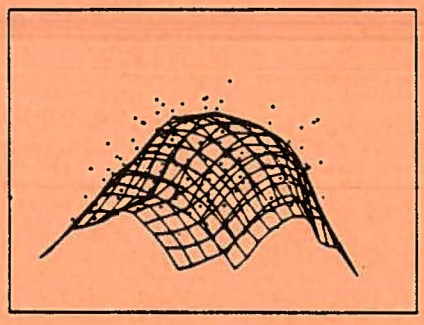
# NONPARAMETRIC CONDITIONAL ESTIMATION

*Arthur B. Owen*

Technical Report No. 25

**February 1987**

**Laboratory for Computational Statistics**



**Department of Statistics**
**Stanford University**

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| **FEB 1987** | | **00-00-1987 to 00-00-1987** |

| 4. TITLE AND SUBTITLE | 5a. CONTRACT NUMBER |
|---|---|
| **Nonparametric Conditional Estimation** | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| **Stanford University,Department of Statistics,Stanford,CA,94309** | |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT

**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

14. ABSTRACT

**Many nonparametric regression techniques (such as kernels, nearest neighbors, and smoothing splines) estimate the conditional mean of Y given X = :.r: by a weighted sum of observed Y values, where observations with X values near :.r: tend to have larger weights. In this report the weights are taken to represent a finite signed measure on the space ofY values. This measure is studied as an estimate of the conditional distribution of Y given X= :.r:. From estimates of the conditional distribution, estimates of cqnditional means, standard deviations, quantiles and other statistical functionals may be computed. Chapter 1 illustrates the computation of conditional quantiles and conditional survival probabilities on the Stanford Heart Transplant data. Chapter 2 contains a survey of nonparametric regression methods and introduces statistical metrics and von Mises' method for later use. Chapter 3 proves some consistency results. The estimated conditional distribution of Y is shown to be consistent in the following sense: the Prohorov ?distance between the estimated and true conditional distributions converges in probability to zero. The required conditions are: that the distribution of Y given X = :.r: vary continuously with :.r:, that the weights regarded as a measure on the X space converge in probability to a point mass at :.r:, and that a measure of the effective local sample size tend to infinity in probability. A slight strengthening of the conditions allows one to establish almost sure consistency. Consistency of Prohorov-continuous (i.e. robust) functionals follows immediately. In the above, the X and Y spaces are complete separable metric spaces. In case Y is the real line, weak and strong consistency results are established for the Kolmogorov-Smirnov and the Vasserstein metrics under stronger conditions. Chapter 4 provides conditions under which the suitably normalized errors in estimating the conditional distribution of Y have a Brownian limit. Using von Mises' method, asymptotic normality is obtained for nonparametric conditional estimates of compactly differentiable statistical functionals.**

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT<br>**unclassified** | b. ABSTRACT<br>**unclassified** | c. THIS PAGE<br>**unclassified** | **Same as Report (SAR)** | **109** | |

**Standard Form 298 (Rev. 8-98)**
Prescribed by ANSI Std Z39-18

NONPARAMETRIC CONDITIONAL ESTIMATION

**Abstract**

Many nonparametric regression techniques (such as kernels, nearest neighbors, and smoothing splines) estimate the conditional mean of $Y$ given $X = x$ by a weighted sum of observed $Y$ values, where observations with $X$ values near $x$ tend to have larger weights. In this report the weights are taken to represent a finite signed measure on the space of $Y$ values. This measure is studied as an estimate of the conditional distribution of $Y$ given $X = x$. From estimates of the conditional distribution, estimates of conditional means, standard deviations, quantiles and other statistical functionals may be computed.

Chapter 1 illustrates the computation of conditional quantiles and conditional survival probabilities on the Stanford Heart Transplant data. Chapter 2 contains a survey of nonparametric regression methods and introduces statistical metrics and von Mises' method for later use.

Chapter 3 proves some consistency results. The estimated conditional distribution of $Y$ is shown to be consistent in the following sense: the Prohorov distance between the estimated and true conditional distributions converges in probability to zero. The required conditions are: that the distribution of $Y$ given $X = x$ vary continuously with $x$, that the weights regarded as a measure on the $X$ space converge in probability to a point mass at $x$, and that a measure of the effective local sample size tend to infinity in probability. A slight strengthening of the conditions allows one to establish almost sure consistency. Consistency of Prohorov-continuous (i.e. robust) functionals follows immediately. In the above, the $X$ and $Y$ spaces are complete separable metric spaces. In case $Y$ is the real line, weak and strong consistency results are established for the Kolmogorov-Smirnov and the Vasserstein metrics under stronger conditions.

Chapter 4 provides conditions under which the suitably normalized errors in estimating the conditional distribution of $Y$ have a Brownian limit. Using von Mises' method, asymptotic normality is obtained for nonparametric conditional estimates of compactly differentiable statistical functionals.

