Helsinki University of Technology
Publications in Computer and Information Science

Report E6
September 2006

HYPERPARAMETER ADAPTATION IN VARIATIONAL BAYES FOR THE GAMMA DISTRIBUTION

Harri Valpola Antti Honkela

TEKNILLINEN KORKEAKOULU
TEKNISKA HÖGSKOLAN
HELSINKI UNIVERSITY OF TECHNOLOGY
TECHNISCHE UNIVERSITÄT HELSINKI
UNIVERSITE DE TECHNOLOGIE D'HELSINKI

Distribution:
Helsinki University of Technology
Department of Computer Science and Engineering
Laboratory of Computer and Information Science
P.O. Box 5400
FI-02015 TKK, Finland
Tel. +358-9-451 3267
Fax +358-9-451 3277

This report is downloadable at
http://www.cis.hut.fi/Publications/

ISBN 951-22-8396-4
ISSN 1796-2803

Hyperparameter Adaptation in Variational Bayes for the Gamma Distribution

Harri Valpola and Antti Honkela

September 14, 2006

Abstract

Gamma distribution is often used as a prior for the precision (inverse variance) of the Gaussian distribution as it is the conjugate prior. If the scale of the underlying Gaussian variables is not known *a priori*, it is sensible to use an empirical Bayes approach to adapt the parameters of the prior. In this note we present a highly convergent fixed point iteration for estimating these parameters using type II maximum likelihood, that is maximising the marginal likelihood, in the context of variational Bayesian learning.

Let us consider the problem of hyperparameter adaptation or type II maximum likelihood estimation of the parameters of the gamma distribution. Given dataset $\mathcal{D}$, we have a model $\mathcal{H}$ with parameters $\boldsymbol{\theta}$. We are interested in a subset $\boldsymbol{\lambda} = (\lambda_i)_{i=1}^N \subset \boldsymbol{\theta}$ of the parameters. These parameters have a common gamma prior with

$$\begin{aligned} p(\lambda_i|\alpha,\beta) &= \text{Gamma}(\lambda_i;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda_i^{\alpha-1}e^{-\beta\lambda_i} \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\exp\big((\alpha-1)\log\lambda_i - \beta\lambda_i\big), \quad i = 1,\ldots,N. \end{aligned} \tag{1}$$

The latter form presents the gamma distribution as an exponential family with natural parameters $\big((\alpha-1),-\beta\big)$ and sufficient statistics $\big(\log\lambda_i,\lambda_i\big)$. Parameters such as λ_i are often used to model the precision (inverse variance) of Gaussian variables, because gamma distribution is the conjugate prior for the precision of a Gaussian. An example of such a model in the context of variational Bayes can be found in [1].

In general learning systems it may be difficult to know the correct scale of values for parameters in advance and thus it is often useful to apply an empirical Bayes approach to adapt the prior parameters. By type II maximum likelihood we mean finding such values for α and β that they maximise the marginal likelihood $p(\mathcal{D}|\mathcal{H}, \alpha, \beta)$. As the exact marginal likelihood is intractable, we use a lower bound obtained by variational Bayes (VB) approximation [1, 2]. We assume a factorial approximation

$$q(\boldsymbol{\lambda}|\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \prod_{i=1}^{N} q(\lambda_i|\hat{\alpha}_i, \hat{\beta}_i), \tag{2}$$

where $\hat{\alpha}_i$ and $\hat{\beta}_i$ are the variational parameters of the gamma distribution of the approximation for λ_i. The approximation is fitted by minimising the variational free energy[1]

$$\mathcal{C} = \left\langle \log \frac{q(\boldsymbol{\theta})}{p(\mathcal{D}, \boldsymbol{\theta}|\mathcal{H})} \right\rangle = \langle \log q(\boldsymbol{\theta}) - \log p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{H}) - \log p(\boldsymbol{\theta}|\mathcal{H}) \rangle \,, \tag{3}$$

where $\langle \cdot \rangle$ denotes expectation over the approximation $q(\boldsymbol{\theta})$. If the likelihood of λ_i is also in the conjugate exponential family, the approximations $q(\lambda_i|\hat{\alpha}_i, \hat{\beta}_i)$ can be easily updated by variational EM [3].

For the purpose of adapting α and β, the relevant terms of the variational free energy can now be derived using the expected values of the sufficient statistics of the gamma distribution,

$$\langle \lambda_i \rangle = \frac{\hat{\alpha}_i}{\hat{\beta}_i} \tag{4}$$

$$\langle \log \lambda_i \rangle = \Psi(\hat{\alpha}_i) - \log \hat{\beta}_i, \tag{5}$$

where $\Psi(x) = \frac{d}{dx} \log \Gamma(x)$ is the digamma function. The part involving α and β thus becomes

$$\mathcal{C}(\alpha, \beta) = \big\langle -\log p(\boldsymbol{\lambda}|\alpha, \beta) \big\rangle_{q(\boldsymbol{\lambda}|\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})} = \\ N \left(\log \Gamma(\alpha) - \alpha \log \beta \right) + \sum_{i=1}^{N} \left[(1 - \alpha) \big(\Psi(\hat{\alpha}_i) - \log \hat{\beta}_i \big) + \beta \frac{\hat{\alpha}_i}{\hat{\beta}_i} \right]. \tag{6}$$

