Prediction and Fault Detection of Environmental Signals with Uncharacterised Faults: Appendix

Michael A. Osborne
Engineering Science
University of Oxford
Oxford OX1 3PJ, UK
mosb@robots.ox.ac.uk

Roman Garnett
Robotics Institute
Carnegie Mellon University
Pittsburgh PA 15213, US
rgarnett@cs.cmu.edu

Kevin Swersky
Aquatic Informatics
Suite 1100, 570 Granville St
Vancouver, V6C 3P1, Canada
kswersky@cs.toronto.edu

Nando de Freitas
Computer Science
University of British Columbia
Vancouver, V6T 1Z4, Canada
nando@cs.ubc.ca

Abstract

Additional mathematical details to supplement OSBORNE, *et al.* (2012).

Fault Bucket

We propose an algorithm that is designed to deal with faults of many different, unspecified types. We use a sequential scheme, applicable for ordered data such as time series, partitioning the data available at any point into old and new halves. We then approximately marginalise the faultiness of old observations, storing and then updating our results for future use. This gives rise to an efficient and fast algorithm. In order to effect our scheme, we make four key approximations:

1. **Fault bucket:** Faulty observations are assumed to be generated from a Gaussian noise distribution with a very wide variance.
2. **Single-Gaussian marginal:** A mixture of Gaussians, weighted by the posterior probabilities of faultiness of old data, is approximated as a single moment-matched Gaussian.
3. **Old/new noise independence:** We assume that noise contributions are independent, and that the contributions for new data are independent of old observations.
4. **Affine precision:** The precision matrix over both old and new halves is assumed to be affine in the precision matrix over the old half.

Approximation 1 represents the state-of-the-art DERESZYNSKI and DIETTERICH (2011). However, using it alone will not give an algorithm that can scale to the real-time problems we consider. Our novel approximations 2-4 permit very fast, fault-tolerant inference. We will detail and justify these approximations further below.

Our single, catch-all, "fault bucket" is expressed by approximation 1. It is built upon the expectation that points that are more likely to have been generated by noise with wide variance than under the normal predictive model of the GP can reasonably be assumed to be corrupted in some way, assuming we have a good understanding of the latent process. It is hoped that a very broad class of faults can be captured in this way. To formalise this idea, we choose an observation noise distribution that models the noise as independent but not identically distributed with separate variances for the non-fault and fault cases:

$$\begin{aligned} p(y \mid f, x, \neg\,\mathrm{fault}, \sigma_n^2) &= \mathcal{N}(y; f, \sigma_n^2) \\ p(y \mid f, x, \mathrm{fault}, \sigma_f^2) &= \mathcal{N}(y; f, \sigma_f^2), \end{aligned} \tag{1}$$

where $\mathrm{fault} \in \{0, 1\}$ is a binary indicator of whether the observation $y(x)$ was faulty and $\sigma_f > \sigma_n$ is the standard deviation around the mean of faulty measurements. The values of both σ_n and σ_f form hyperparameters of our model and are hence included in θ.

Of course, *a priori*, we do not know whether an observation will be faulty. Unfortunately, managing our uncertainty about the faultiness of all available observations is a challenging task. With N observations, there are 2^N possible assignments of faultiness; it is infeasible to consider them all.

Our solution is founded upon approximation 2. For time series, the value to be predicted $f_\star$ typically lies in the future, and old observations are typically less pertinent for this task than new ones. We hence approximately marginalise the faultiness of old observations, representing the mixture of different Gaussian predictions (each given by a different combination of faultiness) as a single Gaussian. We prefer this approximate marginalisation over faultiness to heuristics that would designate all observations as either faulty or not—we acknowledge our uncertainty about faultiness.