This research supported by Office of Naval Research Contract N00014-83-K-0472; supported National Science Foundation Grant DMS86-00235 and issued as Department of Statistics Report No. 265; supported by the Department of Energy Contract DE-AC03-76SF00515 and issued as Stanford Linear Accelerator Center Report No. 309; and by a Natural Sciences and Engineering Research Council of Canada Postgraduate Scholarship.

# Table of Contents

# 1 Introduction

This report is concerned with estimation of aspects of the conditional distribution of a random variable $Y$ given another random variable $X$.

The most familiar example is the estimation of the conditional expectation of $Y$ given $X = x$. When this is carried out for a large number of $x$'s the results can be presented as a curve. The curve is usually plotted together with the data used to estimate it. It then may be used in informal data analysis, or its shape may be used to select or confirm a parametric model, or finally it may be used for the prediction of $Y$ values corresponding to future $X$ values.

No serious analysis of a single sample of data would stop at reporting the sample mean. Similarly in the bivariate case there is a need to go beyond the examination of the estimated conditional mean. Estimating conditional standard deviations is a natural first step in this direction. For the data analyst, a plot with a running mean and with curves equal to the running mean plus or minus two running standard deviations would be useful in assessing whether the data are heteroscedastic. If they are, such a plot would show where and by how much the the variation fluctuates. Much has been written about how hard it is to perceive a conditional mean in a scatterplot without the presence of an estimating curve. Surely the same is true about the perception of conditional variance or skewness. Where conditional variances are equal, they can seem to be larger where the marginal distribution of $X$ is greater, because visual impressions are dominated by extreme observations. For prediction, an estimate of the conditional scale of $Y$ would seem essential in order to provide an interval about the prediction.

Often in one sample situations the mean and standard deviation are not the most

convenient way to study the data. For example in survival analysis the mean of the failure times is difficult to assess in the presence of censoring but the median and other quantiles are readily obtained. Where outliers are suspected the mean is often replaced by a trimmed mean or some other robust estimator. Low quantiles are the natural choice when one studies breaking strengths of materials. In bivariate situations where the $Y$ values are subject to censoring or outliers, or in which extreme $Y$ quantiles are of interest it is natural to compute running quantiles or robust estimators instead of a regression curve.

We suppose that the estimation is performed in two stages. First at each point $x$ in a grid, the conditional distribution of $Y$ given $x$ is estimated. Then at each grid point a function that takes distributions and returns means, variances, quantiles or whatever is applied. The results are plotted against the grid points and joined up to provide the estimate of the curve. The distribution estimators considered are all nonparametric and discrete. They are reweightings of the $Y$ sample adapted from weights used in nonparametric regression techniques.

Figure 1 (page 7) presents the Stanford Heart Transplant data. The horizontal axis is the age at entry to the transplant program of a patient. The vertical axis is the base 10 logarithm of the number of days the patient was observed to survive after the operation. There are 157 data points. Points marked "X" represent times of death and points marked "+" represent censoring times. All that is known about the time of death for a censored patient is that it exceeds the time recorded.

Other variables were recorded, but the survival time is of primary interest and the age at entry is the most useful predictor of it. The most notable feature of this data is the drop-off in survival at entry ages over 50 years. This feature is hard to see in the raw data, especially because of the censoring.

The observed ages were used to form a grid and at each such age a reweighting of the ordered pairs (survival time, censoring indicator) was obtained. (The weights were based

on symmetric triangular nearest neighbors. See Sec. 2.2. The $k = 23$ nearest neighbors on each side of the target point were used.) Because interest centers on the distribution of survival times, the censoring is a nuisance. It is usually handled by calculating the product-limit estimator of the survival function. A convenient way to do this for weighted distributions is via the "redistribute to the right" algorithm of Efron (1967).

In Figure 2 (page 8) there are 5 estimated conditional survival quantiles corresponding to levels (.1,.3,.5,.7,.9). The quantile curve labelled .7 represents an estimate of the (log) time at which 70% of patients will still be alive given their respective ages at entry. Some of the survival quantiles are themselves censored. For example, the time at which only 10% of 25 year olds will remain is censored. This is because there was more than 10% censoring in the data used to estimate the survival time given an age of 25 years at entry.

The sharp drop in the median survival time is also evident in the 70% survival curve and to a lesser extent in the other survival deciles.

Another way to look at the ensemble of estimated survival probabilities is to estimate for each $x$, the conditional probability of survival past a certain time. Figure 3 (page 9) contains a plot of such curves for the probability of survival past 10, 100, and 1000 days. Also plotted are the probabilities of surviving some interpolated times, roughly 3, 32 and 316 days. (The estimated 3 day survival probability is 1 for older patients, so that curve disappears at the right of the figure.) The probability of survival past 100 days drops sharply at the age of 50. So does the probability of survival past 1000 days and the curves are roughly parallel. The probability of survival past 10 days differs markedly from the curves for longer survival times—it is almost flat.

