Heterogeneous Bayes Filters with Sparse Bayesian Models: Application to state estimation in robotics

Alexandre Ravet[1,2] and Simon Lacroix[1,3]

[1] CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, France
[2] Univ de Toulouse, INSA, LAAS, F-31400 Toulouse, France
[3] Univ de Toulouse, LAAS, F-31400 Toulouse, France

Abstract. This study introduces a new augmented Bayes filter model for time-varying, context-dependent emission noise. The envisaged application, robust state estimation for a robot, motivates the use of the Relevance Vector Machine to model the emission noise, because it provides sparsity and fast inference capabilities. Besides the introduction of this new model, this work also aims at comparing the final filter performance when discriminative training is used instead of the prevalent generative training. The theoretical foundations for training and running inference over the model are proposed.

1 Introduction

Bayes filters (BF) have been widely applied to many areas. They are notably a workhorse of robotics, where recursive filtering [6], a fast and simple inference procedure, has provided the most common and reliable method for real-time state estimation over the past decades. BF used for state estimation usually rely on some optimistic assumptions for two main reasons: the real physical system is actually too complex to be perfectly described through a tractable model, and physical exactitude is neglected for the sake of computational efficiency. As a consequence, filter models designed for state estimation usually rely on a minimum system state representing the robot variables required for achieving a specific task. Any unmodeled aspect of the system, among which the effects of the current environment over the filter performance, are then encompassed within additional noise terms.

In practice, when used for robot state estimation, Bayes filters require a substantial tuning phase to provide acceptable performances. This is because capturing all unspecified aspects of the system through the sole introduction of noise often results in a trade-off between output optimality (accuracy of the state estimate) and robustness (to the different unmodeled aspects). An illustration of this problem, and the core motivation of this work, is the case of an autonomous robot navigating through different environments, in which one has to deal with a whole range of alterations in sensor readings, going from average (optimal) observation noise to complete failure (unreliable data). This is especially true

when sensor performances are strongly affected by the different environment characteristics, such as luminosity, texture and materials of surroundings objects, ground and obstacles...

The classical state-space model is defined by:

$$\begin{aligned} x_t &= f(x_{t-1}) + \gamma \\ y_t &= g(x_t) + \nu \end{aligned} \tag{1}$$

where x_t is the latent state at time t with the associated observation y_t, f and g the transition and emission functions respectively, $\gamma \sim \mathcal{N}(0, \Sigma_\gamma)$ the system noise and $\nu \sim \mathcal{N}(0, \Sigma_\nu)$ the observation noise. f and g can either be linear functions (linear dynamical system) or nonlinear (nonlinear dynamical system). Most often, no fixed noise values Σ_ν that would yield an optimal output for a large variety of operating conditions (or environments) can be determined. Other unmodeled effects producing the strongest measurement alterations are often compensated with a rejection scheme, usually relying on a model self-consistency check. In other words, designing a Bayes filter emission model consists in finding the best distribution modeling the emission process for the nominal cases, and reject all the data that does not fit this distribution [13]. One strong consequence of this approach is that the system might converge to erroneous but model-consistent state values [14, 12]. When such divergences are observed, additional parameter tuning is required to improve global robustness, then detrimentally affecting the state estimation performance.

This classical state-space model, whose graphical representation is shown in Fig.1, is known as (time) homogeneous. It can be enhanced by making the model parameters vary in time: the trade-off usually required when tuning the parameters is then no longer needed, and the resulting model can handle the whole spectrum of alterations over measurements. As the research context motivating this study concerns robust and adaptive perception for autonomous robots, this work focuses on the emission distribution – even though the proposed approach can be straightforwardly applied to the prediction distribution of model (1).

This paper aims at developing a model able to compensate for the assumptions made by describing a system through the simplified emission distribution with simple Gaussian noise Σ_ν – while maintaining high computational efficiency. It results in an enhanced Kalman filter capable of dealing with both moderate alterations and outliers, without requiring the implementation of rejection rules. This is done by training an additional model for the emission noise, which relies on contextual information input. One particularly appealing consequence of this approach lies in the introduction of a second order knowledge over measurements reliability, where basic rejection schemes rely on some knowledge about the data properties. The proposed approach is consequently less likely to diverge.

