FIGURES — Ms. Number 1943

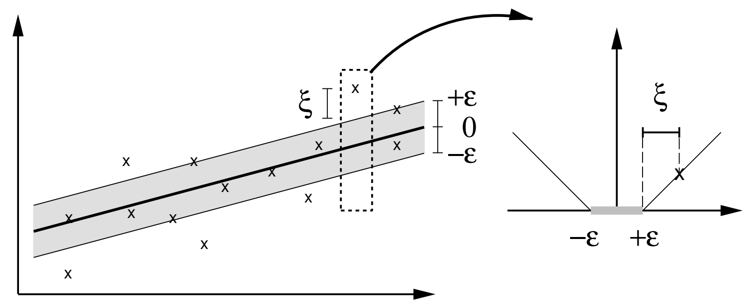

Figure 1: Schölkopf, Ms. Number 1943

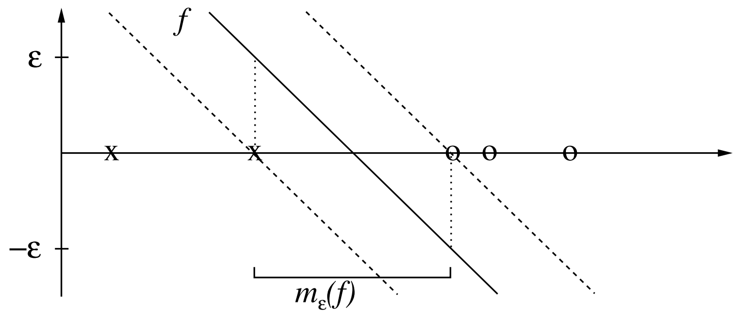

Figure 2: Schölkopf, Ms. Number 1943