The sort of calculations illustrated on this data are similar to those that a data analyst might make on a univariate sample. The next natural step might be to compute conditional hazard functions and plot a hazard surface, using age at time of entry and days since the operation as coordinates. One might also apply Greenwood's formula conditionally to estimate conditional standard deviations of the probabilities in Figure

3. Approximate confidence intervals for conditional probabilities could be used to obtain confidence intervals for conditional quantiles. *Any* functional that a statistician applies to a sample, might in the bivariate case be applied conditionally on $X$.

Methods like these will be analyzed by considering separately the properties of the functional and the distribution estimator. This approach has certain economies: for example, if the distribution estimator is suitably consistent then so are running versions of any functional that is robust at the underlying distributions of $Y$ given $x$. There is no need for specific investigation of the functional beyond that needed to establish its robustness.

The probability model to be used is defined in Chapter 2. It incorporates i.i.d. sampling of $(X, Y)$ pairs and designed sequences of $X$'s. The notation is introduced along with the exposition of the model. Chapter 2 continues with examples of nonparametric regression techniques that can be made into estimates of the conditional distribution of $Y$. Some background material concerning statistical functionals, metrics on spaces of distributions, bivariate probability models and compact differentiability is given in Chapter 2. Some lemmas are presented in Chapter 2, for later use. Nonparametric regression techniques are predicated on an assumption that when $X$ is near $x$, the conditional mean of $Y$ is close to its value at $x$. Usually one can assume more: when $X$ is near $x$, the distribution of $Y$ is near to the distribution it takes at $x$. In Chapter 2 this idea is made precise by placing a metric on the distributions of $Y$ and assuming that the conditional distributions under this metric are continuous in $x$.

Chapter 3 studies consistency. Sufficient conditions for pointwise weak and strong consistency of the estimated conditional distribution of $Y$ are given. Consistency in three of the statistical metrics (Prohorov, Kolmogorov-Smirnov and Vasserstein) from Chapter 2 is obtained. Consistency of running functionals then follows for continuous functionals.

Chapter 4 studies asymptotic normality. First, asymptotic normality of the regression function is developed. This extends to the finite dimensional distributions of the conditional empirical process. A functional central limit theorem is then proved. Asymptotic

normality conditions for the regression may be translated into conditions for running versions of some functionals. The class of compact differentiable functionals is considered. Using von Mises' method the running functional is decomposed into the sum of a regression function and a remainder term. Sufficient conditions for the remainder term to be asymptotically negligible are provided.

The scope is limited as follows: pointwise (not uniform in $x$) results are obtained, problems of bandwidth selection are not considered, and rates of convergence are not computed. These represent three worthwhile directions for extension; perhaps a good starting point for each might be based on the way these extensions are made for regressions. Pointwise results (that hold a.e.) are stronger than global results but not as strong as uniform results. The pointwise approach handles designed as well as sampled predictors whereas global results usually assume i.i.d. sampling of predictor-response pairs. (These distinctions are discussed in Chapter 3.) Bandwidth selection might be tuned to some loss function on distributions or to a particular functional such as the mean. It should be reasonable to select a bandwidth for regression estimation and use it in the associated distribution estimator. Whether one might do better by a direct method is an interesting issue but depends on the loss function imposed on estimates of the distribution. In nonparametric regression the attainable rate of convergence depends on the number of continuous derivatives that the regression function admits. Similar results might be expected to hold for estimators of conditional distribution functions. The models considered here do not go beyond continuity (or Lipschitz continuity) of the $Y$ distributions as a function of $x$. With extensions to differentiability, it would be profitable to consider rates.

In developing theorems and notation, emphasis was placed on getting theorems that applied broadly, with conditions and conclusions that are easy to interpret. Theorems 3.2.2 and 3.3.1 are the most successful in this regard. While minimal assumptions are placed on the estimators of the conditional distribution, there is more structure than usual placed on the conditional distributions of $Y$ given $x$. In particular, some form of continuity is