Discriminative training being known to help in compensating some of the mis-modeled aspects of a system [1], we also aim at analyzing the impact of a discriminative learning method versus a generative one over the performance of the resulting filter.

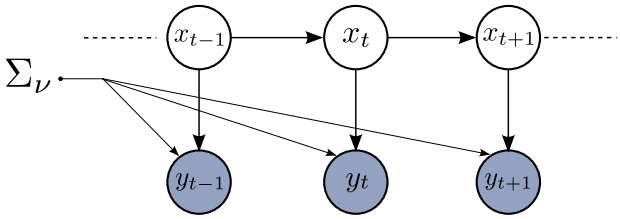

Fig. 1. Graphical model of a Bayes filter with homogeneous emission distributions.

The next section depicts the model basic principle. Sections 3 and 4 respectively describe generative and discriminative training of the model. Inference methods are then provided in section 5, and a discussion concludes the paper.

2 Heterogeneous BF with sparse Bayesian models

2.1 Background

Unless the state of simple models such as (1) encompasses all the exogenous phenomena likely to alter the system behavior, BF are by nature unable to model time varying emission and prediction processes. Recently, extensions have been proposed in order to compensate this unability. Nonparametric models such as Gaussian Processes (GP) have been integrated for modeling transition and emission distributions [3], and extended to fully state-dependent models in [9], through the introduction of a heteroscedastic observation noise. If nonparametric models improve the filter robustness compared to parametric functions, they are usually designed as state-dependent models, and as such are unable to handle contextual influence over the measurement process. For the heteroscedastic observation model proposed in [9], the presence of outliers in the training set is then critical: based on the sole state information, the system is unable to discern the contribution of the noise free model ($g(x_t)$ in (1)) from the noise model (which is then written $\nu(x_t)$) within measurements. An other disadvantage of these approaches is that one has to turn to more complex sparse GP techniques when using a large training set if the system is intended to be used in real time.

From a different perspective, optimizing the parameters of a BF has always been mostly considered as a tuning task, achieved with the intuitive goal of providing the most accurate state estimation. This allows to take into consideration some mis-modeled aspects of the system, even if they are never explicitly described, neither understood. Surprisingly, methods for learning the parameters of a BF appeared quite recently in the literature, and mostly rely on maximum likelihood. Since BF are generative models, this implies that the parameters are not determined with respect to the system ultimate performance, but so as to get the best model for the underlying prediction and emission processes. Conversely, discriminative training is similar to manual tuning, in the sense that

the parameters are optimized with respect to filter performance. But this latter training approach remains uncommon, although it proved to outperform manual tuning and maximum likelihood as well [1]. To our knowledge, combining an augmented model with discriminative training remains untreated, while both approaches serve a similar purpose, *i.e* compensating for unmodeled aspects of the real system.

2.2 Heterogeneous Bayes filters with sparse Bayesian model

To overcome BF unability to deal with context influence, an augmented model is introduced, whose particularity is to explicit the context repercussion over the measurement process. It relies on an additional observation variable c_t relating to the perception context. As suggested in previous work [12, 11], this additional observation can consist in the joint set – or subset – of sensor measurement values y_t possibly extended with any relevant contextual information i_t (any other sensor measurement, robot internal data, or any information that might influence the measurement emission process). It is assumed that this joint set of measurements defines a proper representation space for the contextual influence over measurement noise, *i.e* there exists a mapping from the context input space to the observation noise level.