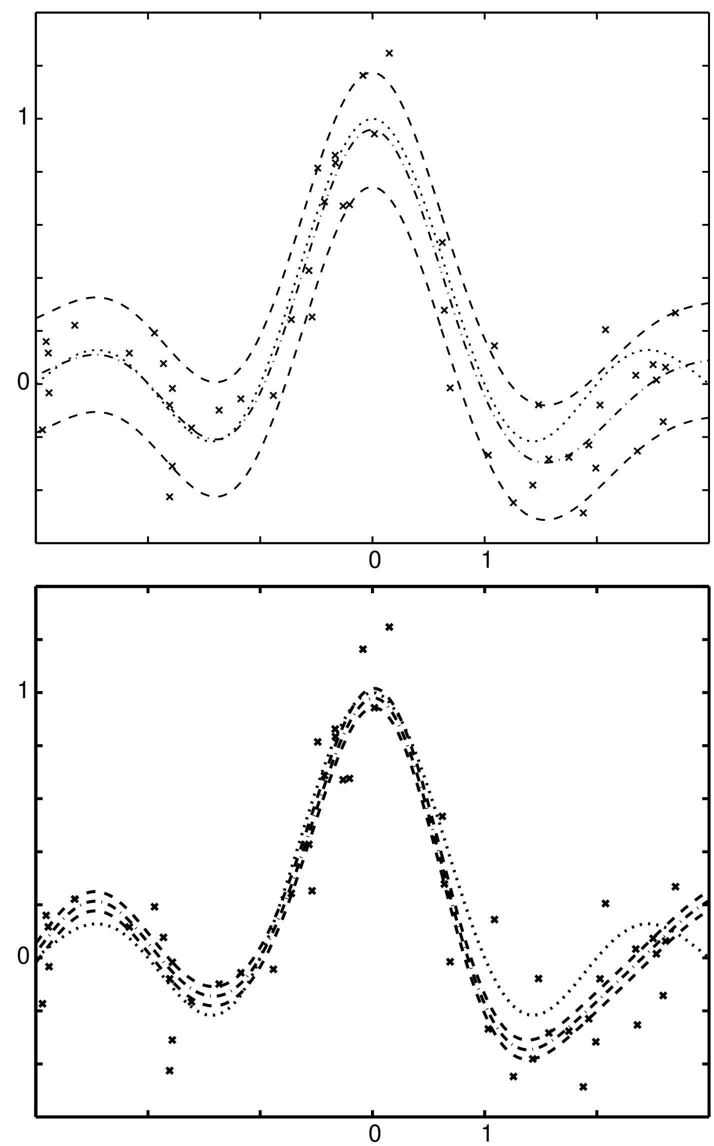

Figure 3: Schölkopf, Ms. Number 1943

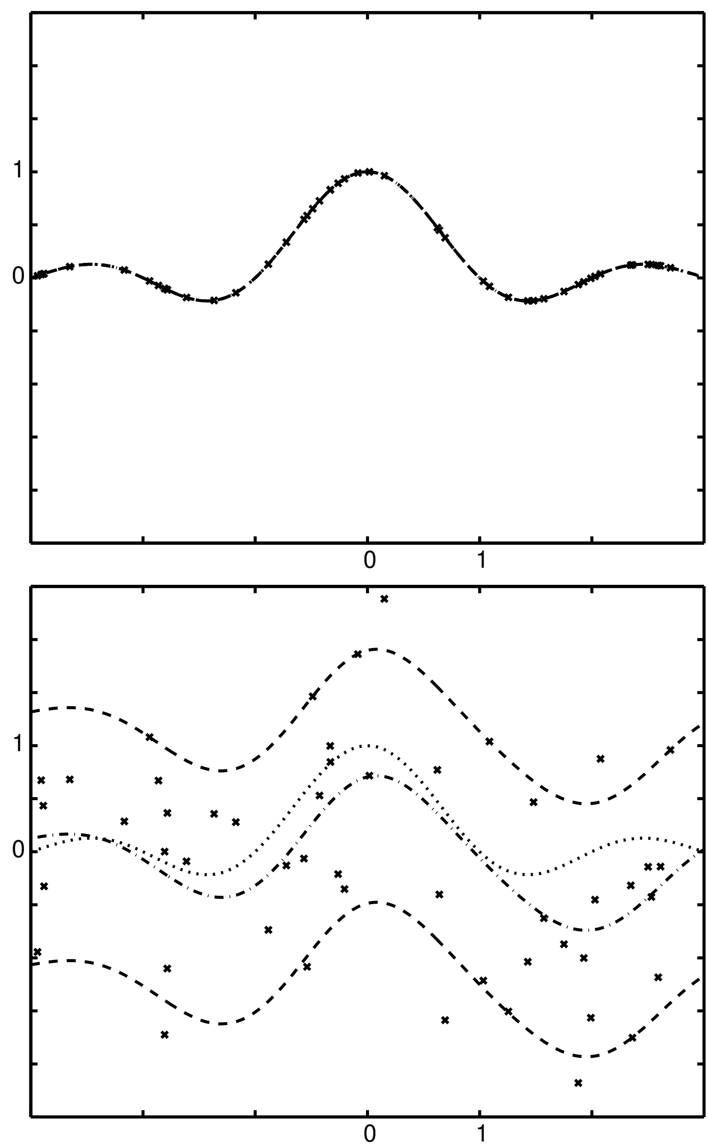

Figure 4: Schölkopf, Ms. Number 1943

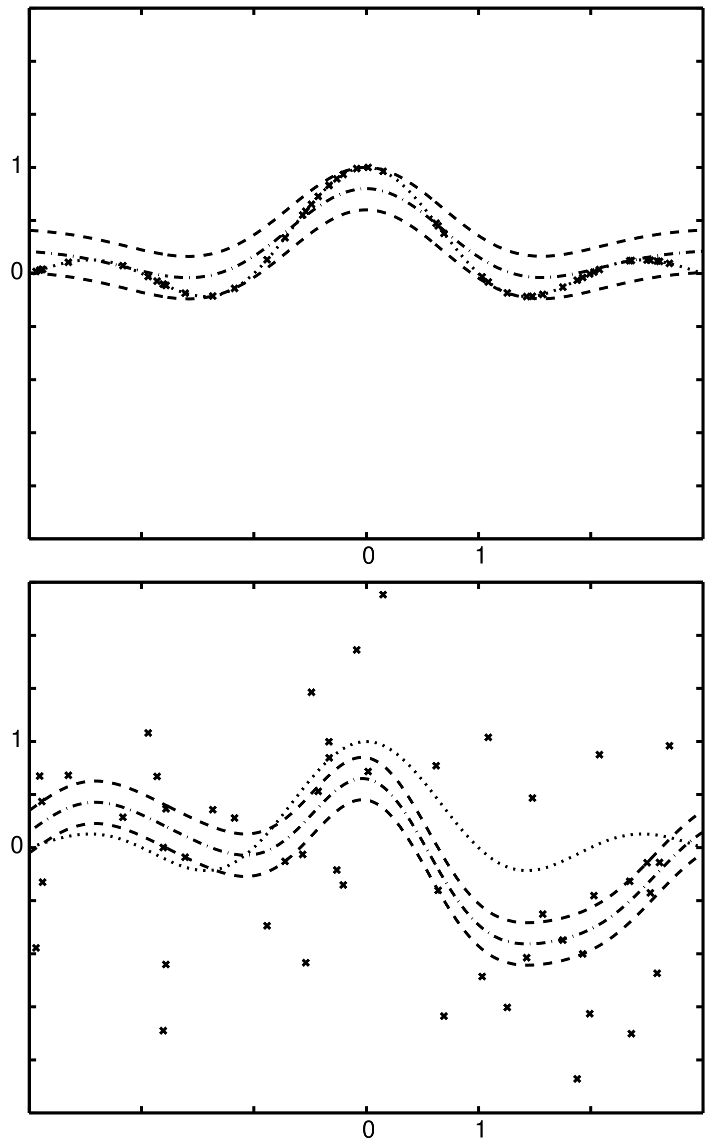