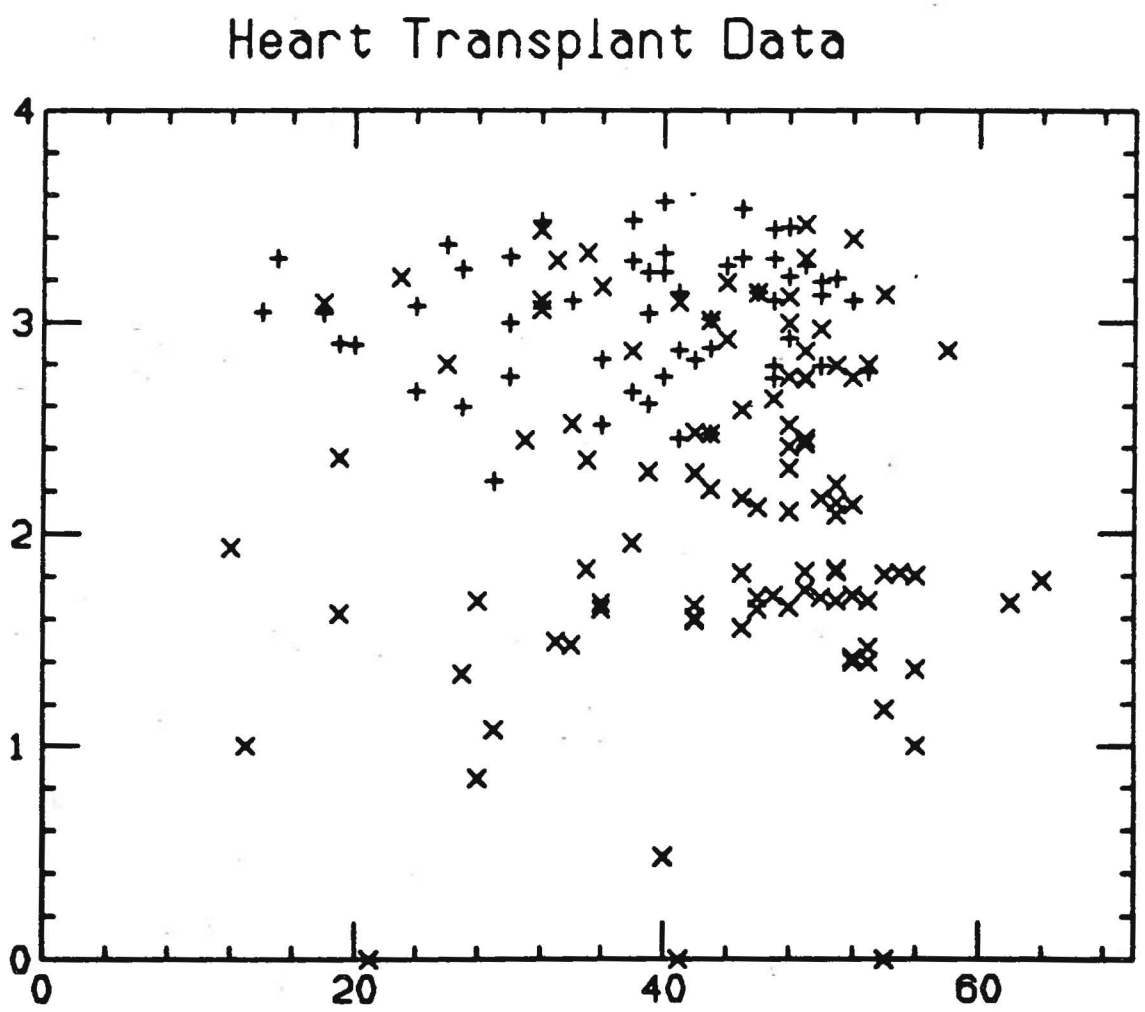
always assumed. The opposite approach is to place (almost) no conditions on the data and impose whatever conditions on the method yield optimal results. This is appropriate when one knows very little about the data because the statistician has complete control over the method. It is especially reasonable when there is a bone fide loss function to which the optimal asympotic result applies. However, when one is reasonably sure that some structure is present, and has reasons unrelated to asymptotics for choosing one estimator over another, then broadly applicable results that exploit some structure are of value. Also, broad conditions can expose similarities between apparently different methods.

The approach taken here—discrete estimation of the conditional distribution followed by the application of a functional is taken from Stone (1977), who uses it to obtain global $L^p$ consistency for nearest neighbor regressions, quantile estimates and conditional Bayes rules. In his discussion of Stone's paper, Brillinger (1977) suggests the application of likelihood functionals to the nonparametrically estimated conditional distributions. Brillinger also suggested extensions to conditional M-estimates which would have advantages of robustness. In his rejoinder Stone proves global weak consistency of the conditional estimate by exploiting its continuity with respect to the Prohorov metric.

Cleveland (1979) uses running versions of conditional M-estimators. Tibshirani (1984) considers local estimation of likelihood-based models. Härdle and Gasser (1984) establish consistency and asymptotic normality of some conditional M-estimators. Stute (1986) obtains a functional central limit theorem for a nearest neighbor estimator. Some other references to results in the literature are made in Chapters 2, 3 and 4.