To further avoid ambiguities in the contribution of two distinct models g and ν in the measurement process, we assume that the noise free component g of the emission model is known and homogeneous in time, since it can generally be obtained directly through physical considerations about the nominal measurement generation process. This results in an emission model of the form:

$$y_t = g(x_t) + \nu(c_t)$$

Reminding the goal of this model is to enhance a Kalman filter, $\nu(c_t)$ is then a zero mean Gaussian noise distribution with context-dependent variance:

$$\nu(c_t) \sim \mathcal{N}\big(0, r(c_t)\big)$$

To avoid making any assumption over the functional form for the variance model, and keep computational efficiency, $r(c_t)$ is modeled with the Relevance Vector Machine (RVM) framework [15], naturally providing sparsity thanks to the *automatic relevance determination* mechanism. For a given training set of T observations $\{c_i\}_{i=1}^{T}$, we define $r(c_t) = exp(z_t)$ to ensure variance positivity where z_t is given by

$$z_t = \sum_{i=1}^{T} w_i K(c_t, c_i) + w_0 + \epsilon \tag{2}$$

with w_0 a bias parameter, K the chosen kernel function, and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. To foster sparsification, a zero-mean Gaussian prior is placed over the weight vector $\mathbf{w} = (w_0, w_1, .., w_N)^T$:

$$p(\mathbf{w}|\alpha) = \prod_{i=0}^{N} \mathcal{N}(w_i|0, \alpha_i^{-1})$$

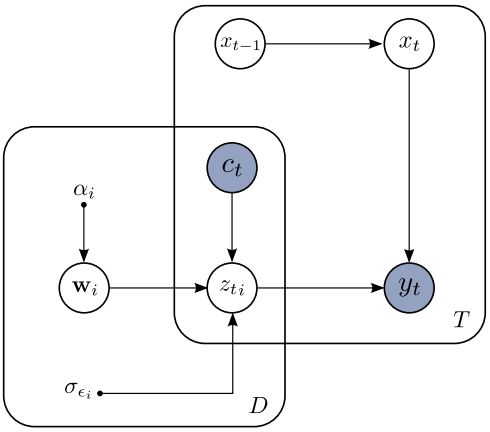

Fig. 2. Bayes Filter with heterogeneous emission noise.

where we define uniform hyperparameters priors over $\alpha = (\alpha_0, \alpha_1, ..., \alpha_N)^T$.

So far, the model has been depicted for a one-dimensional observation space. Real applications however require to consider the multi-dimensional case. For an observation variable $y_t \in \mathbb{R}^D$, D distinct RVM models are then used to model each component of the noise covariance matrix $\nu(c_t)$. For clarity, the next sections only consider the one-dimensional case, the extension to higher dimensions being straightforward. The resulting graphical model of this augmented model is depicted Fig.2.

3 Generative training

Bayes filter parameter optimization is usually done by minimizing the likelihood of the training set [4, 2], considering the latent state variable remains unobserved. In this work, we assume that the training set also contains ground truth data, *i.e* accurate values of the state variables $\mathbf{x} = \{x_1, ..., x_T\}$. The issue of training the emission model then turns to be analogous to the regression task, and more specifically to the heteroscedastic regression task with nonparametric models [5, 7, 10, 8, 9]. Note however that the training task is here a bit simpler since the observation function g is fixed and only the model $\nu(c_t)$ has to be determined. If sampling and variational approximation can be also used, the chosen approach relies on hard-assignment Expectation Maximization (EM) as suggested in [7]. By using hard-assignment EM we iteratively estimate the RVM parameters and predicted log noise level $\mathbf{z}$ at original inputs $C = \{c_i\}_{i=1}^T$. Thanks to this approximation, we are able to make direct use of classical RVM optimization and prediction equations, providing in this context the fastest solution for a real time application.

Following Kersting et al. approach [7], the hard E-step consists in empirical estimation of the noise variance. Based on real observations $\mathbf{y} = \{y_1, ..., y_T\}$ and

samples y_t^k provided by the current observation model (using the parameters α and σ_ϵ found after last EM iteration), the set of values y_t and $\{y_t^k\}_{k=1}^K$ are seen as independent noisy observations of $g(x_t)$. Empirical estimation of the noise variance at x_t is then provided by the mean

$$var_t = \frac{1}{2.K} \sum_{k=1}^{K} (y_t - y_t^k)^2$$

In the subsequent M-step the RVM model is trained with the new training set $D = \{c_t, log(var_t)\}_{t=1}^T$ with a classical optimization procedure [15].

