Supplementary Material for:
Tilted Variational Bayes

A Expanding the Bound

To see the relation given, consider the marginal likelihood written in terms of our bound (10):

$$p(\mathbf{y}) = \sum_{i=1}^{n} \log p(y_i \,|\, \widetilde{\mathbf{y}}_{\setminus i}) + \mathbb{E}_{q(\mathbf{x})}\left[\log \frac{p(\mathbf{x})}{\prod_i p(x_i \,|\, \widetilde{\mathbf{y}}_{\setminus i})}\right] + \mathrm{KL}[q(\mathbf{x}) || p(\mathbf{x} \,|\, \mathbf{y})] \ . \tag{23}$$

Take the following relation, which is a re-arrangement of Bayes' rule

$$p(\mathbf{x}) = \frac{p(\mathbf{x} \,|\, \widetilde{\mathbf{y}}) p(\widetilde{\mathbf{y}})}{p(\widetilde{\mathbf{y}} \,|\, \mathbf{x})} \ , \tag{24}$$

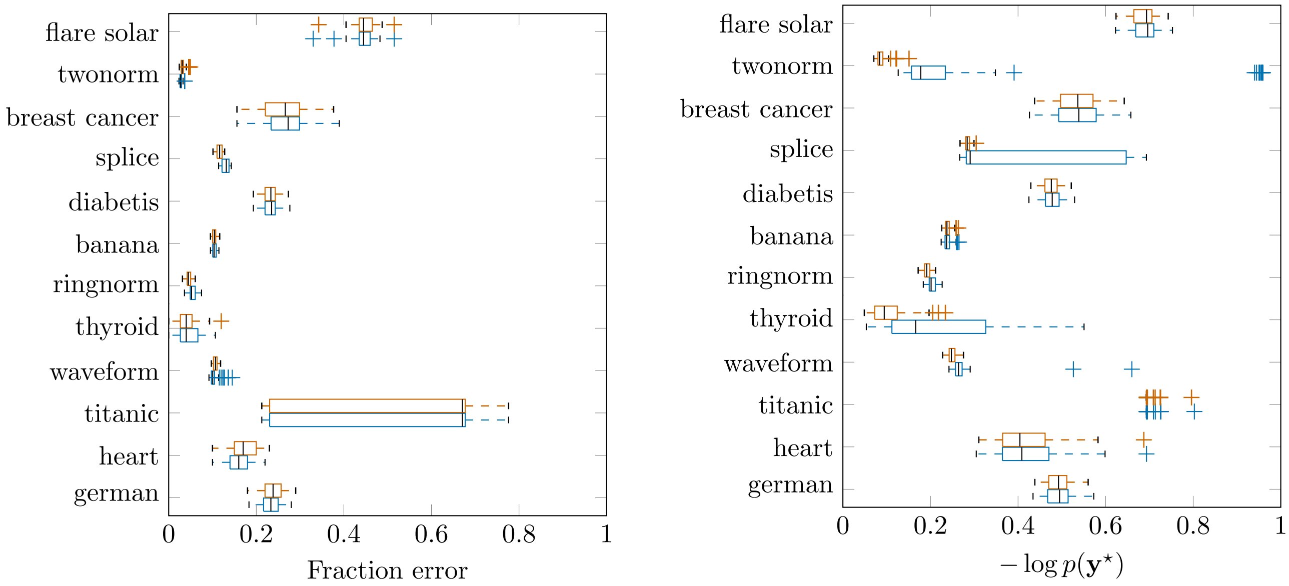
and substitute into the bound for $p(\mathbf{x})$ to give:

$$p(\mathbf{y}) = \log p(\widetilde{\mathbf{y}}) + \sum_{i=1}^{n} \log p(y_i \,|\, \widetilde{\mathbf{y}}_{\setminus i}) + \mathbb{E}_{q(\mathbf{x})}\left[\log \frac{p(\mathbf{x} \,|\, \widetilde{\mathbf{y}})}{p(\widetilde{\mathbf{y}} \,|\, \mathbf{x}) \prod_i p(x_i \,|\, \widetilde{\mathbf{y}}_{\setminus i})}\right] + \mathrm{KL}[q(\mathbf{x}) || p(\mathbf{x} \,|\, \mathbf{y})] \ . \tag{25}$$

Take the terms inside the expectation, and turn them into a pair of KL divergences by multiplying the top and bottom of the fraction by $q(\mathbf{x})$ and separating the terms:

$$\begin{aligned} p(\mathbf{y}) = {} & \log p(\widetilde{\mathbf{y}}) + \sum_{i=1}^{n} \log p(y_i \,|\, \widetilde{\mathbf{y}}_{\setminus i}) + \sum_{i=1}^{n} \log p(\widetilde{y}_i \,|\, \widetilde{\mathbf{y}}_{\setminus i}) \\ & - \sum_{i=1}^{n} \mathrm{KL}[q(x_i) || p(x_i \,|\, \widetilde{y})] - \mathrm{KL}[q(\mathbf{x}) || p(\mathbf{x} \,|\, \widetilde{\mathbf{y}})] + \mathrm{KL}[q(\mathbf{x}) || p(\mathbf{x} \,|\, \mathbf{y})] \ , \end{aligned} \tag{26}$$

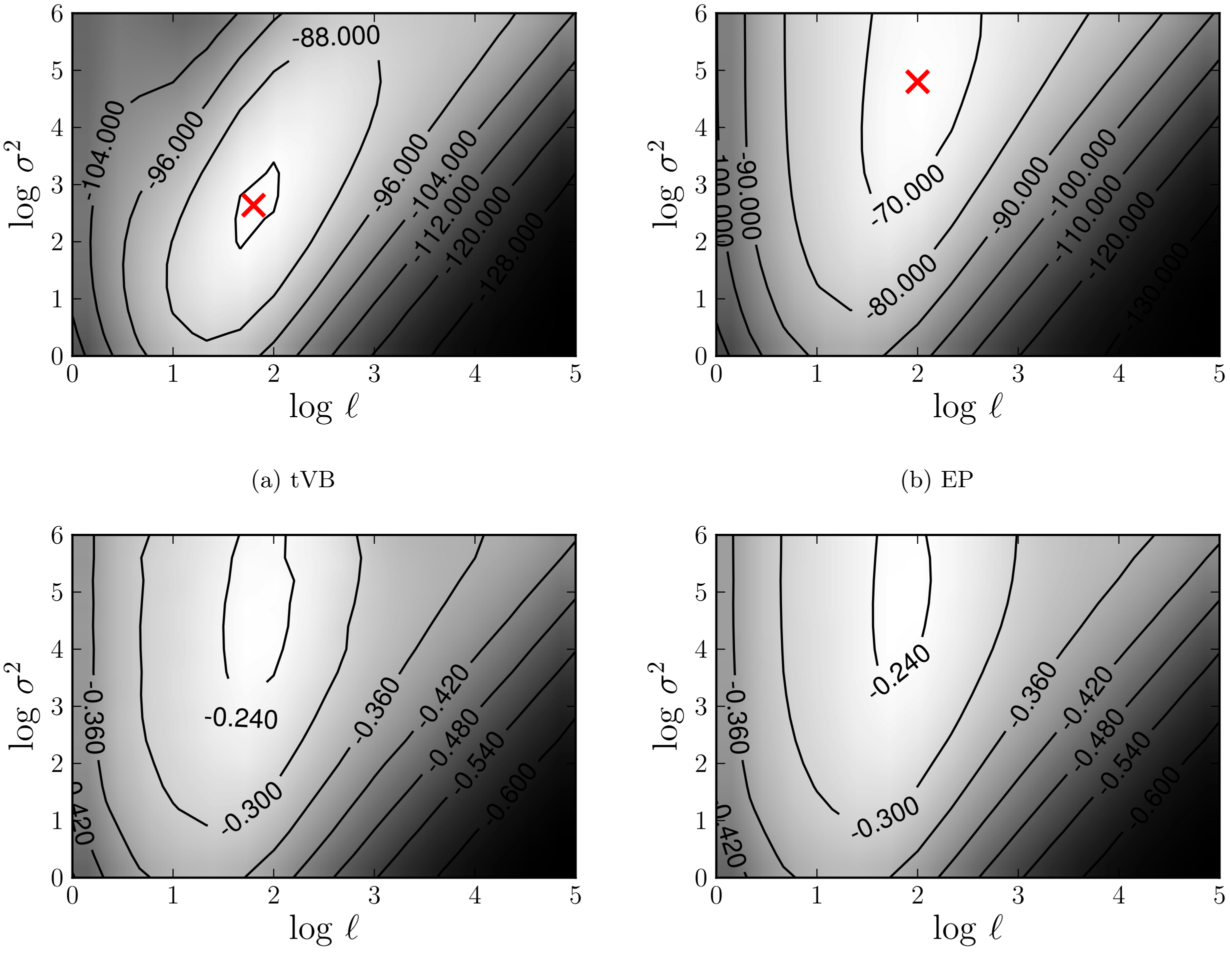
where the extra terms $p(\widetilde{y}_i \,|\, \widetilde{\mathbf{y}}_{\setminus i})$ appear as a consequence of the normalisation of the denominator.