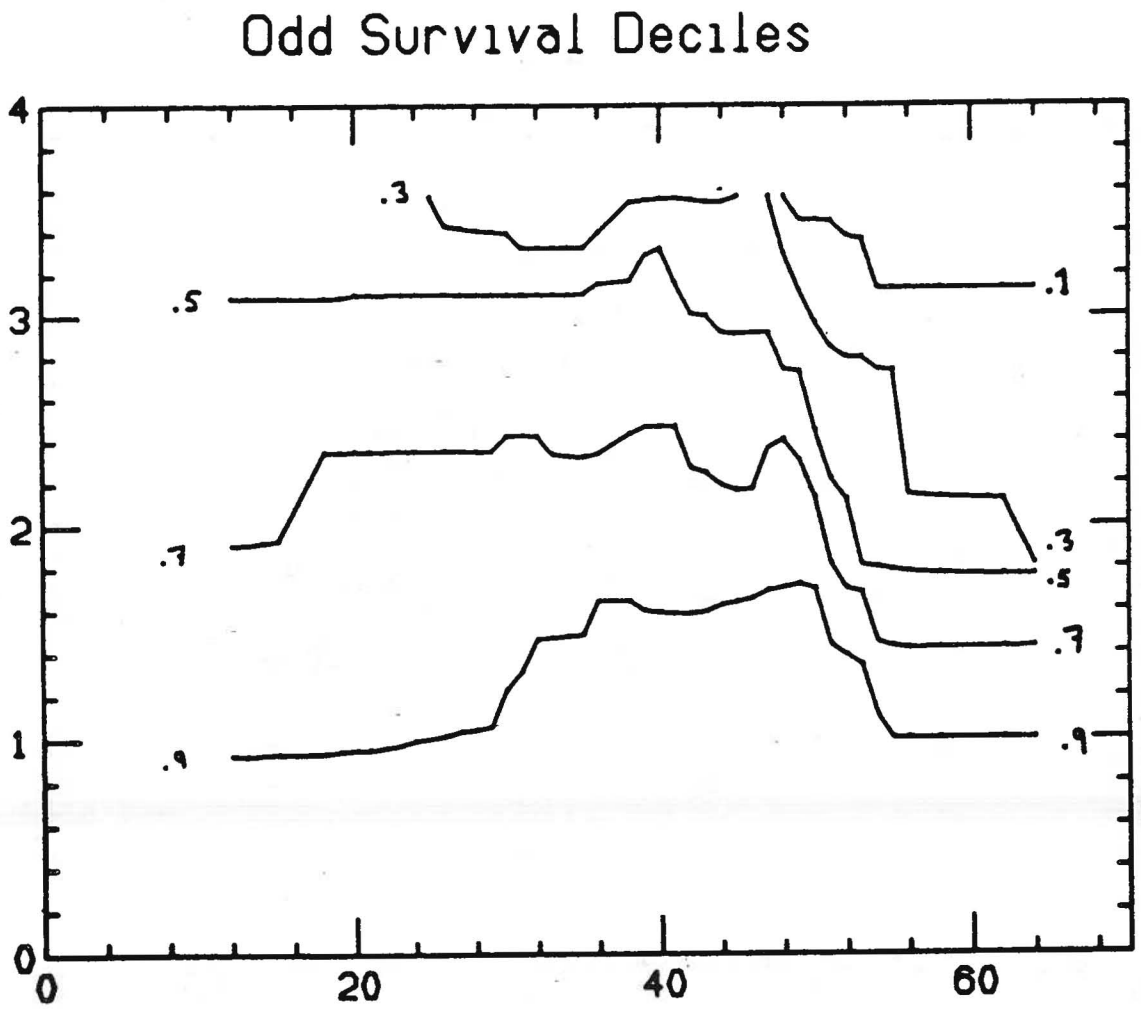
Conditional medians were considered for the heart transplant data by Doksum and Yandell (1983). Tibshirani (1984) computes local proportional hazards models for this data. Segal (1986) develops a rank-based version of the regression trees methodology of Brieman et. al. (1984) that can be applied to censored data. In particular he applies it to the heart transplant data and finds that the first split is made on the age variable at an age of 50.