In other words, the optimization process considers the noise variance as the hidden variable of the model, and iteratively optimize the parameters of the RVM model based on a hard assignment of the estimated noise. This method requires to use a substantial number of samples to empirically estimate the noise variance and, as any hard-assignment EM, is prone to oscillating, requiring to monitor the likelihood of the model over the training set after each algorithm iteration. It however brings an important advantage, since the optimized model, in association with the last noise variance estimation, can be readily used for prediction using classical RVM equations.

4 Discriminative training

The previous learning approach aims at minimizing a loss function corresponding to the emission likelihood. In other words the optimization step finds model parameters explaining at best the measurement generation process. As suggested in [1], it is however better to optimize the parameters with respect to the ultimate system performance, *i.e* the accuracy of filter estimates. Training the model then consists in finding α_{max} and $\sigma_{\epsilon max}$ such that:

$$\langle \alpha_{max}, \sigma_{\epsilon max} \rangle = \arg\max_{\alpha, \sigma_\epsilon} \sum_{t=1}^{T} log(p(x_t|y_{1:t}))$$

where $p(x_t|y_{1:t})$ is provided by Kalman equations. Considering f and g are linear functions corresponding to the matrices F and G respectively (classical Kalman filter case), we have:

$$p(x_t|y_{1:t}) = \mathcal{N}(x_t|\mu_t, \sigma_t)$$

with

$$\begin{aligned} \mu_t &= F\mu_{t-1} + K_t(y_t - GF\mu_{t-1}) \\ \sigma_t &= (I - K_t G)P_{t-1} \end{aligned}$$

$$\begin{aligned} P_{t-1} &= F\sigma_{t-1}F^t + \Sigma_\gamma \\ K_t &= P_{t-1}G^t(GP_{t-1}G^t + \Sigma_{\nu(c_t)}) \end{aligned}$$

Since this distribution requires the evaluation of two latent variables: the log noise level and the RVM weight parameter, a procedure similar to Type II maximum likelihood is employed. z_t and $\mathbf{w}$ are then marginalized out and we now seek for parameters α_{max} and $\sigma_{\epsilon max}$ maximizing:

$$\mathcal{L}_{discr} = \sum_{t=1}^{T} log\Big(\int \int p(x_t|y_{1:t}, z_t) p(z_t|c_t, \mathbf{w}, \sigma_\epsilon) p(\mathbf{w}|\alpha) \, \mathrm{d}\mathbf{w} \, \mathrm{d}z_t \Big)$$

Since $p(z_t|c_t, \mathbf{w}, \sigma_\epsilon)$ and $p(\mathbf{w}|\alpha)$ are both Gaussian, the integral with respect to $\mathbf{w}$ is readily evaluated to give:

$$\mathcal{L}_{discr} = \sum_{t=1}^{T} log\Big(\int p(x_t|y_{1:t}, z_t) \mathcal{N}(z_t|0, D_t) \mathrm{d}z_t \Big)$$

where $D_t = \sigma_\epsilon + K^T \alpha^{-1} K$, with K the vector of kernel functions such that $K_i = K(c_t, c_i)$ as defined in (2), and $A = \mathrm{diag}(\alpha)$. The last equation is analytically intractable and is then approximated with integration by substitution and Gauss-Hermite quadrature.

α_{max} and $\sigma_{\epsilon max}$ are subsequently found by conjugate gradient ascent over $\mathcal{L}_{discr}$. Note that classical optimization of the RVM model requires the computation of a *design matrix* containing all kernel elements evaluated at all original locations $\{c_i\}_{i=1}^{T}$. Here, the optimization is done separately for each kernel vector evaluated at c_i, through their influence over ultimate filter performance. This different form of training (by comparison to the one in [7]) then requires to foster sparsity by thresholding of the α_i values during the optimization process.