(a) Holdout error (fraction) on classification benchmark datasets of Onoda et al. [2001] for our method (blue boxes) compared to standard EP (orange).

(b) Error (fraction) of our method using the parameters learned by EP (blue boxes) compared to standard EP (orange) using the parameters learned by out method.

Figure S.1: Data shown was created applying both methods on the classification benchmark datasets of Onoda et al. [2001].



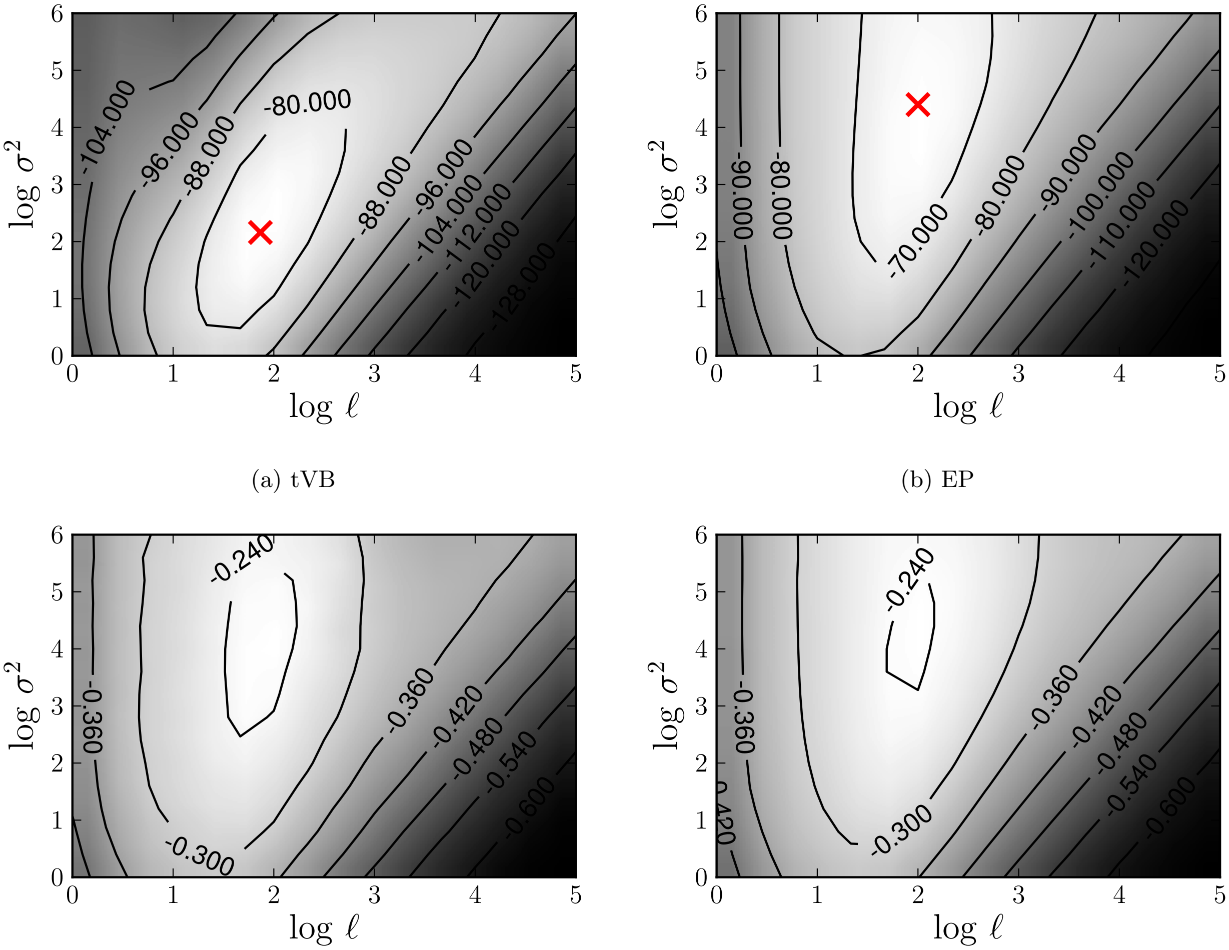
(a) tVB

(b) EP

(c) tVB_info

(d) EP_info

Figure S.2: white1_ionosphere_probit_N200



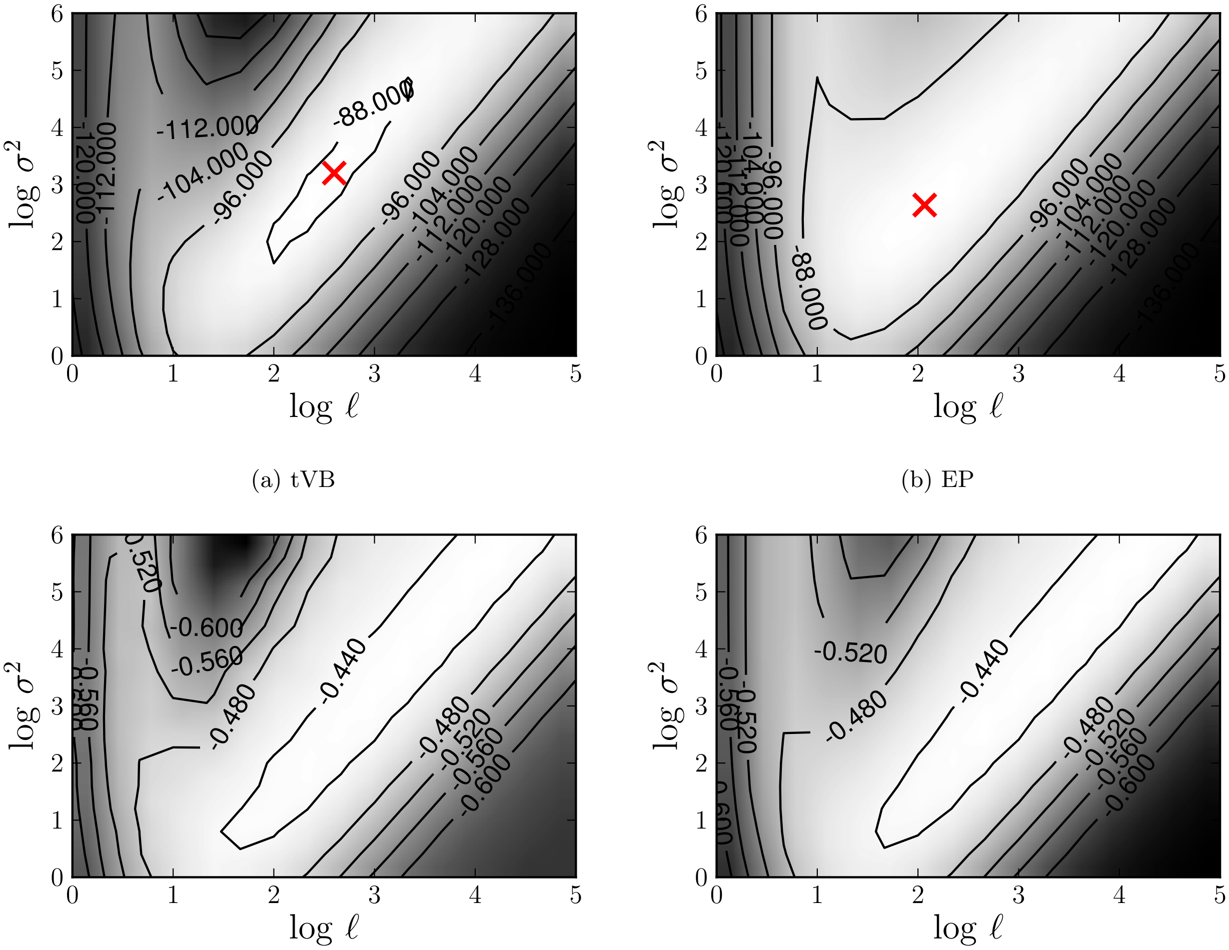
(a) tVB

(b) EP

(c) tVB_info

(d) EP_info

Figure S.3: white1_ionosphere_heaviside_N200



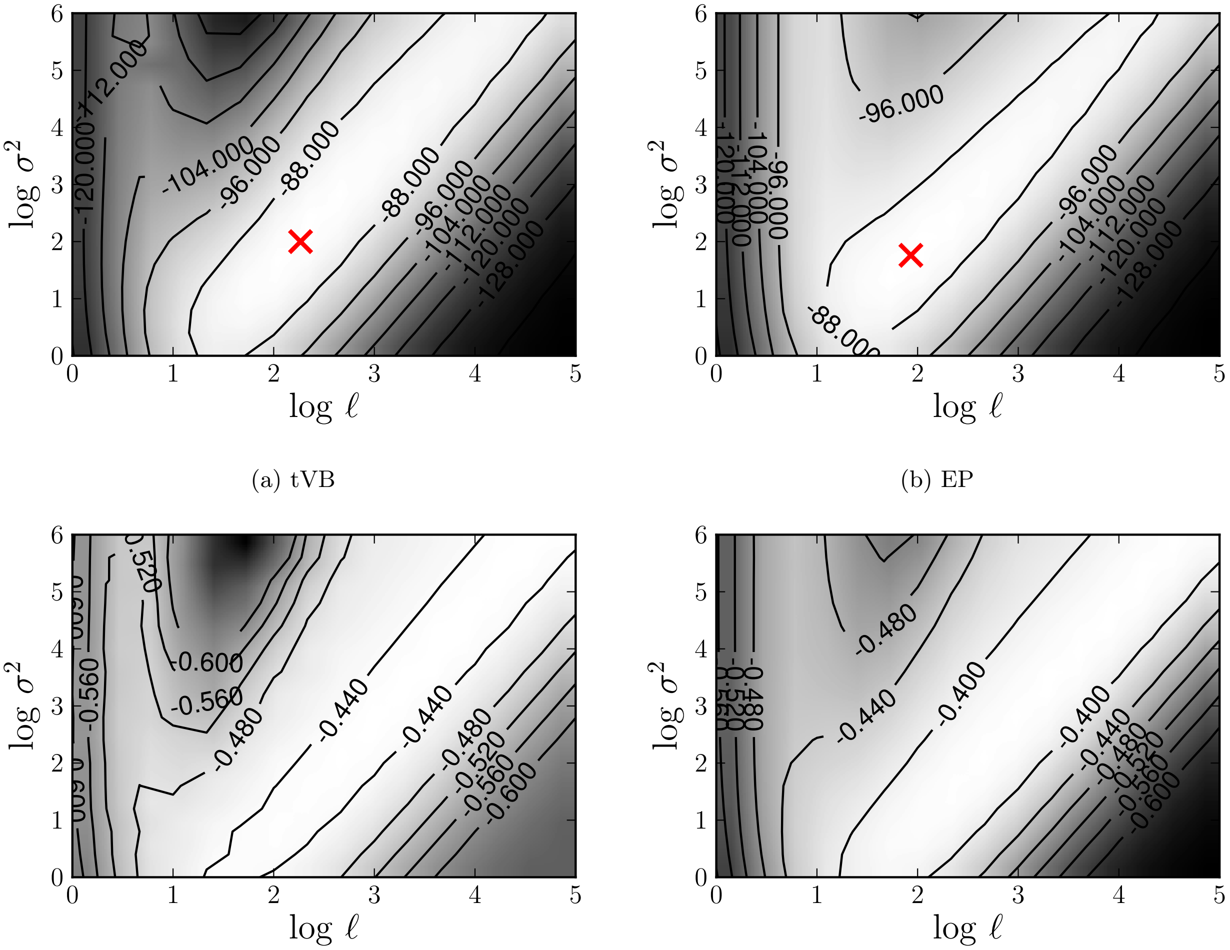
(a) tVB

(b) EP

(c) tVB_info

(d) EP_info

Figure S.4: white1_heart_probit_N200



(a) tVB

(b) EP

(c) tVB_info

(d) EP_info

Figure S.5: white1_heart_heaviside_N200

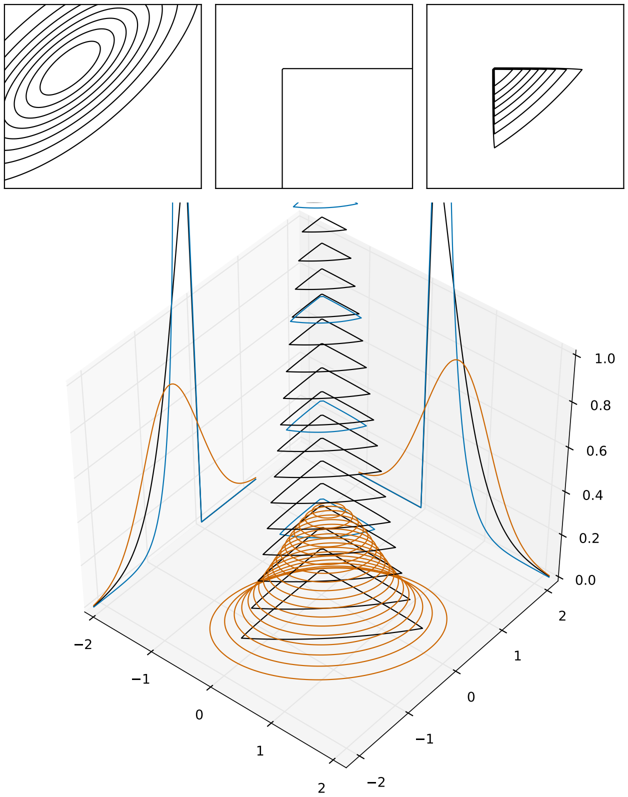


Figure S.6: Approximating Gaussian Process classification with only two data points. Colours and distributions are as for figure 1, but this time the two data points are correlated in the prior (top left), but are assigned different labels, leading to the posterior in the top-right.

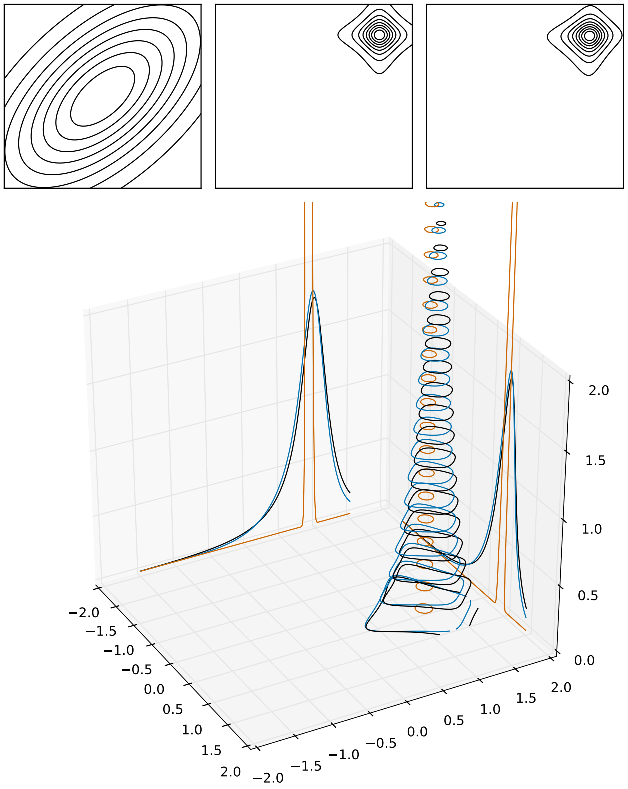


Figure S.7: Approximating robust Gaussian process regression with only two data points. Colours and distributions are as for previous figures. Here, two data that are correlated a-prior are observed at similar values, leading to a uni-modal posterior.