More formally, imagine that we have partitioned our observations $\mathcal{D}_{a,b}$ into a set of old observations $\mathcal{D}_a = (\mathbf{x}_a, \mathbf{y}_a)$ and a set of new observations $\mathcal{D}_b = (\mathbf{x}_b, \mathbf{y}_b)$. Define σ_a to be the (unknown) vector of all noise variances at observations $\mathbf{y}_a$, and define σ_b similarly. Because we have to sum over all possible values for these vectors, we will index the possible values of σ_a by i (each given by a different combination of faultiness over $\mathcal{D}_a$) and the values of σ_b similarly by j. We now define the covariances $V_a^i = K_{a,a} + \operatorname{diag} \sigma_a^i$, $V_b^j = K_{b,b} + \operatorname{diag} \sigma_b^j$ and $V_{a,b}^{i,j} = K_{\{a,b\},\{a,b\}} + \operatorname{diag}\{\sigma_a^i, \sigma_b^j\}$, where $\operatorname{diag} \sigma$ is the diagonal matrix with diagonal σ.

To initialise our algorithm, imagine that a identifies a small set of data, such that we can readily compute the like-

lihood of our hyperparameters

$$p(\mathbf{y}_a) = \sum_i p(\mathbf{y}_a|\sigma_a^i)p(\sigma_a^i) = \sum_i \mathcal{N}(\mathbf{y}_a; 0, V_a^i)p(\sigma_a^i) \quad (2)$$

and hence the hyperparameter posterior, $p(\sigma_a|\mathbf{y}_a)$. This distribution specifies the probability of our observations $\mathcal{D}_a$ being faulty; for a single observation $\mathcal{D}_a$, $p\big(\text{fault}(\mathcal{D}_a)\,|\,\mathbf{y}_a\big) = p(\sigma_a = \sigma_f\,|\,\mathbf{y}_a)$. If we were to perform predictions for some $f_\star$ using $\mathcal{D}_a$ alone, we would need to evaluate

$$\begin{aligned} &p(f_\star\,|\,\mathbf{y}_a) \\ &= \sum_i p(\sigma_a^i\,|\,\mathbf{y}_a)p(f_\star\,|\,\mathbf{y}_a, \sigma_a^i) \\ &= \sum_i p(\sigma_a^i\,|\,\mathbf{y}_a)\mathcal{N}\big(f_\star; m(f_\star\,|\,\mathbf{y}_a, \sigma_a^i), C(f_\star\,|\,\mathbf{y}_a, \sigma_a^i)\big), \end{aligned}$$

the weighted sum of Gaussian predictions made using the different possible values for σ_a. We now use approximation 2. It is our hope that our predictions for $f_\star$ are not so sensitive to the noise in our observations that all the Gaussians in this sum become dramatically different. In any case, the quality of this approximation will improve over time—if $f_\star$ is far removed from our old data $\mathcal{D}_a$, then our predictions really will not be very sensitive to σ_a. So, we take

$$\begin{aligned} p(f_\star\,|\,\mathbf{y}_a) \simeq \mathcal{N}\big(&f_\star; K_{\star,a}\tilde{V}_a^{-1}\mathbf{y}_a, \\ &K_{\star,\star} - K_{\star,a}(\tilde{V}_a^{-1} - \tilde{W}_a^{-1})K_{a,\star} - (K_{\star,a}\tilde{V}_a^{-1}\mathbf{y}_a)^2\big), \end{aligned}$$

where[1]

$$\begin{aligned} \tilde{V}_a^{-1} &= \sum_i p(\sigma_a^i\,|\,\mathbf{y}_a)(V_a^i)^{-1}, \\ \tilde{W}_a^{-1} &= \sum_i p(\sigma_a^i\,|\,\mathbf{y}_a)(V_a^i)^{-1}\mathbf{y}_a\mathbf{y}_a^\mathsf{T}(V_a^i)^{-1}. \end{aligned} \quad (4)$$

With these calculations performed, imagine receiving further data $\mathcal{D}_b$. To progress, we make approximation 3; we assume that faults will not persist longer than $|\mathcal{D}_b|$. To be precise, we assume

$$p(\sigma_{a,b}^{i,j}, \mathbf{y}_{a,b}) \simeq p(\mathbf{y}_a)\,p(\sigma_a^i\,|\,\mathbf{y}_a)\,p(\sigma_b^j)\,p(\mathbf{y}_b\,|\,\sigma_{a,b}^{i,j}, \mathbf{y}_a) \quad (5)$$