5 Inference

5.1 Generative training case

For the classical model of Fig.1, filtering consists in evaluating normalized marginal distributions $\hat{\alpha}(x_t) = p(x_t|y_1, .., y_t)$ with the recursion equation of the form:

$$\eta_t \hat{\alpha}(x_t) = p(y_t|x_t) \int \hat{\alpha}(x_{t-1}) p(x_t|x_{t-1}) \mathrm{d}x_{t-1} \tag{3}$$

where $\eta_t = p(y_t|y_1, ..., y_{t-1})$ the scaling factor and $p(y_t|x_t)$ and $p(x_t|x_{t-1})$ the emission and prediction distribution considered as Gaussian for kalman filtering. Since in the new model the noise level is considered as an additional latent variable, the emission distribution required for the evaluation of (3) is now given by:

$$p(y_t|x_t, c_t) = \int \mathcal{N}\big(y_t|g(x_t), exp(z_t)\big) p(z_t|c_t, C, \mathbf{z}, \alpha, \sigma_\epsilon) \mathrm{d}z_t \tag{4}$$

where $\mathbf{z}$ is the predicted log noise level at original inputs $\{c_i\}_{i=1}^{T}$, and $p(z_t|c_t, C, \mathbf{z}, \alpha, \sigma_\epsilon)$ the predictive distribution given by:

$$p(z_t|c_t, C, \mathbf{z}, \alpha, \sigma_\epsilon) = \int p(z_t|c_t, \mathbf{w}, \sigma_\epsilon) p(\mathbf{w}|C, \mathbf{z}, \alpha, \sigma_\epsilon) \mathrm{d}\mathbf{w} \tag{5}$$

This familiar predictive distribution [15] is also Gaussian. The evaluation of the integral (4) is hence analytically intractable, and requires approximation. The fastest approach is the most likely approximation, where we replace the integral by $\mathcal{N}\big(y_t|g(x_t), exp(z_t^*)\big)$ with $z_t^* = \arg\max_{z_t} p(z_t|c_t, C, \mathbf{z}, \alpha, \sigma_\epsilon)$. Note that this approximation allows $p(y_t|x_t, c_t)$ to be a Gaussian distribution, a necessary condition for using Kalman recursive equations.

5.2 Discriminative training case

Since discriminative training did not involve evaluation of the posterior distribution of the noise level $\mathbf{z}$, we can not make straightforward use of the RVM prediction equation. We then replace equation (5) by:

$$p(z_t|c_t, C, \alpha, \sigma_\epsilon) = \int\int p(z_t|c_t, \mathbf{w}, \sigma_\epsilon) p(\mathbf{w}|C, \mathbf{z}, \alpha, \sigma_\epsilon) p(\mathbf{z}|C, \alpha, \sigma_\epsilon) \mathrm{d}\mathbf{w}\mathrm{d}\mathbf{z} \quad (6)$$

We turn to MCMC sampling in order to evaluate at first the posterior $p(\mathbf{w}|C, \alpha, \sigma_\epsilon)$ reminding that

$$p(\mathbf{w}|C, \alpha, \sigma_\epsilon) = \int p(\mathbf{w}|C, \mathbf{z}, \alpha, \sigma_\epsilon) p(\mathbf{z}|C, \alpha, \sigma_\epsilon) \mathrm{d}\mathbf{z}$$

and as described in [15],

$$p(\mathbf{w}|C, \mathbf{z}, \alpha, \sigma_\epsilon) = \mathcal{N}(\mathbf{w}|m, \Sigma)$$
$$p(\mathbf{z}|C, \alpha, \sigma_\epsilon) = \mathcal{N}(\mathbf{z}|0, E)$$

where $m = \sigma_\epsilon \Sigma \Phi^T \mathbf{z}$, $\Sigma = (A + \sigma_\epsilon \Phi^T \Phi)^{-1}$ and $E = \sigma_\epsilon^{-1} I + \Phi A^{-1} \Phi^T$, with Φ the *design matrix*. For computational efficiency, evaluation of the predictive distribution over the noise level for a new input c_t is then done by replacing the integral over $\mathbf{w}$ in (6) by the evaluation of $p(z_t|c_t, \mathbf{w}, \sigma_\epsilon)$ at the point estimate value of $\mathbf{w}$ provided by the sampling procedure, which is done offline. The remaining of the inference is then similar to the one depicted in the generative training case.

6 Discussion

The aim of this work is to define a new Bayes filter model able to encompass a variety of contexts, and to analyse different training approaches and their expected consequences over the system performance. Its specificities rely in the introduction of an additional observation used for context identification, and in the use of a sparse RVM model for context-dependent observation noise prediction. The theoretical foundations have been presented and system performances are currently being investigated. Note that the idea of introducing an additional context variable has been tested in our previous work [12, 11] and proved to be relevant for the simple task of context-dependent sensor selection (equivalent to rejection). However experiments conducted so far concerned the simple task of

altitude estimation for an UAV, and the approach has still to be tested on more complex scenarios, involving numerous sensors and a broad range of contexts.

While augmenting Bayes filters with time-varying noise model plays a central role in trying to compensate the optimistic assumptions usually made by the classical model, the training method might also have major consequences over the system performance. As such, discriminative training seems promising in that it requires to *run* the filter during optimization while generative training focuses on the underlying emission and prediction processes. Discriminative training nevertheless brings some particular issues, since at first, it does not allow to use classical training (and sparsification) method for the RVM model, but also because the optimization process of $\mathcal{L}_{discr}$ is much more complex. Indeed, each term of the discriminative loss function is strongly related to the preceding one as a direct consequence of the recursive equations used for state estimation. Classical RVM models already require the optimization of a nonconvex function, and we still need to study the consequences of the additional complexity introduced along with this specific loss function. Besides this particular issue, the use of discriminative training also ends up with some additional approximations during inference. Experiments will provide a better insight on how expected benefits and inherent drawbacks of the discriminative method impact the system performance, by comparison to the much simpler generative approach.

References

1. Abbeel, P., Coates, A., Montemerlo, M., Ng, A.Y., Thrun, S.: Discriminative training of kalman filters. In: Proceedings of Robotics: Science and Systems. Cambridge, USA (June 2005)
2. Bishop, C.: Pattern recognition and machine learning, vol. 4. Springer New York (2006)
3. Deisenroth, M.P., Turner, R.D., Huber, M.F., Hanebeck, U.D., Rasmussen, C.E.: Robust filtering and smoothing with gaussian processes. CoRR abs/1203.4345 (2012)
4. Ghahramani, Z., Hinton, G.E.: Parameter estimation for linear dynamical systems. Tech. rep. (1996)
5. Goldberg, P.W., Williams, C.K.I., Bishop, C.M.: Regression with input-dependent noise: A gaussian process treatment. In: In Advances in Neural Information Processing Systems 10. pp. 493–499. MIT Press (1998)
6. H. E. Rauch, F. Tung, C.T.S.: Maximum likelihood estimates of linear dynamic systems, (1965)
7. Kersting, K., Plagemann, C., Pfaff, P., Burgard, W.: Most likely heteroscedastic gaussian process regression. In: Proceedings of the 24th International Conference on Machine Learning. pp. 393–400. ICML '07, ACM, New York, NY, USA (2007)
8. Khashabi, D., Ziyadi, M., Liang, F.: Heteroscedastic relevance vector machine. CoRR abs/1301.2015 (2013)
9. Ko, J., Fox, D.: Gp-bayesfilters: Bayesian filtering using gaussian process prediction and observation models. Autonomous Robots 27(1), 75–90 (2009)
10. Lzaro-gredilla, M., Titsias, M.K.: Variational heteroscedastic gaussian process regression. In: In 28th International Conference on Machine Learning (ICML-11. pp. 841–848. ACM (2011)

11. Ravet, A., Lacroix, S., Hattenberger, G.: Context-dependent bayesian filtering: an approach based on measurement selection and supervised learning. In: To be published in Intelligent Robots and Systems (IROS), 2014 IEEE/RSJ International Conference on (2014)
12. Ravet, A., Lacroix, S., Hattenberger, G., Vandeportaele, B.: Learning to combine multi-sensor information for context dependent state estimation. In: Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on. pp. 5221–5226 (2013)
13. Sivia, D., Skilling, J.: Data Analysis: A Bayesian Tutorial. Oxford science publications, OUP Oxford (2006), http://books.google.fr/books?id=zN-yliq6eZ4C
14. Thrun, S., Burgard, W., Fox, D.: Probabilistic Robotics (Intelligent Robotics and Autonomous Agents). The MIT Press (2005)
15. Tipping, M.E.: Sparse bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1, 211–244 (Sep 2001)