**Figure 1** Stanford Heart Transplant Data.



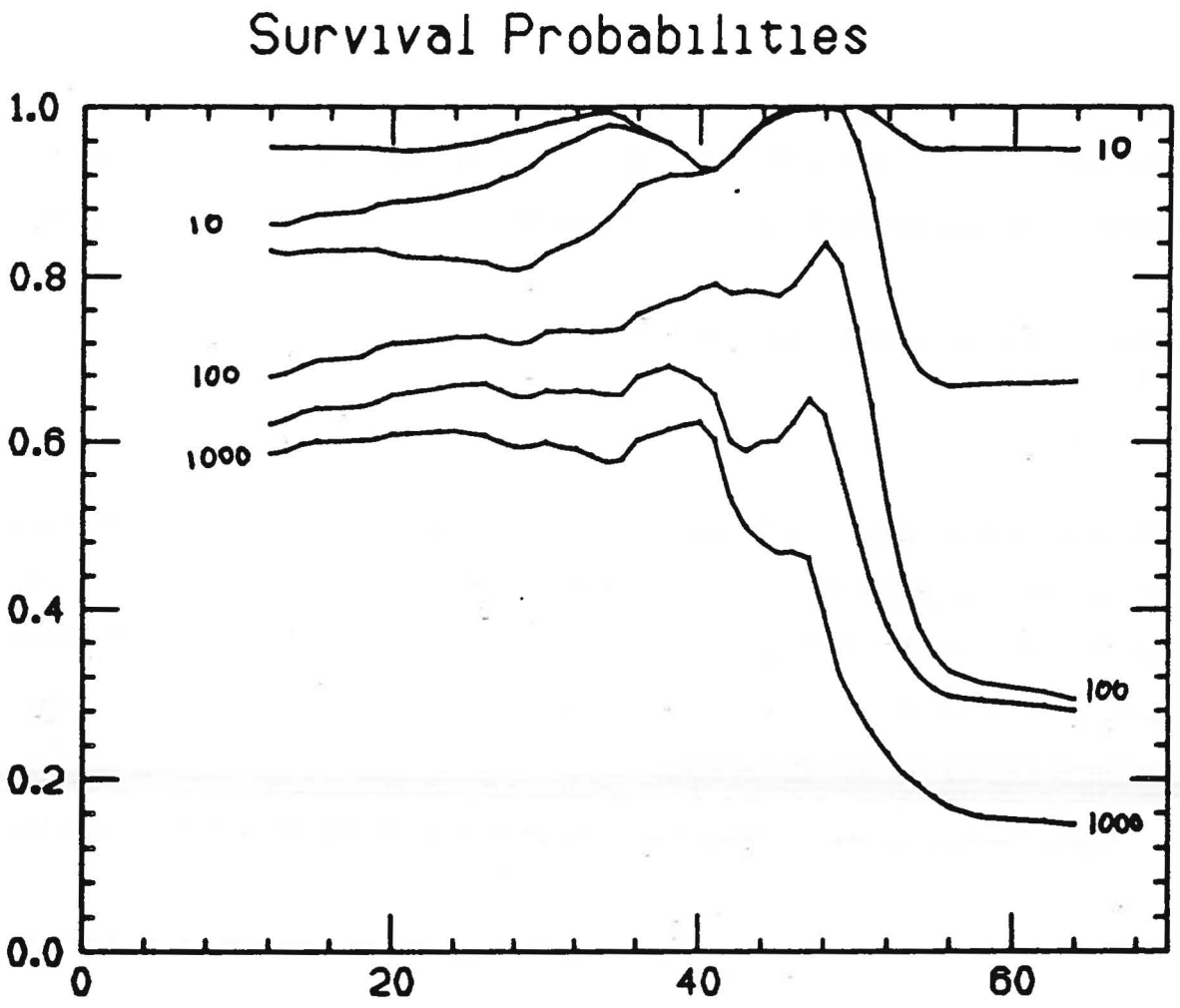

The horizontal axis is the age of a patient on the date of entry to the transplant program. The vertical axis is the logarithm (base 10) of the number of days the patient was observed to survive after the operation. There are 157 data points. Points marked "X" represent times of death and points marked "+" represent censoring times.

**Figure 2** Odd Survival Deciles.




The axes are as in Figure 1. The curve labelled ".5" is an estimate of the conditional median log survival time of a heart transplant patient, given the patient's age at entry. The other curves correspond to the estimated log times to which 10%, 30%, 70% and 90% of patients will survive given their age at entry. Portions of the 10% and 30% curves are censored. For example, the time at which only 10% of 25 year olds will remain is censored because there was more than 10% censoring in the data used to estimate the survival time given an age of 25 years at entry.

**Figure 3** Survival Probabilities




The horizontal axis gives the age at entry of a patient to the Stanford Heart Transplant program. The vertical axis gives the estimated conditional probability of survival past 10, 100, and 1000 days, given the age at entry. Also plotted are the probabilities of surviving some interpolated times, roughly 3, 32 and 316 days. (The estimated 3 day survival probability is 1 for older patients, so that curve disappears at the right of the figure.)

# 2 Preliminaries

This chapter introduces the notation used throughout, and provides some examples of estimators for conditional distributions. It also includes a discussion of statistical functionals, of metrics on distributions, of models for conditional distributions and of von Mises' method and compact differentiability of statistical functionals.

## 2.1 Notation

The data consist of pairs $(X_i, Y_i)$ where $i = 1, \ldots, n$. The $X_i$ take values in a set $\mathcal{X}$ and are thought of as predictors. The response variable $Y_i$ is a member of the set $\mathcal{Y}$. Unless otherwise specified $\mathcal{X} \subset \mathbb{R}$ and $\mathcal{Y} = \mathbb{R}$ and both are endowed with the usual Euclidean distance and topology. $X$ and $Y$ are used as typical data values that do not necessarily correspond to any specific observation.

Interest centers on the conditional behavior of $Y$ given $X$. To this end it is convenient to consider

$$F_x(y) = P(Y \leq y \,|\, X = x) \tag{1}$$

which for fixed $x \in \mathcal{X}$ is a distribution function on $\mathcal{Y}$ and for fixed $y \in \mathcal{Y}$ is a function on $\mathcal{X}$. Given that $X_i = x_i$, $Y_i$ has distribution function $F_{x_i}$. The random distribution $F_{X_i}$ is equal to $F_{x_i}$ when $X_i = x_i$.