Our predictions are now

$$\begin{aligned} p(f_\star\,|\,\mathbf{y}_{a,b}) \simeq &\sum_j p(\sigma_b^j\,|\,\mathbf{y}_{a,b}) \sum_i p(\sigma_a^i\,|\,\mathbf{y}_a) \\ &\mathcal{N}\big(f_\star; m(f_\star\,|\,\mathbf{y}_{a,b}, \sigma_{a,b}^{i,j}), C(f_\star\,|\,\mathbf{y}_{a,b}, \sigma_{a,b}^{i,j})\big). \end{aligned} \quad (6)$$

[1]Note that for $\tilde{W}_a$, explicitly computing (unstable) matrix inverses can be avoided by solving the appropriate linear equations using Cholesky factors. For $\tilde{V}_a$, we can rewrite $(A^{-1}+B^{-1})^{-1} = A(A+B)^{-1}B$. If $i \in \{0, 1\}$ (as it would be if a identified a single observation which could be either faulty or not),

$$\tilde{V}_a = V_a^0\big(p(\sigma_a^1\,|\,\mathbf{y}_a)V_a^0 + p(\sigma_a^0\,|\,\mathbf{y}_a)V_a^1\big)^{-1}V_a^1. \quad (3)$$

If i takes more than two values, we can simply iterate using the same technique. We can then use the Cholesky factor of $\tilde{V}_a$ to compute our required equations.

Before trying to manage these sums, we will determine $p(\sigma_b \mid \mathbf{y}_{a,b})$. As before, this distribution gives us the probability of the observations $\mathcal{D}_b$ being faulty. For example, if we have only a single observation $\mathcal{D}_b$, $p\big(\text{fault}(\mathcal{D}_b)\,|\,\mathbf{y}_{a,b}\big) = p(\sigma_b = \sigma_f\,|\,\mathbf{y}_{a,b})$. We define

$$\begin{aligned} \tilde{m}(\mathbf{y}_b\,|\,\mathbf{y}_a) &= K_{b,a}\tilde{V}_a^{-1}\mathbf{y}_a \\ \tilde{C}(\mathbf{y}_b\,|\,\mathbf{y}_a, \sigma_b) &= V_b - K_{b,a}(\tilde{V}_a^{-1} - \tilde{W}_a^{-1})K_{a,b} \\ &\qquad - \tilde{m}(\mathbf{y}_b\,|\,\mathbf{y}_{a,b})^2, \end{aligned}$$

where both $\tilde{V}_a$ (or its Cholesky factor) and $\tilde{W}_a^{-1}$ were computed previously. By using approximations 2 and 3,

$$\begin{aligned} p(\sigma_b\,|\,\mathbf{y}_{a,b}) &= \frac{\sum_i p(\mathbf{y}_b\,|\,\mathbf{y}_a, \sigma_{a,b}^i)p(\mathbf{y}_a, \sigma_{a,b}^i)}{p(\mathbf{y}_{a,b})} \\ &\simeq \frac{\mathcal{N}\big(\mathbf{y}_b; \tilde{m}(\mathbf{y}_b\,|\,\mathbf{y}_a), \tilde{C}(\mathbf{y}_b\,|\,\mathbf{y}_a, \sigma_b)\big)p(\sigma_b)}{p(\mathbf{y}_b\,|\,\mathbf{y}_a)}, \end{aligned} \quad (7)$$

where we have

$$\begin{aligned} p(\mathbf{y}_b\,|\,\mathbf{y}_a) &= \sum_i \sum_j p(\mathbf{y}_b\,|\,\mathbf{y}_a, \sigma_{a,b}^i)p(\sigma_{a,b}^{i,j}\,|\,\mathbf{y}_a) \\ &\simeq \sum_j \mathcal{N}\big(\mathbf{y}_b; \tilde{m}(\mathbf{y}_b\,|\,\mathbf{y}_a), \tilde{C}(\mathbf{y}_b\,|\,\mathbf{y}_a, \sigma_b^j)\big)p(\sigma_b^j). \end{aligned} \quad (8)$$