Figure 5: Schölkopf, Ms. Number 1943

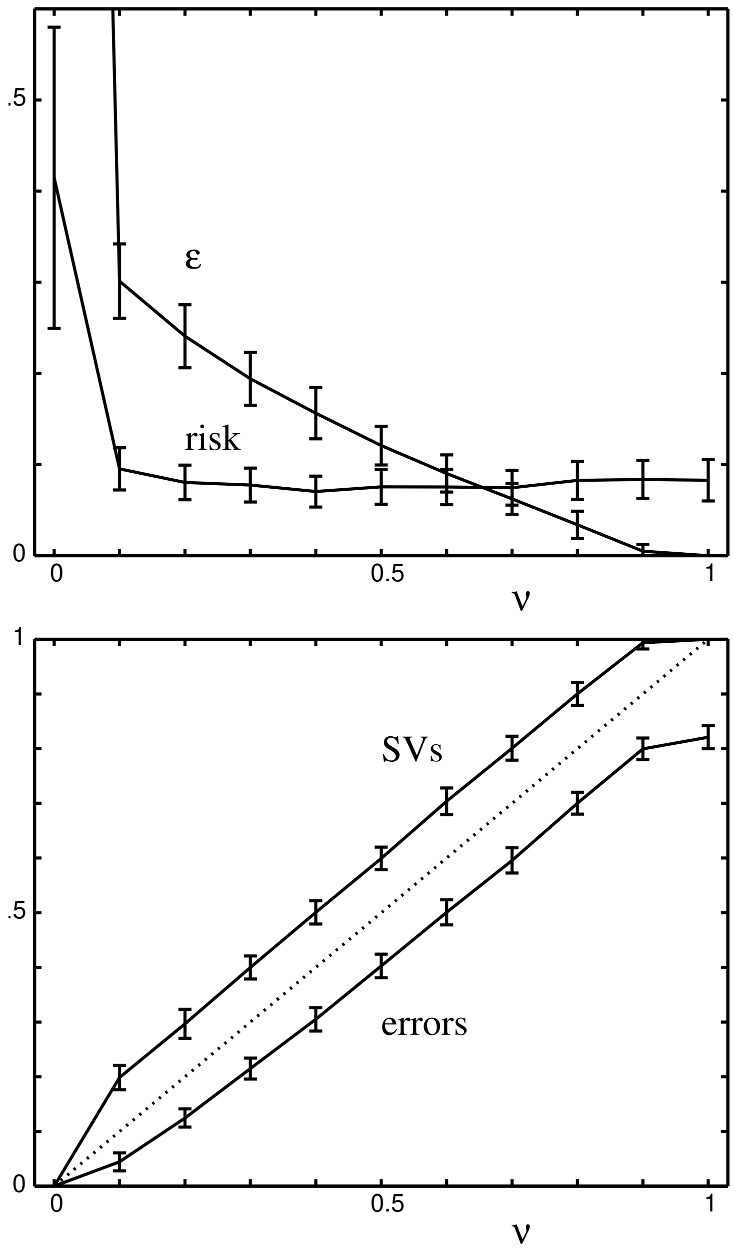

Figure 6: Schölkopf, Ms. Number 1943

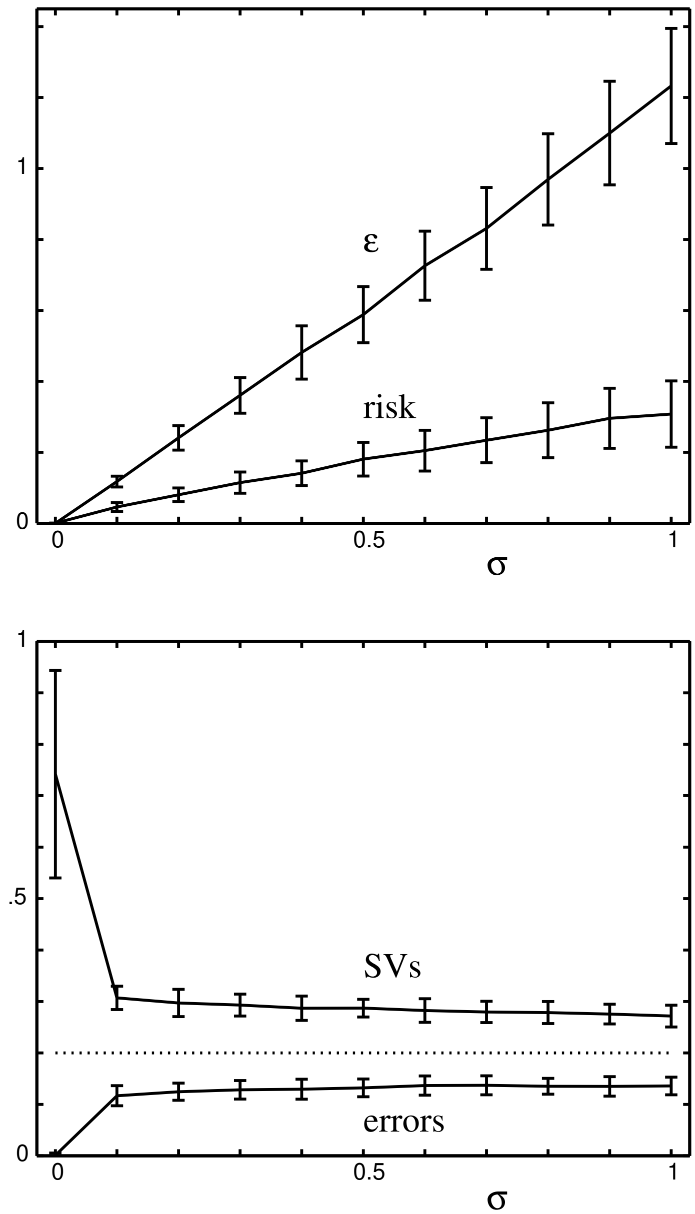

Figure 7: Schölkopf, Ms. Number 1943

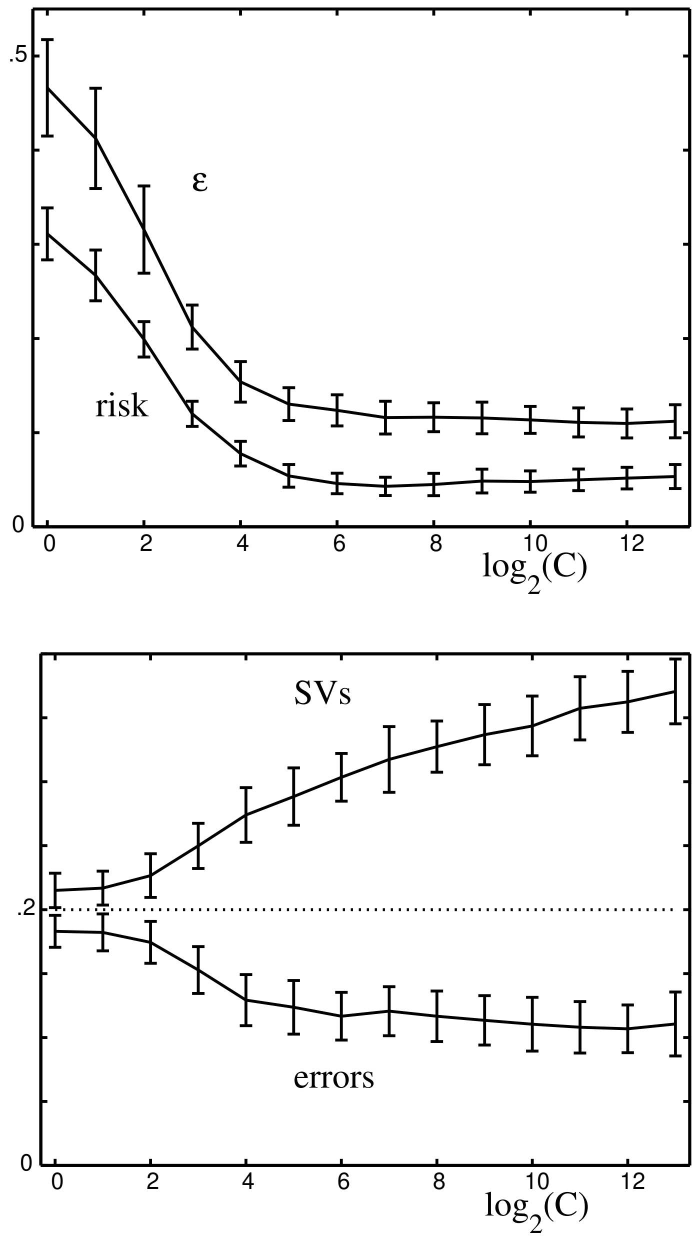

Figure 8: Schölkopf, Ms. Number 1943

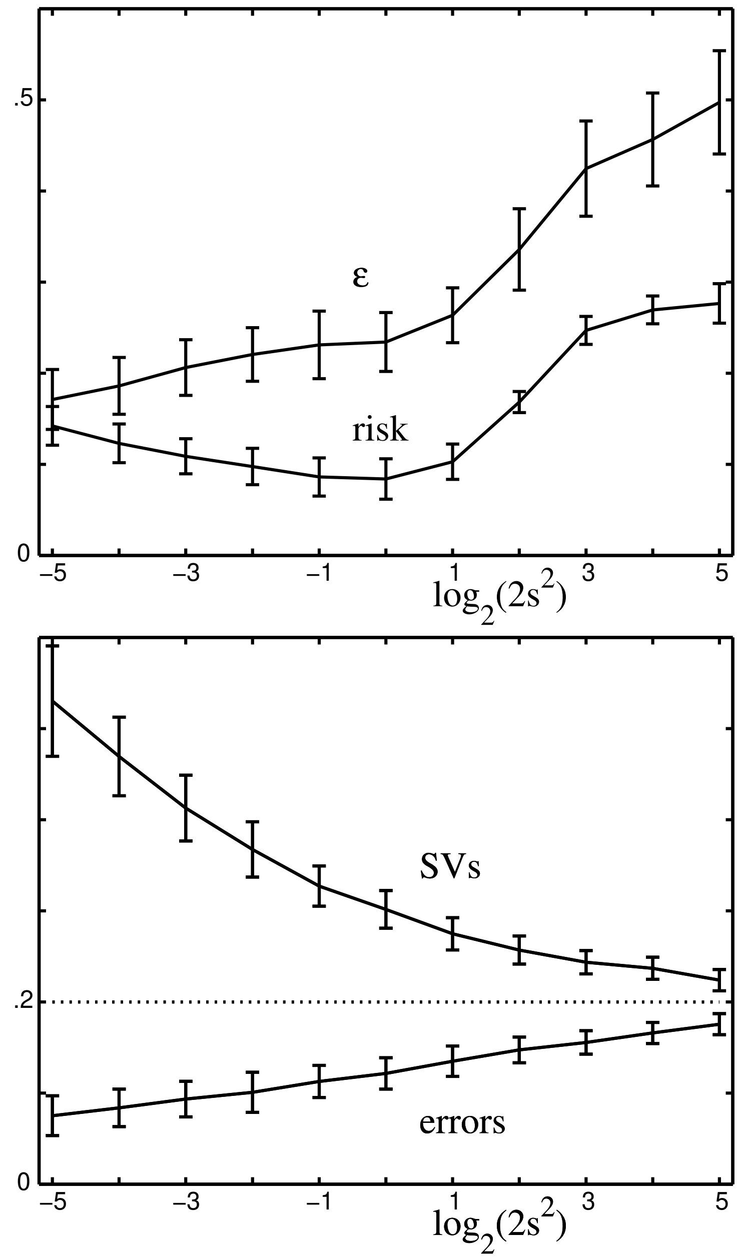

Figure 9: Schölkopf, Ms. Number 1943

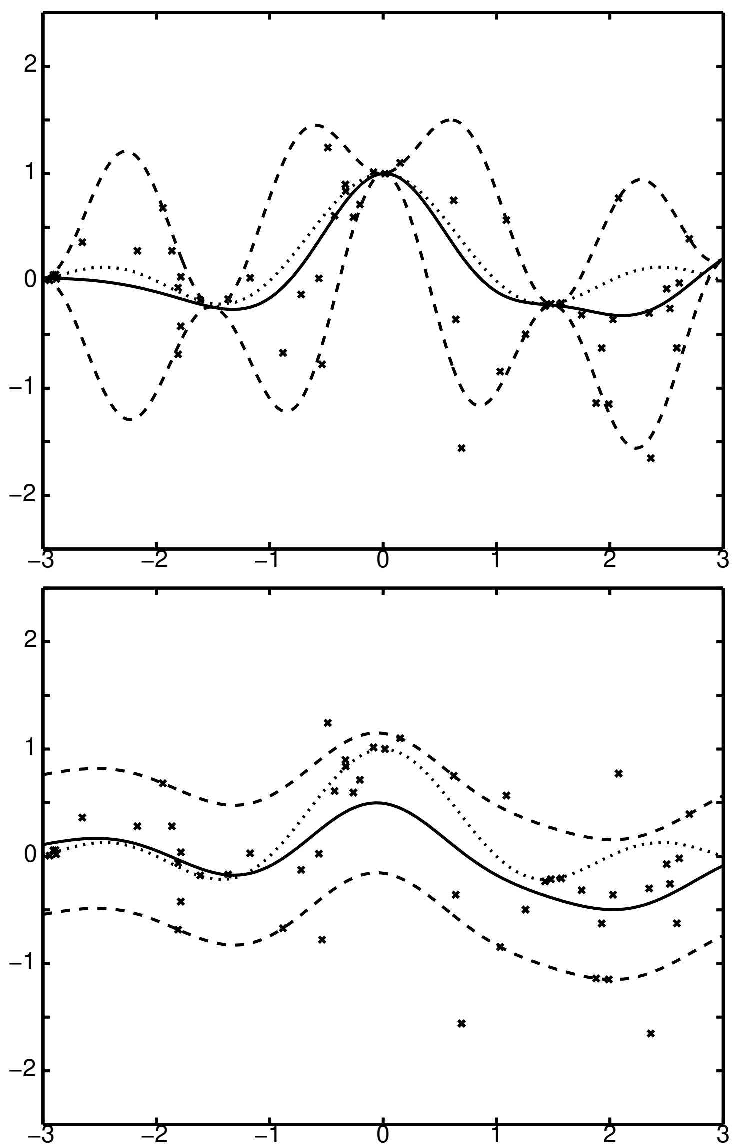

