Qualitative Robustness of Bootstrap Approximations for Kernel Based Methods *

Andreas Christmann[1], Matías Salibián-Barrera[2]
and Stefan Van Aelst[3]
[1] *University of Bayreuth*, [2] *University of British Columbia*
and [3] *Ghent University*

April 15, 2013

Abstract

Support vector machines (SVMs) are a very popular method in modern statistical learning theory and practice, where they are typically used for classification and regression purposes. SVMs can be thought of as penalized M-estimators, and their robustness properties have been explored in recent years. For example, it is know that if the loss function is Lipschitz continuous and the kernel is bounded, the resulting SVMs have a bounded influence function, bounded maxbias, and are qualitatively robust. Although there are many theoretical results available dealing with the consistency of SVMs and their rate of convergence, less is known about their asymptotic distribution and how to estimate it. The bootstrap (Efron, 1979) provides a consistent estimator for the distribution of a wide range of statistics. Recently it has been shown that this is also the case for SVMs. Here we study the robustness properties of these bootstrap distribution estimators for support vector machines. More specifically, we show that if T is an estimator based on a continuous operator from the space of probability measures over a compact metric space into a complete separable metric space, then bootstrap approximations for the distribution of T are stable, in the sense of being qualitatively robust. Intuitively, this means that the bootstrap distribution estimates are not severely affected by the presence of outliers in the data.

Keywords: Kernel-based methods, Support Vector Machines, Bootstrap, Robustness.

1 Introduction

Statistical learning refers to a relatively large collection of statistical methods designed to extract information from data, see for example, Vapnik (1995, 1998). In many cases where these techniques are applied one is interested in predicting the value of a specific "response" variable, based on a number of potentially relevant

*This research was partially supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada.

explanatory variables. Assume that our data consist of n observations (y_1, x_1), ..., (y_n, x_n) which we assume can be modelled as independent realizations of a random pair $(Y, X) \in (\mathcal{Y}, \mathcal{X})$, where $\mathcal{Y}$ and $\mathcal{X}$ denote the spaces on which the response and predictor variables take their values, respectively. Using the available data we want to find a function $f : \mathcal{X} \to \mathcal{Y}$ such that $f(x)$ is a good predictor for Y when $X = x$. It is convenient to introduce a loss function to measure the quality of different functions f as predictors for Y. Since Y is a random object, it is natural to consider the expected loss (or risk):

$$R(\mathrm{P}, f) = \mathbb{E}_{\mathrm{P}} \left[L(X, Y, f(X)) \right] , \tag{1}$$

where P denotes the joint distribution of the pair (Y, X) and $L : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is the loss function. Given a fixed loss function one can then try to find the optimal f in terms of its risk. For example, if $\mathcal{Y} \subset \mathbb{R}$ and $L(X, Y, f(X)) = (Y - f(X))^2$, then the optimal predictor is given by $\hat{f}(X) = \mathbb{E}_{\mathrm{P}}(Y|X)$, the conditional expectation of Y given X. Since the distribution P is typically unknown, one works with the empirical distribution P_n based on the data set, and (1) becomes $R(\mathrm{P}_n, f) = 1/n \sum_{i=1}^{n} L(x_i, y_i, f(x_i))$. It is clear that any function f that interpolates the data will minimize $R(\mathrm{P}_n, f)$. However, these functions will generally result in poor predictors when X takes values other than the observed x_1, ..., x_n. To avoid this problem, support vector machines (SVM) consider predictors $f \in \mathcal{H}$, where $\mathcal{H}$ is the reproducing kernel Hilbert functional space associated with a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Moreover, to avoid over-fitting, SVMs are defined as the solution to the following regularized risk minimization problem:

$$f_{(L,P,\lambda)} \;=\; \arg\min_{f \in \mathcal{H}} \;\; \mathbb{E}_P \left[L(X, Y, f(X)) \right] \;+\; \lambda \, \|f\|_{\mathcal{H}} \, , \tag{2}$$

where $\lambda \geq 0$ is a penalty parameter. The empirical version of this problem is:

$$f_{(L,\mathrm{P}_n,\lambda)} \;=\; \arg\min_{f \in \mathcal{H}} \;\; \frac{1}{n} \sum_{i=1}^{n} L(x_i, y_i, f(x_i)) \;+\; \lambda \, \|f\|_{\mathcal{H}} \, . \tag{3}$$

In what follows, we will fix the loss function L and penalty parameter λ and denote the solution of (3) by $\hat{f}_n$. For a comprehensive discussion of Support Vector Machines, see, for example, Steinwart and Christmann (2008).