Note that

$$p(\mathbf{y}_{a,b}) = p(\mathbf{y}_b\,|\,\mathbf{y}_a)\,p(\mathbf{y}_a), \quad (9)$$

the product of (8) and (2), gives the likelihood of our hyperparameters, useful if we want to learn such hyperparameters from data using, for example, maximum marginal likelihood. Now, returning to (6), we will once again use approximation 2. We aim to reuse our previously evaluated sums over i to resolve future sums over i. As we gain more data, the faultiness of old data becomes less important. We arrive at

$$\begin{aligned} p(f_\star\,|\,\mathbf{y}_{a,b}) \simeq \mathcal{N}\big(&f_\star;\ K_{\star,\{a,b\}}\tilde{V}_{a,b}^{-1}\mathbf{y}_{a,b}, \\ &K_{\star,\star} - K_{\star,a}(\tilde{V}_{a,b}^{-1} - \tilde{W}_{a,b}^{-1})K_{a,\star} - (K_{\star,\{a,b\}}\tilde{V}_{a,b}^{-1}\mathbf{y}_{a,b})^2\big), \end{aligned} \quad (10)$$

where we have

$$\begin{aligned} \tilde{V}_{a,b}^{-1} &= \sum_j p(\sigma_b^j\,|\,\mathbf{y}_{a,b}) \sum_i p(\sigma_a^i\,|\,\mathbf{y}_a)(V_{a,b}^{i,j})^{-1} \\ \tilde{W}_{a,b}^{-1} &= \sum_j p(\sigma_b^j\,|\,\mathbf{y}_{a,b}) \sum_i p(\sigma_a^i\,|\,\mathbf{y}_a) \\ &\qquad (V_{a,b}^{i,j})^{-1}\mathbf{y}_{a,b}\mathbf{y}_{a,b}^\mathsf{T}(V_{a,b}^{i,j})^{-1}. \end{aligned}$$

Now, using the inversion by partitioning formula (PRESS, *et al.*, 1992, Section 2.7),

$$(V_{a,b}^{i,j})^{-1} = \begin{bmatrix} S_a^{i,j} & -S_a^{i,j}K_{a,b}(V_b^j)^{-1} \\ -(V_b^j)^{-1}K_{b,a}S_a^{i,j} & (V_b^j)^{-1} + (V_b^j)^{-1}K_{b,a}S_a^{i,j}K_{a,b}(V_b^j)^{-1} \end{bmatrix}$$

where $S_a^{i,j} = (V_a^i - K_{a,b}(V_b^j)^{-1}K_{b,a})^{-1}$. Note that $(V_{a,b}^{i,j})^{-1}$ is affine in $S_a^{i,j}$, so that when $V_a \gg K_{a,b}V_b^{-1}K_{b,a}$, $(V_{a,b}^{i,j})^{-1}$ is effectively affine in $(V_a^i)^{-1}$. This is true if given $\mathcal{D}_b$, it is impossible to accurately predict $\mathcal{D}_a$. This might be the case if $\mathcal{D}_a$ represents a lot of information relative to $\mathcal{D}_b$ (if, for example, $\mathcal{D}_a$ is our entire history of observations where $\mathcal{D}_b$ is simply the most recent observation), or if $\mathcal{D}_b$ and $\mathcal{D}_a$ are simply not particularly well correlated. On this basis, we make approximation 4. Additionally noting that $\sum_i p(\sigma_a^i \,|\, \mathbf{y}_a) = 1$, we have [2]

$$\tilde{V}_{a,b}^{-1} \simeq \sum_j p(\sigma_b^j \,|\, \mathbf{y}_{a,b}) \begin{bmatrix} \tilde{V}_a & K_{a,b} \\ K_{b,a} & V_b^j \end{bmatrix}^{-1},$$

$$\tilde{V}_{a,b} \simeq \begin{bmatrix} \tilde{V}_a & K_{a,b} \\ K_{b,a} & \tilde{V}_{b|a} + K_{b,a}\tilde{V}_a^{-1}K_{a,b} \end{bmatrix}$$

$$\tilde{V}_{b|a}^{-1} = \sum_j p(\sigma_b^j \,|\, \mathbf{y}_{a,b})(V_b^j - K_{b,a}\tilde{V}_a^{-1}K_{a,b})^{-1}.$$

Note that the lower right hand element of $\tilde{V}_{a,b}$ defines the noise variance to be associated with observations $\mathcal{D}_b$. In effect, we represent each observation as having a known variance lying between σ_n^2 and σ_f^2. The more likely an observation's faultiness, the closer its assigned variance will be to the (large) fault variance and the less relevant it will become for inference about the latent process. This approximate observation is then used for future predictions; we need never consider the full sum over all observations.

We now turn to $\tilde{W}_{a,b}^{-1}$. Unfortunately, even if $V_a \gg K_{a,b}V_b^{-1}K_{b,a}$, $\tilde{W}_{a,b}^{-1}$ is quadratic in $(V_a^i)^{-1}$. We will nonetheless again make approximation 4 and assume that $\tilde{W}_{a,b}^{-1}$ is affine in $(V_a^i)^{-1}$. The quality of our approximation for $\tilde{W}_{a,b}^{-1}$ is much less critical than for $\tilde{V}_{a,b}^{-1}$, because the former only influences the variance of our predictions for the current predictant; any flaws in that approximation will not be propagated forward. Further, of course, if one probability dominates, $p(\sigma_a^i \mid \mathbf{y}_a) \gg p(\sigma_a^{i'} \mid \mathbf{y}_a), \forall i' \neq i$, then the approximation is valid. With this,

$$\tilde{W}_{a,b}^{-1} \simeq \sum_j p(\sigma_b^j \,|\, \mathbf{y}_{a,b}) \begin{bmatrix} \tilde{V}_a & K_{a,b} \\ K_{b,a} & V_b^j \end{bmatrix}^{-1} \mathbf{y}_{a,b}\mathbf{y}_{a,b}^{\mathsf{T}} \begin{bmatrix} \tilde{V}_a & K_{a,b} \\ K_{b,a} & V_b^j \end{bmatrix}^{-1}.$$

If we now receive further data $\mathcal{D}_c$, our existing data is simply treated as old data ($a \leftarrow \{a, b\}$, $b \leftarrow c$), and another iteration of our algorithm performed. At each iteration, we are able to return the predictions for the latent variable using (10) and the posterior probability of an observation's faultiness using (7). We can also return the marginal likelihood (9) for the purposes of training hyperparameters.

[2] $\tilde{V}_{b|a}^{-1}$ can be computed using the same trick as in (3) if b identifies a single observation and $j \in \{0, 1\}$. The Cholesky factor of $\tilde{V}_{a,b}$ required to solve the linear equations for our predictions can be efficiently determined (OSBORNE, 2010) using the previously evaluated Cholesky factor of $\tilde{V}_a$.

References

DERESZYNSKI, E. and DIETTERICH, T.G. (2011). Spatiotemporal models for anomaly detection in dynamic environmental monitoring campaigns. *ACM Transactions on Sensor Networks*.

OSBORNE, M.A. (2010). *Bayesian Gaussian Processes for Sequential Prediction, Optimisation and Quadrature*. Ph.D. thesis, University of Oxford. Available at `www.robots.ox.ac.uk/~mosb/full_thesis.pdf`.

OSBORNE, M.A., GARNETT, R., SWERSKY, K. and DE FREITAS, N. (2012). Prediction and fault detection of environmental signals with uncharacterised faults. In: *Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12)*.

PRESS, W.H., TEUKOLSKY, S.A., VETTERLING, W.T. and FLANNERY, B.P. (1992). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, New York, NY , USA.