Figure 10: Schölkopf, Ms. Number 1943

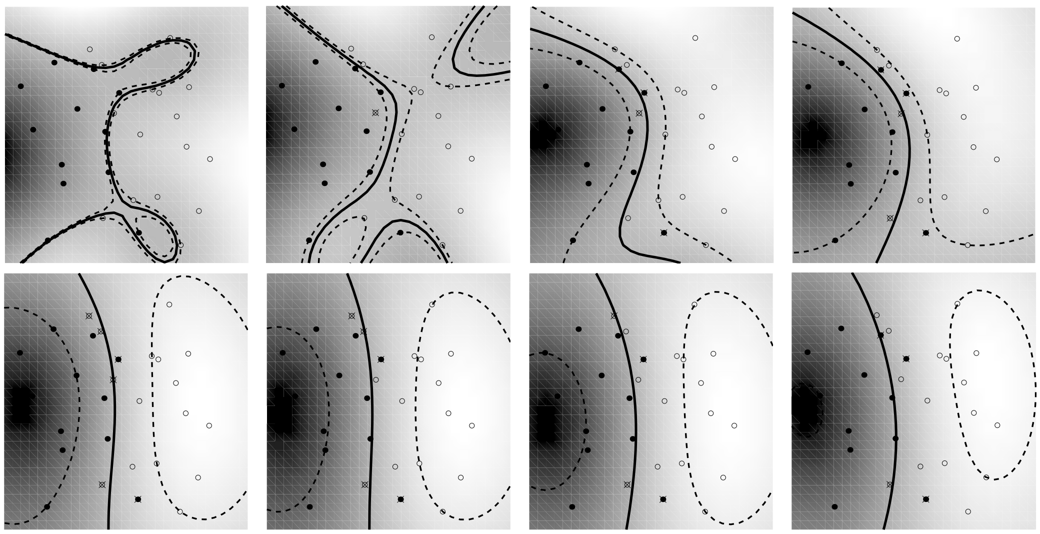

Figure 11: Schölkopf, Ms. Number 1943

Figure Captions

1. In SV regression, a desired accuracy ε is specified a priori. It is then attempted to fit a tube with radius ε to the data. The trade-off between model complexity and points lying outside of the tube (with positive slack variables ξ) is determined by minimizing (5).