$F_\bullet$ represents the mapping $x \to F_x$ from $\mathcal{X}$ to the space of distributions on $\mathcal{Y}$. Regularity conditions about the behavior of the distribution of $Y$ for varying $X$ will be expressed in terms of $F_\bullet$. This will generally mean that $F_\bullet$ is continuous or Lipschitz continuous when the distributions on $\mathcal{Y}$ are given an appropriate metric.

The probability model for the data is as follows: the $X$'s are drawn according to a design measure (that does not depend on the $Y$ values), and the $Y$'s are drawn from the corresponding conditional distributions and are conditionally independent given the $X$'s.

The design measure for the $X$'s could be a prescribed design sequence (design case) or it could be i.i.d. sampling from some distribution on $\mathcal{X}$ (sampling case) or it could be more complicated involving, say, a randomized choice among designs or dependent sampling that tends to fill in gaps left in $\mathcal{X}$ by the previous observations. The stipulation that the design measure not depend on the $Y$ rules out some sequential methods that might be of value.

A convenient construction to describe the conditional independence of the $Y_i$ given the $X_i$ is obtained as follows: introduce i.i.d. standard uniform random variables $U_i$, $i = 1, \ldots, n$ that are independent of the $X$'s, then put

$$Y_i = F_{X_i}^{-1}(U_i).$$

We define inverses of distribution functions as follows: $G^{-1}(u) = \inf\{y : G(y) \geq u\}$ and $G^{-1}\{u\} = \{y : G(y) = u\}$.

For some fixed point $x \in \mathcal{X}$ let

$$Y_i^x = F_x^{-1}(U_i). \tag{2}$$

Then $Y_i^x$, $i = 1, \ldots, n$ are i.i.d. random variables with distribution $F_x$. Intuitively, $Y_i^x$ is what $Y_i$ would have been if $X_i$ had been $x$. This construction will be handy in bias-variance-like decompositions.

The focus of interest will often be one or more functionals

$$T(\mathcal{L}(Y|X = x)) = T(F_x);$$

commonly considered functionals are the mean, mode, median, other quantiles, M-estimates, m.l.e.'s and variance estimates of the aforementioned. $T(F_x)$ can be thought of as a function on $\mathcal{X}$ as $x$ varies. The regression function arises for $T(\cdot) = m(\cdot)$, where

$$m(F_x) = \int y dF_x(y) \tag{3}$$

is the mean. It should cause no confusion to use $m(x)$ for $m(F_x)$.

To analyze $T(\cdot)$, consider it as a mapping. Its domain $\mathcal{D}_T$ must naturally contain $F_x$ for all $x \in \mathcal{X}$. It will also have to contain estimates of $F_x$ obtained from the data. These will be distributions with support in a finite set. Unless otherwise stated the range of $T$ is $\mathbb{R}$. The domain $\mathcal{D}_T$ also comes equipped with a topology. Most of the topologies considered are metrizable. The basic open sets in a metrized topological space can be taken to be the open balls

$$B_\epsilon(F) = \{G \in \mathcal{D}_T \mid \rho(F, G) < \epsilon\}$$

where $\epsilon > 0$ and $F \in \mathcal{D}_T$, and $\rho(\cdot, \cdot)$ is a metric on $\mathcal{D}_T$. The one non-metrizable topology considered is the topology of weak convergence for finite signed measures used in Sec. 3.2.

The emphasis will be on the Kolmogorov-Smirnov metric, the Prohorov metric, and the Vasserstein metrics. See Sec. 2.4 for a discussion of statistical metrics.

The running functional $T(F_x)$ is estimated by $T(\hat{F}_x)$ where $\hat{F}_x$ is an estimate of $F_x$ based on the data. $\hat{F}_x$ will depend on $n$ and $(X_i, Y_i), i = 1, \ldots, n$ although this dependence is suppressed for notational convenience. $\hat{F}_x$ is not in general a statistical functional by virtue of its dependence upon $n$, but may be thought of as a sequence of such functionals. $\hat{F}_x$ need not be a probability measure on $\mathcal{Y}$ in which case it may be necessary to extend the definition of $T(\cdot)$.

Following Stone (1977) consider estimators $\hat{F}_x$ of the form

$$\hat{F}_x(y) = \sum_{i=1}^{n} W_{ni}(x)\delta_{Y_i}$$

where $\delta_{Y_i} = 1_{Y_i \leq y}$ is a point-mass at $Y_i$ and $W_{ni}(x)$ is a weight attached to the $i$'th observation out of the first $n$ observations. $W_{ni}(x)$ depends on $X_1, \ldots, X_n$ and on $n$ but not on the $Y$ values. To keep the notation uncluttered, denote the weight on the $i$'th observation by $W_i$ with $n$ and the target point $x$ understood. That is

$$\hat{F}_x = \sum W_i 1_{Y_i \leq y} \tag{4}$$

in terse notation. It is natural to denote $m(\hat{F}_x)$ by $\hat{m}(x)$.

