Bayesian Generalized Linear Models in a Terabyte World

Onno Zoeter
Applied Games Group
Microsoft Research Cambridge
7 JJ Thomson Avenue, Cambridge, CB3 0FB, United Kingdom
onnoz@microsoft.com

Abstract

This paper introduces extremely fast approximate inference schemes for Bayesian treatments of dynamic generalized linear models. The approximations are tailored variants of quadrature EP. The first forward pass of this fixed point iteration algorithm can be interpreted as a one-step unscented Kalman filter. For on-line applications this filter can handle tens of thousands of updates a second on a current day desktop machine.

1. Introduction

Generalized linear models (GLMs) form a flexible class of regression models and are often used in classical statistics [5]. In Bayesian treatments of the GLM, instead of point estimates, full beliefs over regression weights are kept. This ensures that weights are adjusted at the correct rates, and perhaps more importantly, it prevents overfitting. For most GLMs the non-linear link functions lead to posteriors that have no compact analytical representation.

In many applications Markov Chain Monte Carlo methods are used to approximate the exact posteriors (see e.g. [2] for a detailed discussion of Bayesian GLMs and sampling approximations).

The typical use of these sampling methods is in an off-line batch mode with a number of datapoints in the tens or hundreds. In this paper we will be interested in extremely large scale applications with millions of observations and millions of weights that (perhaps after an initial off-line batch training period) should be practical for real-time use. In addition we do not expect the domain to be stationary and assume an auto-regressive (AR) process on the latent weights (Section 1.1 describes the model in detail).

In Section 2 we introduce an assumed density filter for the dynamic GLM. This approximation keeps Gaussian beliefs over individual weights. Gaussian quadrature integration rules lead to a general approximation scheme, applicable to all choices within the GLM framework. Section 3 introduces a tailored variant of quadrature EP [8]. Quadrature EP provides a smoother that is symmetric to (i.e. is based on the same principles and does not make more approximations than) the assumed density filter. For GLMs the standard quadrature EP can be adapted to use special kernels in the quadrature rules, such that less non-linearities are approximated.

1.1. Dynamic generalized linear models

In a GLM the observation for a one-dimensional output y given input vector x follows

$$p(y|x) = f(y; g^{-1}(w^\top x)) \ ,$$

where w is a vector containing linear regression weights, $f(y;\theta)$ is a one parameter distribution, and $g(\cdot)$ the so-called link function. The inverse of the link function maps the weighted sum of inputs to the parameter space of f.

Strictly speaking there is no need to give f and g separate names, we could just as well define $\psi(y; w^\top x) = f(y; g(w^\top x))$. In fact we will do so for notational convenience later in the paper. From an exposition/model construction point of view it might be beneficial to think of the noise model and the link function separately.

Different choices for $f(y;\theta)$ and $g(\cdot)$ yield different GLMs. Well known cases are Poisson, logit, and probit models. The last two models both have $f(y;\theta) = Bi(y;\theta)$ a binomial. The probit has $g(\theta) = \Phi(\theta)$, the normal cumulative distribution. The logit has $g(\theta) = \ln\left(\frac{\theta}{1-\theta}\right)$, the log-odds function. Poisson models use $f(y;\theta) = Po(y;\theta)$ and $g(\theta) = ln(\theta)$.

In non-stationary domains the weights can be assumed to follow a simple first order drift $w_{t+1} = w_t + \epsilon_t$, where subindeces denote discrete time steps, and ϵ_t are iid noise disturbances drawn from a zero-mean normal distribution with known variance. This is a special case of the dynamics in Kalman filter models. Since this part of the model can be treated exactly and is very well known its discussion is largely suppressed in this paper. Instead we will focus on the intractable factors in the model, the non-linear measurement updates.