2. 1-D toy problem: separate 'x' from 'o'. The SV classification algorithm constructs a linear function $f(x) = w \cdot x + b$ satisfying (45) ($\varepsilon = 1$). To maximize the margin $m_\varepsilon(f)$, one has to minimize $|w|$.

3. ν-SV regression with $\nu = 0.2$ (top) and $\nu = 0.8$ (bottom). The larger ν allows more points to lie outside the tube (see section 2). The algorithm automatically adjusts ε to 0.22 (top) and 0.04 (bottom). Shown are the sinc function (dotted), the regression f and the tube $f \pm \varepsilon$.

4. ν-SV regression on data with noise $\sigma = 0$ (top) and $\sigma = 1$ (bottom). In both cases, $\nu = 0.2$. The tube width automatically adjusts to the noise (top: $\varepsilon = 0$, bottom: $\varepsilon = 1.19$).

5. ε-SV regression (Vapnik, 1995) on data with noise $\sigma = 0$ (top) and $\sigma = 1$ (bottom). In both cases, $\varepsilon = 0.2$ — this choice, which has to be specified a priori, is ideal for neither case: in the upper figure, the regression estimate is biased; in the lower figure, ε does not match the external noise (Smola et al., 1998).

6. ν-SVR for different values of the error constant ν. Notice how ε decreases when more errors are allowed (large ν), and that over a large range of ν, the test error (risk) is insensitive towards changes in ν.

7. ν-SVR for different values of the noise σ. The tube radius ε increases linearly with σ (largely due to the fact that both ε and the $\xi_i^{(*)}$ enter the cost function linearly). Due to the automatic adaptation of ε, the number of SVs and of points outside the tube (errors) is, except for the noise-free case $\sigma = 0$, largely independent of σ.

8. ν-SVR for different values of the constant C. The top figure shows that ε decreases when the regularization is decreased (large C). Only very little, if any, overfitting occurs. In the bottom figure, note that ν upper bounds the fraction of errors, and lower bounds the fraction of SVs (cf. Proposition 1). The bound gets looser as C increases — this corresponds to a smaller number of examples ℓ relative to C (cf. table 2).

9. ν-SVR for different values of the Gaussian kernel width $2s^2$, using $k(x, x') = \exp(-|x - x'|^2/(2s^2))$. Using a kernel which is too wide results in underfitting; moreover, since the tube becomes too rigid as $2s^2$ gets larger than 1, the ε which is needed to accomodate a fraction $(1 - \nu)$ of the points increases significantly. In the bottom figure, it can again be seen that the speed of the uniform convergence responsible for the asymptotic statement given in Proposition 1 depends on the capacity of the underlying model. Increasing the kernel width leads to smaller covering numbers (Williamson et al., 1998) and therefore faster convergence.

10. Toy example, using prior knowledge about an x-dependence of the noise. Additive noise ($\sigma = 1$) was multiplied by the function $\sin^2((2\pi/3)x)$. *Top:* the *same* function was used as ζ as a parametric insensitivity tube (section 5). *Bottom:* ν-SVR with standard tube.

11. Toy problem (task: separate circles from disks) solved using ν-SV classification, using parameter values ranging from $\nu = 0.1$ (top left) to $\nu = 0.8$ (bottom right). The larger we select ν, the more points are allowed to lie inside the margin (depicted by dotted lines). As a kernel, we used the Gaussian $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2)$.

References

A. Smola, N. Murata, B. Schölkopf, and K.-R. Müller. Asymptotically optimal choice of ε-loss for support vector machines. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *Proceedings of the 8th International Conference on Artificial Neural Networks*, Perspectives in Neural Computing, pages 105 – 110, Berlin, 1998. Springer Verlag.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

R. C. Williamson, A. J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. Technical Report 19, NeuroCOLT, http://www.neurocolt.com, 1998.