We now seek to minimise Eq. (6) by finding a zero of the gradient. Differentiating Eq. (6) with respect to β yields

$$\frac{\partial \mathcal{C}}{\partial \beta} = -N \frac{\alpha}{\beta} + \sum_{i=1}^{N} \frac{\hat{\alpha}_i}{\hat{\beta}_i}. \tag{7}$$

[1] Free energy is equal to minus the variational lower bound.

Setting this to zero yields

$$\beta = \alpha \frac{N}{\sum_{i=1}^{N} \frac{\hat{\alpha}_i}{\hat{\beta}_i}} = \alpha T, \tag{8}$$

where $T = \frac{N}{\sum_{i=1}^{N} \frac{\hat{\alpha}_i}{\hat{\beta}_i}}$.

Substituting this back to Eq. (6) yields

$$\mathcal{C}(\alpha, T\alpha) = N\big(\log\Gamma(\alpha) - \alpha\log(T\alpha)\big) + \sum_{i=1}^{N}\left[(1-\alpha)\big(\Psi(\hat{\alpha}_i) - \log\hat{\beta}_i\big) + T\alpha\frac{\hat{\alpha}_i}{\hat{\beta}_i}\right]. \tag{9}$$

Differentiating Eq. (9) with respect to α yields

$$\begin{aligned}\frac{d\mathcal{C}}{d\alpha} &= N\big(\Psi(\alpha) - \log(T\alpha) - 1\big) + \sum_{i=1}^{N}\left[T\frac{\hat{\alpha}_i}{\hat{\beta}_i} - \big(\Psi(\hat{\alpha}_i) - \log\hat{\beta}_i\big)\right] \\ &= N\big(\Psi(\alpha) - \log\alpha - \log T - 1 + 1\big) - \sum_{i=1}^{N}\big(\Psi(\hat{\alpha}_i) - \log\hat{\beta}_i\big).\end{aligned} \tag{10}$$

Setting this to zero and dividing by N yields

$$\log\alpha - \Psi(\alpha) = \frac{1}{N}\sum_{i=1}^{N}\big(\log\hat{\beta}_i - \Psi(\hat{\alpha}_i)\big) - \log T. \tag{11}$$

Since $\log\alpha - \Psi(\alpha) \approx \frac{1}{2\alpha}$, the following form could be used as an approximate solution:

$$(\alpha \approx)\frac{1}{2}\left(\log\alpha - \Psi(\alpha)\right)^{-1} = \frac{1}{2}\left(\frac{1}{N}\sum_{i=1}^{N}\big(\log\hat{\beta}_i - \Psi(\hat{\alpha}_i)\big) - \log T\right)^{-1}. \tag{12}$$

However, by moving the terms to the same side and adding α we get a fixed point equation which yields an exact solution iteratively:

$$\begin{aligned}\alpha &= f(\alpha) \\ &= \alpha + \frac{1}{2}\left(\Psi(\alpha) - \log\alpha\right)^{-1} + \frac{1}{2}\left(\frac{1}{N}\sum_{i=1}^{N}\big(\log\hat{\beta}_i - \Psi(\hat{\alpha}_i)\big) - \log T\right)^{-1}.\end{aligned} \tag{13}$$

This converges, because $0 < f'(\alpha) < \frac{1}{2}$ for all $\alpha > 0$.

Thus we get a global fixed point iteration:

$$\begin{aligned}\alpha \leftarrow & \alpha + \frac{1}{2}\left(\Psi(\alpha) - \log(\alpha)\right)^{-1} \\ & + \frac{1}{2}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\log(\hat{\beta}_i) - \Psi(\hat{\alpha}_i)\right) - \log\left(\frac{N}{\sum_{i=1}^{N}\frac{\hat{\alpha}_i}{\hat{\beta}_i}}\right)\right)^{-1} \qquad (14)\end{aligned}$$

$$\beta \leftarrow \alpha\frac{N}{\sum_{i=1}^{N}\frac{\hat{\alpha}_i}{\hat{\beta}_i}} \qquad (15)$$

References

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

[2] T. S. Jaakkola, "Tutorial on variational approximation methods," in *Advanced Mean Field Methods: Theory and Practice* (M. Opper and D. Saad, eds.), pp. 129–159, Cambridge, MA, USA: The MIT Press, 2001.

[3] Z. Ghahramani and M. Beal, "Propagation algorithms for variational Bayesian learning," in *Advances in Neural Information Processing Systems 13* (T. Leen, T. Dietterich, and V. Tresp, eds.), pp. 507–513, Cambridge, MA, USA: The MIT Press, 2001.