2 Inference based on SVMs

An important component of a statistical analysis deals with quantifying the uncertainty associated with the estimate $\hat{f}_n$ and its associated predictions. For example, one may want to compute point-wise confidence bounds around the predictions $\hat{f}_n(x_i)$, $1 \leq i \leq n$. Recently, Hable (2012) showed that $\sqrt{n}(\hat{f}_n - f_{(L,P,\lambda)})$ converges to a zero-mean Gaussian process on $\mathcal{H}$ (the result is, in fact, slightly more general, allowing λ in (3) to depend on the sample). Although this result implies that each prediction $\hat{f}_n(x_i)$ has an asymptotic normal distribution, its variance is not easy to estimate. Hable (2013) gives a consistent estimator which seems to work well

for very large samples. Furthermore, confidence intervals based on a normal asymptotic distribution are symmetric, although asymmetric ones may sometimes be more appropriate, particularly for samples sizes that are not too large.

In this paper we consider using Efron's bootstrap (Efron, 1979) to approximate the finite-sample distribution of support vector machines. To fix ideas, consider a functional $S : \mathcal{M} \to \mathcal{W}$, where $\mathcal{M}$ is a set of probability measures and $\mathcal{W}$ denotes a metric space. Many estimators can be included in this framework. Simple examples include the sample mean (with functional $S(\mathrm{P}) = \int Z \, d\mathrm{P}$) and M-estimators (Huber, 1981). Let $\mathcal{B}(\mathcal{Z})$ be the Borel σ-algebra on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and denote the set of all Borel probability measures on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ by $\mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$. Then, equation (2) defines an operator $S : \mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z})) \to \mathcal{H}$ whose value is given by $S(P) = f_{(L,\mathrm{P},\lambda)}$. Furthermore, the estimator in (3) satisfies $f_{L,\mathrm{D}_n,\lambda} = S(\mathrm{P}_n)$.

More generally, let $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n$, be independent and identically distributed (i.i.d.) random variables with distribution P, and let

$$S_n(Z_1, \ldots, Z_n) = S(\mathrm{P}_n)$$

be the corresponding estimator, where P_n denotes the empirical distribution of the sample $Z_1, \ldots, Z_n$. Let $\mathscr{L}_n(S; \mathrm{P}) = \mathscr{L}(S(\mathrm{P}_n))$ be the sampling distribution of $S(\mathrm{P}_n)$. If P was known, one could construct a Monte Carlo estimate of this sampling distribution by drawing a large number of samples from P and repeatedly computing the estimator S_n. The bootstrap proposes to replace the unknown distribution P by an estimate $\hat{\mathrm{P}}$ in this process. In this paper we will consider $\hat{\mathrm{P}} = \mathrm{P}_n$. In other words, we will approximate the distribution of the estimator of interest by its sampling distribution when the data are generated by P_n. In symbols, the bootstrap proposes to use $\widehat{\mathscr{L}_n(S; \mathrm{P})} = \mathscr{L}_n(S; \mathrm{P}_n)$. Since this latter distribution is typically unknown, in practice one uses the Monte Carlo simulation method described above with P replaced by P_n. Note that obtaining a random sample from P_n is equivalent to drawing n observations with replacement from the original sample $Z_1, \ldots, Z_n$. Christmann and Hable (2013) have recently shown that the bootstrap is consistent for SVMs.

3 Qualitative robustness of the bootstrap for SVMs

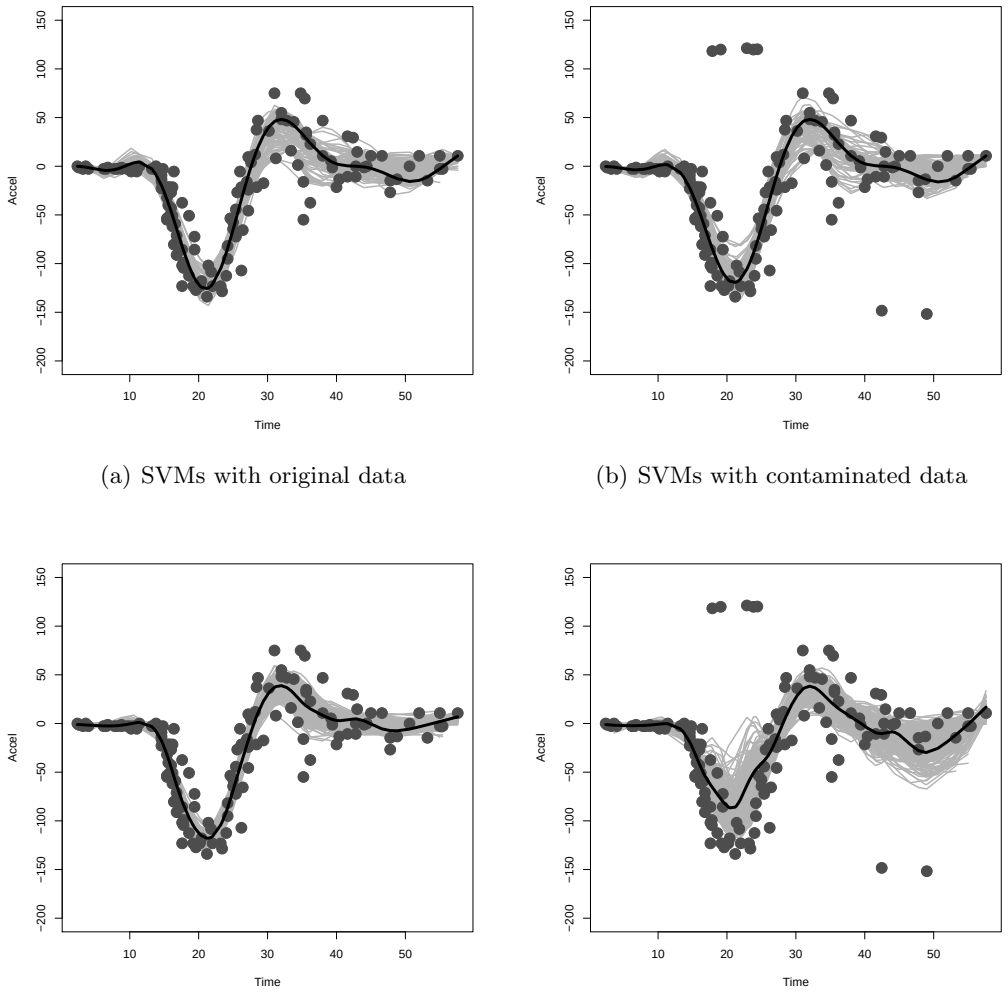
It is well known that many statistical methodologies are highly vulnerable to the presence of small proportions of observations deviating from the assumed model. Techniques that are able to remain informative even when the data may contain outliers are called robust. In the last 50 years many robust estimators have been proposed for a variety of different models and situations. Under relatively weak regularity conditions support vector machines have been shown to posses certain robustness properties (e.g. Christmann and Van Messem, 2008; Hable and Christmann, 2011). However, it is easy to see that a small proportion of outliers might severely damage the bootstrap distribution of estimators, even when these are robust. Intuitively, when the data contain outliers, they might be overly represented in the bootstrap samples, which may in turn negatively affect the estimated distribution of the estimator.

Different robustness properties have been studied in the literature. In this paper we focus on the concept of qualitative robustness (Hampel, 1971) for the bootstrap estimator of the sampling distribution (Cuevas and Romo, 1993). Informally, the bootstrap estimators are qualitatively robust if a small perturbation of the distribution that generated the data only produces, for a sample size large enough, a small deviation in the resulting bootstrap distribution estimator. In this sense, qualitative robustness of the bootstrap distribution relates to its infinitesimal stability.

Extending the work of Cuevas and Romo (1993), Christmann et al. (2013) show that the bootstrap distribution estimates of estimators defined by a functional that is continuous uniformly over neighbourhoods of distributions are qualitatively robust. Furthermore, Christmann et al. (2013) show that the following are sufficient conditions for this result to hold for SVMs: $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ is a compact metric space with $\mathcal{Y} \subset \mathbb{R}$; the loss function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$ is convex and uniformly Lipschitz continuous with respect to its third argument (i.e. there exists a constant $0 < C < +\infty$ such that $\sup_{x,y} |L(x, y, t) - L(x, y, t')| \leq C|t - t'|$ for any $t, t' \in \mathbb{R}$); the penalty parameter $0 < \lambda < \infty$; and the kernel is continuous and bounded by $\|k\|_\infty = (\sup_x k(x, x))^{1/2} < +\infty$. Since these conditions only involve the loss and kernel functions, but not the unknown distribution P, they are easy to check. In particular, they are satisfied by the hinge and logistic loss functions for classification problems, and by the L_1, logistic, ϵ-insensitive, pinball and Huber loss functions for regression problems. Admissible kernels include the Gaussian, Laplacian and Wendland radial basis function families (see Christmann et al., 2013).

4 Example

To illustrate the practical implications of our results we use the motorcycle data (Silverman, 1985). The data consist of 133 observations from a simulated motorcycle accident. The response is the head acceleration (in g's) and the predictor variable is the time elapsed from impact, in milliseconds. We consider a SVM with an ϵ-insensitive loss ($\epsilon = 0.001$) and a Gaussian radial basis function (RBF) kernel with $\gamma = 0.015$. We use the function `svm` in the `e1071` package for `R`, and set the `cost` parameter to 110. We also fit a penalized cubic spline as implemented in the package `SemiPar`. The optimal penalty term obtained with these data is `spar=4.83` and we keep it fixed throughout the rest of our analysis. Panels (a) and (c) in Figure 1 contain 200 bootstrapped fits for each estimator on the original data. We then added 7 mild outliers (around 5% of atypical observations) and re-computed both estimators on 200 bootstrap samples. The tuning parameters were kept fixed at the same values used with the "clean" data. Panels (b) and (d) display both sets of bootstrapped estimators when the data include these few mild outliers. Note that the bootstrap estimator of the distribution of the SVM estimates barely changes when outliers are present in the data. Penalized regression splines fits, however, are much more sensitive. Although these plots only represent one realization of the bootstrap estimate of the distribution of these regression methods, they illustrate the different degrees of sensitivity to the presence of outliers of the bootstrap distribution estimates.



(a) SVMs with original data

(b) SVMs with contaminated data

(c) Cubic splines with original data

(d) Cubic splines with contaminated data

Figure 1: Illustration of the stability of bootstrapped support vector machines when the data contain a small proportion of outliers. The first row contains 200 bootstrapped SVMs with and without 7 outliers (among 133 "good" points). The second row displays the corresponding results for a penalized cubic spline fit. The black lines show the corresponding estimated regression function for each case. The penalty parameters of both the SVMs and the penalized cubic splines were the same for the clean and contaminated data sets.

References

[1] Christmann, A. and Van Messem, A. (2008) Bouligand derivatives and robustness of support vector machines for regression. *Journal of Machine Learning Research*, **9**, 915-936.

[2] Christmann, A., Salibian-Barrera, M., Van Aelst, S. (2013) On the Stability of Bootstrap Estimators. To appear in *Robustness and Complex Data Structures*, Becker, C., Kuhnt, S. and Fried, R. Eds., Springer, Heidelberg, New York.

[3] Cuevas, A. and Romo, R. (1993) On robustness properties of bootstrap approximations. *Journal of Statistical Planning and Inference*, **37**, 181-191.

[4] Efron, b. (1979) Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, **7**, 1-26.

[5] Hable, R. (2012) Asymptotic Confidence Sets for General Nonparametric Regression and Classification by Regularized Kernel Methods. `arXiv:1203.4354v1 [stat.ML]`.

[6] Hable, R. (2012) Asymptotic Normality of Support Vector Machine Variants and Other Regularized Kernel Methods. *Journal of Multivariate Analysis*, **106**, 92117.

[7] Hable, R. (2013) Asymptotic Confidence Sets for General Nonparametric Regression and Classification by Regularized Kernel Methods.

[8] Hable, R. and Christmann, A. (2011) On Qualitative Robustness of Support Vector Machines. *Journal of Multivariate Analysis*, **102**, 993-1007.

[9] Hampel, F. R. (1971) A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, **42**, 1887-1896.

[10] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986) *Robust Statistics*. John Wiley and Sons, New York.

[11] Huber, P. J. (1981) *Robust Statistics*. John Wiley and Sons, New York.

[12] Silverman, B. W. (1985) Some aspects of the spline smoothing approach to non-parametric curve fitting. Journal of the Royal Statistical Society series B 47, 1-52.

[13] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*, Springer, New York.

[14] Vapnik, V. N. (1995) *The Nature of Statistical Learning Theory*. Springer, New York.

[15] Vapnik, V. N. (1998) *Statistical Learning Theory*. John Wiley and Sons, New York.