2. Assumed density filtering

Our initial beliefs over weights is given by a factorized Gaussian:

$$p(w_1) = \prod_{i=1}^{N} N(w_{1,i}; \mu_{1,i}, \sigma^2_{1,i}) \ , \tag{1}$$

where $N(x; \mu, \sigma^2)$ denotes the Gaussian density function with mean μ and variance σ^2. Subindeces denote discrete time points, possibly augmented with an index in the vector of weights. For example w_1 is the vector of weights at time step 1, $w_{1,i}$ is the scalar weight for input i at time step 1.

As mentioned before, since the dynamics in latent space is linear Gaussian, standard Kalman filter updates can be used in the prediction steps in the filter. In this section we will describe a one-step unscented Kalman filter (one-step UKF) update [9] to approximate the measurement update. The one-step UKF is a first forward pass in the quadrature EP fixed point iteration algorithm that is introduced in Section 3. As a stand-alone filter it has been proposed independently several times in recent literature. We are currently aware of [4] and [1].

The one-step UKF derives its name from the unscented Kalman filter (UKF) [3]. The UKF uses quadrature to first linearize the observation model and then makes an update using the standard Kalman filter update (i.e. the measurement update consists of two steps), the one-step update approximates the posterior directly using quadrature, there by giving significant improvements if the observation model is far from linear. In particular, as is shown in [9], if x and y are uncorrelated in the observation model (but still dependent), the UKF provably yields no measurement update at all, whereas the UKF usually tracks the state adequately.

In the one-step UKF we will greedily, at each measurement update step, approximate the posterior by a factorized Gaussian of the same form as the prior (1). The approximation is such that it minimizes the Kullback-Leibler (KL) divergence between the unapproximated posterior and its factorized Gaussian approximation.

If we define the weight prior before the update as $q(w_t) = \prod_{i=1}^{N} N(w_{t,i}; \mu_{t,i}, \sigma^2_{t,i})$, and the unapproximated result of the measurement update as $p(w_t|y_t) \propto \psi_t(w_t; x_t, y_t) q(w_t)$, we are looking for $q^{\text{new}}(w_t) = \prod_{i=1}^{N} N(w_{t,i}; \mu^{\text{new}}_{t,i}, \sigma^{2\ \text{new}}_{t,i})$ in

$$\begin{aligned} q^{\text{new}}(w_t) &= \operatorname*{argmin}_{\tilde{p}(w_t) \in \mathcal{Q}} KL(p(w_t|y_t) || \tilde{p}(w_t)) && (2) \\ &= \operatorname*{argmin}_{\tilde{p}(w_t) \in \mathcal{Q}} \int p(w_t|y_t) \log \frac{p(w_t|y_t)}{\tilde{p}(w_t)} dw_t && (3) \end{aligned}$$

In the above $\mathcal{Q}$ is the family of factorized Gaussians over w_t. Since the KL minimization in (3) is over an exponential family, the optimal $\tilde{p}(w_t)$ is the factorized Gaussian that matches the natural moments (the means and variances) of $p(w_t|y_t)$. In other words, to find $q^{\text{new}}(w_t)$ we need to compute the individual means and variances of $p(w_t|y_t)$. In the one-step UKF we numerically compute these integrals (expectations) using Gaussian quadrature.

For simplicity we introduce a sum-node s_t:

$$\begin{aligned} s_t &= w_t^\top x_t \ , \quad \text{which implies} && (4) \\ s_t &\sim N(s_t; \mu_{s_t}, \sigma^2_{s_t}) \equiv q(s_t) \ , \quad \text{with} && (5) \\ \mu_{s_t} &= \sum_{i=1}^{N} x_{t,i} \mu_{t,i} && (6) \\ \sigma^2_{s_t} &= \sum_{i=1}^{N} x^2_{t,i} \sigma^2_{t,i} \ . && (7) \end{aligned}$$

Since the interaction between the individual $w_{t,i}$'s and s_t is linear Gaussian, this part can be treated exactly, and we can concentrate on finding the mean and variance of s_t after the observation.

The new approximated marginal for the sum-node is given by

$$\begin{aligned} q^{\text{new}}(s_t) &= N(s_t; \mu^{\text{new}}_{s_t}, \sigma^{2\ \text{new}}_{s_t}) \ , \quad \text{with} && (8) \\ \mu^{\text{new}}_{s_t} &= \int s_t \frac{1}{z_t} f(y_t; s_t) q(s_t) ds_t && (9) \\ z_t &= \int f(y_t; s_t) q(s_t) ds_t && (10) \\ \sigma^{2\ \text{new}}_{s_t} &= \int s_t^2 \frac{1}{z_t} f(y_t; s_t) q(s_t) ds_t - \left(\mu^{\text{new}}_{s_t}\right)^2 && (11) \end{aligned}$$

In the one-step UKF, the integrals (9)-(11) are numerically approximated using Gaussian quadrature (this method is due to Gauss, see e.g. [7] for an introduction).

To be precise if $\mathcal{X}'_i$ and $\mathcal{W}'_i$ are tabulated Gauss-Hermite points and weights [7], then $\mathcal{W}_i = \frac{\mathcal{W}'_i}{\sqrt{\pi}}$ and $\mathcal{X}_i = \mathcal{X}'_i \sqrt{2} \sigma_{s_t} + \mu_{s_t}$ are points and weights such that

$$\int_{-\infty}^{\infty} h(s_t) N\left(s_t; \mu_{s_t}, \sigma^2_{s_t}\right) = \sum_i h\left(\mathcal{X}_i\right) \mathcal{W}_i$$

for polynomials up to degree $2k-1$, if k quadrature points are used.

We use these points to find the following numerical approximations

$$\begin{aligned} \tilde{z}_t &= \sum_i f\left(y_t; \mathcal{X}_i\right) \mathcal{W}_i && (12) \\ \tilde{\mu}^{\text{new}}_{s_t} &= \frac{1}{\tilde{z}_t} \sum_i \mathcal{X}_i f\left(y_t; \mathcal{X}_i\right) \mathcal{W}_i && (13) \\ \tilde{\sigma}^{2\ \text{new}}_{s_t} &= \frac{1}{\tilde{z}_t} \sum_i \mathcal{X}_i^2 f\left(y_t; \mathcal{X}_i\right) \mathcal{W}_i - \left(\tilde{\mu}^{\text{new}}_{s_t}\right)^2, && (14) \end{aligned}$$

and use these as the parameters in $q^{\text{new}}(s_t)$. Section 4 shows in experiments the quality of these approximations for a logit model.

The updated Gaussian marginals over weights follow from the change in the sum node s_t. If we define the change in mean δ_μ and the change in variance δ_{σ^2} such that

$$\begin{aligned} \mu^{\text{new}}_{s_t} &= \mu_{s_t} + \delta_\mu \\ \sigma^{2\ \text{new}}_{s_t} &= \sigma^2_{s_t} + \delta_{\sigma^2} \ , \end{aligned}$$

then, using standard manipulations of Gaussians, we find that the new Gaussian marginals over $q^{\text{new}}(w_{t,i}) = N(w_{t,i}; \mu_{i,t}^{\text{new}}, \sigma_{i,t}^{2\,\text{new}})$ have means and variances given by

$$\mu_{t,i}^{\text{new}} = \mu_{t,i} + a_i \delta_\mu \quad (15)$$

$$\sigma^{\text{new}} 2_{t,i} = \sigma_{t,i}^2 + a_i^2 \delta_{\sigma^2}\,, \quad \text{where} \quad (16)$$

$$a_i = \frac{x_i \sigma_i^2}{\sum_{j=1}^N x_j^2 \sigma_j^2}\,. \quad (17)$$

From the above equations we see that the change in the marginal over the sum node s after seeing the observation is divided and distributed over the individual weights. The distribution is such that weights corresponding to larger inputs, and weights corresponding to larger prior uncertainty are adjusted more.

3 Quadrature EP for the GLM

Quadrature EP forms a general method of using Gaussian quadrature approximations for low dimensional integrals within the expectation propagation framework [6]. In this section we will discuss a tailored adaptation of quadrature EP that exploits properties of the GLM.

For a general discussion of expectation propagation and quadrature EP we refer the reader to [6] and [8]. Here we will only introduce the relevant notation to describe the extension.

The exact posterior we are looking for can be written as a product of factors

$$p(w_{1:T}|x_{1:T}, y_{1:T}) \propto \prod_{t=1}^T \psi_t(w_{t-1,t})\psi_t(s_t, w_t; x_t)\psi_t(w_t; y_t)\,.$$

The factors are simply the conditional distributions in the model

$$\begin{aligned} \psi_t(w_{t-1,t}) &\equiv p(w_{t-1}|w_t) \\ \psi_t(s_t, w_t; x_t) &\equiv p(s_t|w_t, x_t) \\ \psi_t(y_t, s_t) &\equiv p(y_t|s_t)\,. \end{aligned}$$

In EP the joint posterior is approximated by a product of approximate factors

$$p(w_{1:T}|x_{1:T}, y_{1:T}) \approx \prod_{t=1}^T \tilde{\psi}_t(w_{t-1,t})\tilde{\psi}_t(s_t, w_t; x_t)\tilde{\psi}_t(w_t; y_t)\,.$$

This product is chosen to be in a tractable form. For the GLM a sensible choice is a fully factorized Gaussian approximation

$$q(w_{1:T}, s_{1:T}) = \prod_t q_t(s_t) \prod_i q_{t,i}(w_{t,i})\,. \quad (18)$$

It must be emphasized that the approximation is far less coarse than the form in (18) may at first seem to imply. The algorithm will iteratively refine the approximation using the exact factors, essentially trying to find the marginals (18) that best match the exact posterior. This procedure is a strict generalization of the approximate filter. In the filter when making a greedy local approximation, there is no information coming from the future (multiplication by a uniform distribution). As presented in Section 3.1 the smoother and iteration scheme perform the local approximations in the light of more information.

The choice in (18) implies that approximate factors can, without loss of generality be written as a product of contributions to each of the variables in their domain. We will refer to these contributions as 'messages' to stress the similarity with belief propagation and loopy belief propagation (both are special cases). In the remainder we will be particularly interested in $\tilde{\psi}_t(s_t; y_t)$. Since the only variable is s_t (y_t is observed) we immediately get the shorthand $\tilde{\psi}_t(s_t; y_t) \equiv m_{y_t \to s_t}(s_t)$, the message from y_t to s_t, a Gaussian potential.

3.1 Refinement of $\tilde{\psi}(s_t; y_t)$ using quadrature EP

The general applicability of quadrature EP comes from the fact that old variable beliefs $q_i(x_i)$ can be used as kernels in quadrature integration rules during updates. The variable beliefs remain by construction in a single exponential family form. Most of the exponential families can be identified with well studied kernel forms such that the procedures for finding suitable points and weights can be found in the literature.

For the GLM we can use a tailored kernel in each update step: the belief divided by the message. In general models this need not be normalizable. However in the GLM all beliefs divided by messages still approximate a proper belief

$$\frac{q(s_t)}{m_{y_t \to s_t}(s_t)} \approx p\left(s_t | y_{\{1,\ldots,T\}\setminus t}\right)\,. \quad (19)$$

This special kernel 'absorbs' the (non-linear) inverse message part and hence we can expect the approximation to improve at no extra cost. It must be noted that, just as in the regular EP iteration, in rare cases skipping of updates [6] may be necessary.

With a kernel as in (19) an update of a the non-linear observation factors in the modified quadrature EP is as follows.

1. Determine the normalizing constant
$$Z_1 = \int_{-\infty}^{\infty} \frac{q(s_t)}{m_{y_t \to s_t}(s_t)} dx\,.$$
2. Determine points $\mathcal{X}_i$ and weights $\mathcal{W}_i$ for $r(s_t) \equiv \frac{1}{Z_1}\frac{q(s_t)}{m_{y_t \to s_t}(s_t)}$.
3. Approximate the normalization constant by
$$\tilde{Z} = \sum_i \psi_t(\mathcal{X}_i; x_t, y_t)\,\mathcal{W}_i Z_1\,.$$

4. Find mean and variance of new Gaussian s_t belief $q^{\text{new}}(s_t)$ using

$$\tilde{m} = \frac{1}{\tilde{Z}} \sum_i \mathcal{X}_i \psi_n(\mathcal{X}_i) \mathcal{W}_i Z_1 \text{ and}$$
$$\tilde{v} = \frac{1}{\tilde{Z}} \sum_i \mathcal{X}_i^2 \psi_n(\mathcal{X}_i) \mathcal{W}_i Z_1 - \tilde{m}^2 .$$

5. Infer the new message $m^{\text{new}}_{y_t \to s_t}(s_t)$ by division using

$$m^{\text{new}}_{y_t \to s_t}(s_t) = \frac{q^{\text{new}}(s_t)}{\frac{q(s_t)}{m_{y_t \to s_t}(s_t)}} = \frac{q^{\text{new}}(s_t)}{q(s_t)} m_{y_t \to s_t}(s_t) .$$

The updates of the linear-Gaussian interactions are simpler and can be done in closed form [9]. We initialize all messages with 1 and iterate in forward-backward passes. It is easy to see that with initial messages at 1 the first forward pass coincides with the one-step UKF of Section 2.

3.2 Checks for numerical problems

The Achilles heel of quadrature EP (and also for its special case the one-step UKF) is identical to that of importance sampling and particle filtering. The quadrature points and weights are determined based on a proposal distribution (the prior in the vanilla version of quadrature EP described above). If the proposal distribution has poor support in the region of the exact posterior the standard method breaks down and additional steps are required.

For now we can test a wide range of possible prior settings. Due to the symmetry of a positive and a negative observation we effectively have one possible observation factor to test. We can take a specific number quadrature points and check what ranges of priors lead to reasonable approximations. Section 4 demonstrates this for the logit model.

Any suggested trick from the importance sampling and particle filter literature can be tried if extreme prior/link function combinations are used.

4. Experiments

4.1 One-step UKF approximations

The one-step UKF and quadrature EP algorithms (i) approximate the posterior by a Gaussian, and (ii) find the mean and variance of this Gaussian using quadrature. In this section we show the effects of both approximations for the logit model. The logit model has as inverse link function the well known sigmoid $g^{-1}(s) = \frac{1}{1+\exp(-s)}$. Figure 1 shows an example of an update of the prior $q(s; 0, 25)$. The "exact posterior" shown in the plot is based on a fine grid with 1 million bins (grids with 10.000 or more bins give indistinguishable plots). Despite the fact that the exact posterior is asymmetric, the Gaussian approximation using 50 points is reasonably accurate. If we look at many possible prior distributions we see that the example of Figure 1, which corresponds with the point $\mu = 0$ and $\sigma = 5$ in Figure 2 is one of the worst approximations if we look at the KL between the non-parametric posterior based on the grid with 1 million bins and the Gaussian posterior. Figure 3 shows a similar plot, but in terms of the KL in the predictive distribution for the next observation. It shows that the approximations in a space that one might argue is more relevant, are even closer. Figure 4 shows a similar comparison, but now between the Gaussian found using the grid and the Gaussian found using the one-step UKF. These plots shows that, for the logit model at least, on the range of priors that we can reasonably expect to encounter, the one-step UKF approximation is reasonable, and that the approximation incurred by quadrature is small compared to the Gaussian approximation. Using only 10 points leads to a maximal KL of 0.08 in the equivalent of Figure 2 and a maximal KL of 0.018 in the equivalent of Figure 3 (not shown).

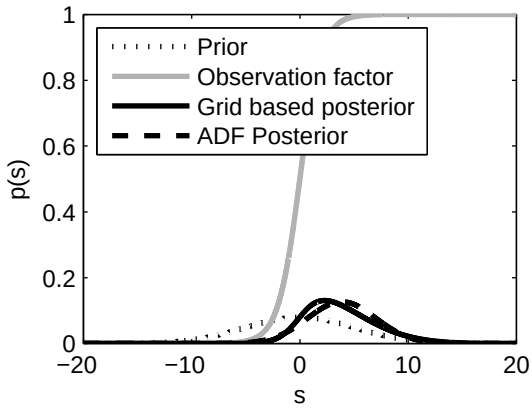

Figure 1. Grid based and one-step UKF posteriors for a logit observation.

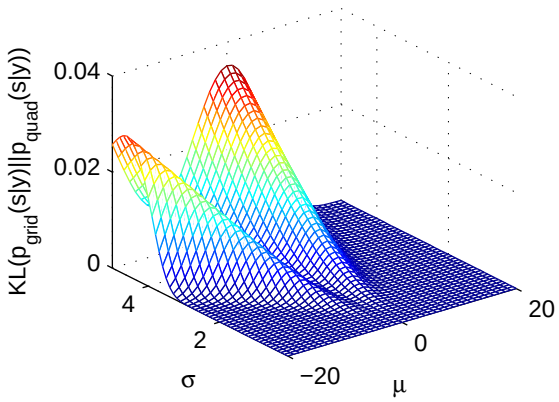

Figure 2. Difference between "exact" and one-step UKF posteriors.

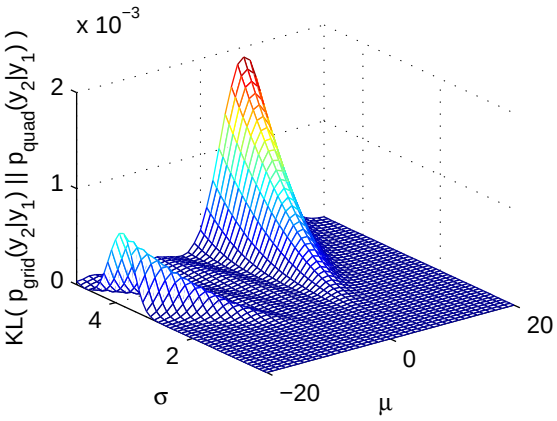

Figure 3. Difference between "exact" and one-step UKF predictions.

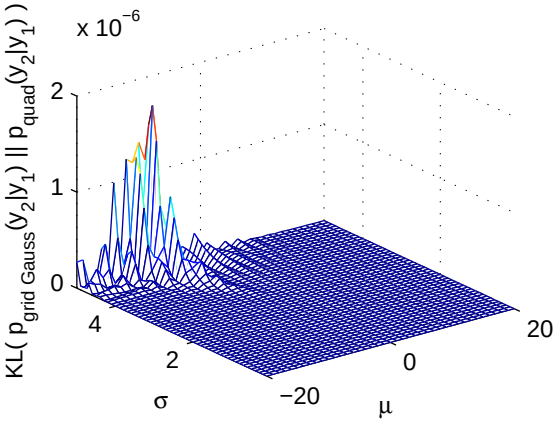

Figure 4. Difference between "exact Gauss" and one-step UKF posteriors.

4.2 Quadrature EP approximations

To study the effects of multiple measurement updates we look at a static probit model. This model has as link function $\Phi(s)$, the cumulative normal. With a prior $p(s) = N(0, 1)$ we have, after the inverse link function transformation $\theta = \Phi^{-1}(s)$, a uniform distribution on θ, the parameter in a Binomial model. Since the uniform distribution is a special case of the Beta distribution, and the Beta and Binomial are conjugate families, we know that the exact distribution after several updates is still Beta. This gives us a rare opportunity to test the approximation against the exact ground truth (it of course also tells us that for this very particular example it would not be wise to use the approximation).

Figure 5 shows the predicted probability of success in the 11th observation after seeing 1 success in 10 Bernoulli observations. The predictions are computed using (i) an exact analysis using a Beta distribution, (ii) the quadrature EP

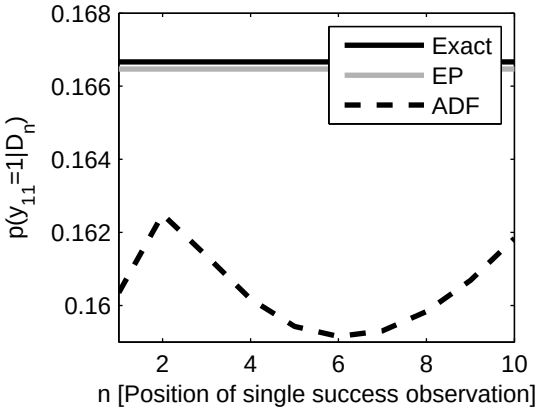

Figure 5. Quality of approximations and influence of dataset permutations

of Section 3, and (iii) the one-step UKF filter of Section 2. All methods assume a uniform prior on θ.

In the experiments we found that taking the prior as quadrature kernel or the prior divided by the message as suggested in Section 3.1 gave essentially the same results (not shown). So for probit and logit models where we studied the differences both choices appear to work fine. For other link functions a practicioner needs to assess for a particular choice of quadrature points the range of priors that lead to stable updates. At the same time the alternative kernels can be compared.

To show the effect of smoothing and iterating, the experiment is replicated 10 times, with the single success observation at 10 different positions. The x-axis denotes this position. Since quadrature EP iterates it is not influenced by the ordering of observations. Two and more iterations of EP give indistinguishable plots (only the results of 100 iterations is shown here).

5. Discussion

We have introduced a simple yet effective way of approximating GLMs using quadrature EP and one-step unscented Kalman filters. In the experimental results presented in Section 4 we find that the approximations work remarkably well. Due to their computational simplicity these methods open the way for the on-line use of GLMs with a number of inputs and a number observations in the millions and beyond.

References

[1] E. Bølviken and G. Storvik, Deterministic and stochastic particle filters in state space models, In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo methods in practice*, Springer-Verlag, 2001.

[2] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis*, Chapman and Hall, 1995.

[3] S. Julier and J. Uhlmann, "A new extension of the Kalman filter to nonlinear systems", In *Proceedings of AeroSense: The 11th International Symposium on Aerospace/Defense Sensing, Simulations and Controls*, 1997.

[4] H. Kushner and A. Budhiraja, "A nonlinear filtering algorithm based on an approximation of the conditional distribution", *IEEE Transactions on Automatic Control*, 45(3), 2000.

[5] P. McCullagh and J. Nelder, *Generalized Linear Models*, Chapman and Hall, 1989.

[6] T. Minka, "Expectation propagation for approximate Bayesian inference", In *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence*. Morgan Kauffman Publishers, 2001.

[7] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C: The Art of Scientific Programming*, Cambridge University Press, 2nd edition, 1992.

[8] O. Zoeter and T. Heskes, "Gaussian quadrature based expectation propagation", In Z. Ghahramani and R. Cowell, editors, *Proceedings AISTATS*, 2005.

[9] O. Zoeter, A. Ypma, and T. Heskes, "Improved unscented Kalman smoothing for stock volatility estimation", In A. Barros, J. Principe, J. Larsen, T. Adali, and S. Douglas, editors, *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing*, 2004.