The weights form a discrete signed measure on $\mathcal{X}$ with atoms of size $W_i$ at $X_i$. This measure is denoted $W_x$, so that

$$W_x(A) = \sum W_i 1_{X_i \in A}. \tag{5}$$

Many conditions on the weights can be expressed in terms of $W_x$. For large $n$, $W_x$ should be close to $\delta_x$, the point-mass at $x$. That notion is made precise by putting a metric $\rho$ on the distributions on $\mathcal{X}$ and requiring $\rho(W_x, \delta_x) \to 0$ in some mode of stochastic convergence.

For the regression function $m(\hat{F}_x) = \sum W_i Y_i$ and (4) incorporates many of the commonly used nonparametric regression techniques including smoothing splines, kernel estimators, nearest neighbour estimators, and running linear regressions. Sec. 2.2 discusses the choice of the $W_i$ in more detail. These weights are distinguished from adaptive weights which depend on the $Y$'s. For example, if the smoothing parameter in spline regression or running linear regression is determined by cross-validation the resulting regression estimate is adaptive and hence not covered by (4).

For a given set of weights put

$$n_x = [\sum W_i^2]^{-1}. \tag{6}$$

If each $F_{x_i}$ has variance $\sigma^2$, then conditionally on the observed $X$'s, $\sum W_i Y_i$ has variance $\sigma^2/n_x$. In this sense $n_x$ is an effective sample size at $x$. The $X_i$ that contribute to $\hat{F}_x$ are thought of as a sample of size $n_x$ from $W_x$ and the locally reweighted $Y_i$ are thought of as a biased sample of size $n_x$ from $F_x$. In asymptotic considerations, it will be necessary for $n_x \to \infty$ to control the variance. Typically $n_x/n \to 0$ as $n \to \infty$ and this allows $W_x$ to converge to $\delta_x$ to control the bias. For a fixed sample, $n_x$ regarded as a function of $x$ can be used to compare precision over the range of $\mathcal{X}$. It can also be used in heuristic degree of freedom calculations for pointwise t-tests and intervals.

Most consistency proofs for nonparametric regressions start with a decomposition

$$\hat{m}(x) - m(x) = \sum W_i(Y_i - m(X_i)) + \sum W_i(m(X_i) - m(x)) - m(x)(1 - \sum W_i).$$

Conditionally on the $X$'s, the first term is a sum of mean zero random vairables, and differs from zero because of sampling variability in the $Y_i$ and the second term is conditionally constant and differs from zero because the $X_i$'s are not exactly at $x$. It is natural to call the first term a variance term and the second term a bias term, though strictly speaking these labels refer to the second moment of the first term and the first moment of the second term respectively. The decomposition considered here is of the form

$$\hat{m}(x) - m(x) = \sum W_i(Y_i^x - m(x)) + \sum W_i(Y_i - Y_i^x) - m(x)(1 - \sum W_i). \qquad (7)$$

In this decomposition the first and second terms will still be refered to as variance and bias terms, but the variance term in (7) is conditionally a weighted sum of i.i.d. mean zero terms and moreover, the terms $Y_i^x - m(x)$ are independent of the $X_i$'s and hence also of the $W_i$'s. This makes the variance term easier to handle, at the expense of some complication in the bias term. However the bias term in (7) is tractable, and may be conveniently analyzed via Vasserstein metrics. A similar decomposition of $\hat{F}_x$ will be used in Chapter 3.

## 2.2 Examples of Weights

This section presents some examples of weights that fit into the framework of the Sec. 2.1. Most of the weight schemes discussed here were developed for estimating regression functions. Similar ideas have been used in density estimation and in the estimation of spectral densities. The discussion covers in turn the following methods: kernels, nearest neighbors, symmetric nearest neighbors, local linear regressions, and smoothing splines. A final subsection discusses some other related ideas. For a comprehensive bibliography of nonparametric regression techniques see Collomb (1985).

### 2.2.1 Kernel Smoothers

Kernel estimates of the regression function were introduced by Nadaraya (1964) and

Watson (1964). For the kernel estimate:

$$W_i = \frac{K\left(\frac{x-X_i}{b_n}\right)}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{b_n}\right)} \tag{1}$$

where $K(v)$ is a function called the kernel and $b_n > 0$ is a constant called the bandwidth. We assume that

$$\int |K(v)|dv < \infty$$

and

$$\int K(v)dv = 1.$$

The latter is a convenient normalization—multiplying $K$ by a (nonzero) constant would not change $W_i$ and might make computation easier. Consistency of the kernel regression estimate generally requires that $b_n \to 0$ at an appropriate rate.

Kernel regression estimators were preceded by kernel density estimators. