $\mathcal{M}$ from our notation.

Models usually have hierarchical structure, which means that the prior of a set of unknown variables is defined in terms of another set of unknown variables. This may lead to extremely complex posterior inference unless priors have convenient forms to simplify calculations. In particular, the prior for an unknown variable can be chosen such that the resulting posterior distribution conditioned on all other unknown variables is in the same family as the prior. This type of prior distribution is called a conjugate prior for the likelihood. In addition, if the distributions are from the exponential family, the model is said to be from the conjugate exponential family.

The main challenge in Bayesian inference is that the posterior distribution (2.2) is often analytically intractable. Therefore, one has to resort to methods that approximate the posterior. These methods can roughly be divided into two categories: deterministic and stochastic techniques ([4]). Both of these techniques have their advantages and disadvantages.

Deterministic methods use analytic approximations to the posterior. The resulting approximate distribution is often evaluated efficiently, but it usually requires extra work because some formulas must be derived analytically. The approximate distribution does not, in general, recover the true posterior distribution exactly. Important deterministic approximations include: maximum likelihood and maximum a posteriori methods, which approximate the posterior distribution with a point estimate; Laplace method, which fits a Gaussian distribution to a mode of the posterior probability density function; variational Bayes ([13]) and expectation propagation ([15]), which find an approximate distribution by minimizing an information-theoretic dissimilarity to the true distribution; and integrated nested Laplace approximations for latent Gaussian models ([17]).

Stochastic techniques approximate the posterior distribution with a finite number of samples. The samples from the intractable posterior may be obtained in several ways depending on the problem. These stochastic techniques are covered comprehensively, for instance, in the book by [7]. In complex problems, sampling is often implemented with random-walk type algorithms, called Markov chain Monte Carlo (MCMC). In general, stochastic methods have the property that the approximation approaches the true posterior at the limit of infinite computation time. However, for large and complex problems, the convergence can be extremely slow.

CHAPTER 3

Variational Bayesian approximation

3.1 Minimizing Kullback-Leibler divergence

- lower bound

3.2 Mean-field, fixed form

3.3 Variational Bayesian expectation maximization algorithm

(conjugate exponential family models)

Todo

Split and fix the text below into the sections

In variational Bayesian (VB) methods, the idea is to find an approximate distribution $q(Z)$ which is close to the true posterior distribution $p(Z|Y)$ (see, e.g., [13][1][4][6]). The dissimilarity is defined as the Kullback-Leibler (KL) divergence of $p(Z|Y)$ from $q(Z)$:

$$\mathrm{KL}(q||p) = -\int q(Z) \log \frac{p(Z|Y)}{q(Z)} \mathrm{d}Z.$$

The divergence is always nonnegative and zero only when $q(Z) = p(Z|Y)$. However, the divergence cannot typically be evaluated because the true posterior distribution is intractable.

The key idea in VB is that the divergence can be minimized indirectly by maximizing another function which is tractable. This function is a lower bound of the log marginal likelihood. It can be found by decomposing the log

marginal likelihood as

$$\begin{aligned}
\log p(Y) &= \int q(Z) \log p(Y) \mathrm{d}Z \\
&= \int q(Z) \log \frac{p(Y,Z)}{p(Z|Y)} \mathrm{d}Z \\
&= \int q(Z) \log \frac{p(Y,Z) q(Z)}{p(Z|Y) q(Z)} \mathrm{d}Z \\
&= \int q(Z) \log \frac{p(Y,Z)}{q(Z)} \mathrm{d}Z - \int q(Z) \log \frac{p(Z|Y)}{q(Z)} \mathrm{d}Z \\
&= \int q(Z) \log \frac{p(Y,Z)}{q(Z)} \mathrm{d}Z + \mathrm{KL}(q \| p) \\
&\geq \int q(Z) \log \frac{p(Y,Z)}{q(Z)} \mathrm{d}Z \\
&\equiv \mathcal{L}(q).
\end{aligned}$$

Because the KL divergence is always non-negative, $\mathcal{L}(q)$ is a lower bound for $\log p(Y)$. Furthermore, because the sum of $\mathcal{L}(q)$ and $\mathrm{KL}(q\|p)$ is constant with respect to q, maximizing $\mathcal{L}(q)$ is equivalent to minimizing $\mathrm{KL}(q\|p)$. In some cases, even $\mathcal{L}(q)$ may be intractable and further approximations are needed to find a tractable lower bound.

Thus far, there is nothing approximate in the procedure, because the optimal solution which minimizes the divergence is the true posterior distribution $q(Z) = p(Z|Y)$. In order to find a tractable solution, the range of functions needs to be restricted in some way. However, the range must be as rich and flexible as possible in order to find as good an approximation as possible. This can be achieved by assuming a fixed functional or factorial form for the distribution.

This work restricts the class of approximate distributions by assuming that the q distribution factorizes with respect to some grouping of the variables:

$$q(Z) = \prod_{m=1}^{M} q_m(Z_m), \tag{3.1}$$

where $Z_1, \ldots, Z_M$ form a partition of Z. The notation is kept less cluttered by ignoring the subscript on q. The lower bound $\mathcal{L}(q)$ can be maximized with respect to one factor at a time. The optimal factor $q(Z_m)$ can be found by inserting the approximate distribution eqref{eq:VB_factorized} to $\mathcal{L}(q)$ and maximizing $\mathcal{L}(q)$ with respect to $q(Z_m)$. This yields

$$q(Z_m) = \exp\left(\langle \log p(Y,Z) \rangle_{\setminus m} \right), \tag{3.2}$$

where the expectation is taken over all the other factors except $q(Z_m)$. An iterative update of the factors until convergence is called the variational Bayesian expectation maximization (VB-EM) algorithm ([1][3]). Alternatively, it is also possible to use, for instance, gradient-based optimization methods to optimize the parameters of the approximate q distributions (see, e.g., [11] or stochastic variational inference ([10]) in order to improve scaling to large datasets with stochastic optimization.

Fig. 3.1 illustrates typical effects of the factorizing VB approximation: only one mode of the true posterior is captured, dependencies between variables are lost and marginal distributions are too narrow.

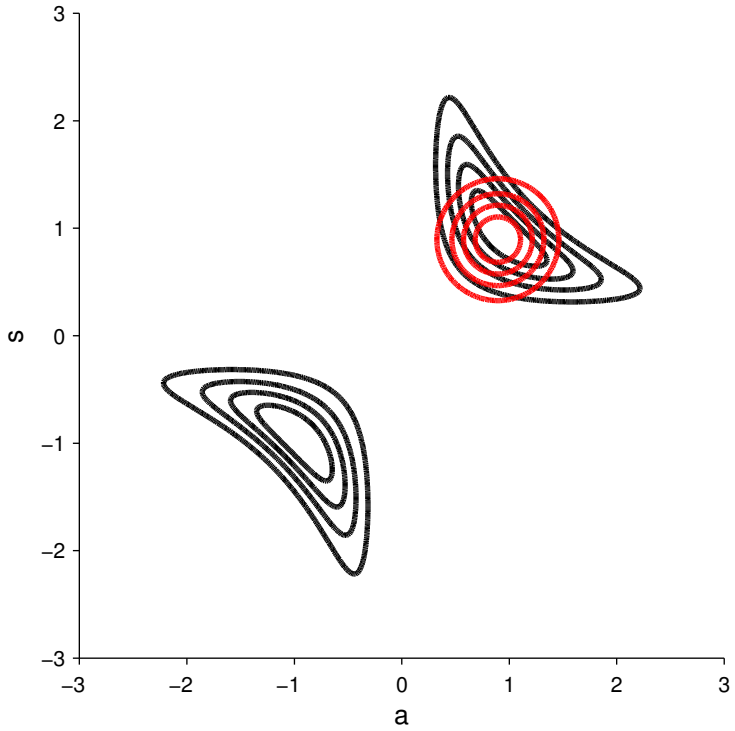

Figure 3.1: An illustration of typical effects of the factorizing approximation. A true posterior (in black) and the optimal factorizing VB posterior (in red).

CHAPTER 4

Variational message passing

If the model is in the conjugate exponential family, the VB-EM algorithm can be implemented as the variational message passing (VMP) algorithm ([19]). The algorithm is based on local computations, in which each factor is represented by a node. When a node is updated, it receives messages from its children and parents and uses those messages to compute the new approximate distribution and relevant expectations. The advantage of this message passing formulation is that the computations are local and depend on well-defined messages. This makes it easy to modify the model by adding or changing nodes.

CHAPTER 5

Stochastic variational inference

Stochastic variational inference ([10]) learns shared latent variables with stochastic optimization. It uses subsets of the data to compute noisy estimates of the gradients. The method can be used to scale the inference on large datasets, if the model has a specific structure.

CHAPTER 6

Non-conjugate methods

6.1 Lower bounding

- logistic regression

6.2 Numerical integration

numerical integration of expectations

- logistic regression

6.3 Conjugate exponential family interpretation

- Student-t

6.4 Tilted VB

It is also possible to extend VB inference to a wide range of non-conjugate models (see, e.g., [14][9]).

6.5 Other methods

In later chapters:

- black box variational inference
- inference or recognition model
- Riemannian conjugate learning

CHAPTER 7

Riemannian conjugate gradient learning

7.1 Riemannian manifold

- basics about Riemannian manifold

7.2 Algorithm

- riemannian conjugate gradient learning

CHAPTER 8

Improving optimization

CHAPTER 9

Gaussian processes

9.1 Gaussian process regression

Rasmussen:2006

9.2 Variational sparse approximation

variational sparse approximation Titsias:2009

9.3 Uncertain inputs

9.4 Stochastic inference

“GPs for big data”

9.5 Variational approximation of hyperparameters? Elsewhere?

VB approximation for hyperparameters (Titsias:2014) or maybe this elsewhere

9.6 GP-LVM?

GP-LVM? does this contain something interesting and specific to VB?

9.7 Deep GPs?

again: does this contain something interesting and specific to VB?

CHAPTER 10

Markov models

10.1 Smoothing in HMM

10.2 Smoothing in LSSM

10.3 Gaussian Markov random fields

CHAPTER 11

General black-box framework

11.1 "Black box"

In order to have minimal amount of model specific computations, one can use black box variational inference ([Paisley:2013][Ranganath:2014]). It uses stochastic optimization and computes noisy gradients of the VB lower bound by sampling from the approximate posterior distribution $q(Z)$ to estimate the relevant expectations. In principle, the method can be applied to any model for which the (unnormalized) joint density $p(Y, Z)$ can be computed.

11.2 Variational approximation as linear regression

[Salimans:2013]

11.3 Gradient-based approximation

Titsias:2014

continuous variable and differentiable density

CHAPTER 12

Inference model

- inference or recognition model
- deep learning / neural networks

Bibliography

[1] Hagai Attias. A variational Bayesian framework for graphical models. In Sara A. Solla, Todd K. Leen, and Klaus-Robert Müller, editors, *Advances in Neural Information Processing Systems 12*, 209–215. Denver, Colorado, USA, 2000.

[2] David Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

[3] Matthew J. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[4] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2nd edition, 2006.

[5] Richard T. Cox. *The Algebra of Probable Inference*. Johns Hopkins University Press, 1961.

[6] Charles W. Fox and Stephen J. Roberts. A tutorial on variational Bayesian inference. *Artificial Intelligence Review*, 38(2):85–95, jun 2012. doi:10.1007/s10462-011-9236-8.

[7] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Texts in Statistical Science. Chapman and Hall/CRC, Florida, 2nd edition, 2003.

[8] James Hensman, Magnus Rattray, and Neil D. Lawrence. Fast variational inference in the conjugate exponential family. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, 2888–2896. Lake Tahoe, Nevada, USA, 2012.

[9] James Hensman, Max Zwiessele, and Neil Lawrence. Tilted variational Bayes. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, 356–364. Reykjavik, Iceland, apr 2014.

[10] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–47, 2013.

[11] Antti Honkela, Tapani Raiko, Mikael Kuusela, Matti Tornio, and Juha Karhunen. Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes. *Journal of Machine Learning Research*, 11:3235–3268, 2010.

[12] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[13] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical methods. In *Machine Learning*, 183–233. MIT Press, 1998.

[14] David A. Knowles and Tom Minka. Non-conjugate variational message passing for multinomial and binary regression. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, 1701–1709. Granada, Spain, 2011.

[15] Thomas Minka. Expectation propagation for approximate Bayesian inference. In Jack S. Breese and Daphne Koller, editors, *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, 362–369. Seattle, Washington, USA, aug 2001.

[16] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[17] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B*, 71(2):319–392, apr 2009.

[18] Yee W. Teh, David Newman, and Max Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In Bernhard Schölkopf, John Platt, and Thomas Hoffman, editors, *Advances in Neural Information Processing Systems 19*, 1353–1360. Vancouver, Canada, 2007.

[19] John Winn and Christopher M. Bishop. Variational message passing. *Journal of Machine Learning Research*, 6:661–694, 2005.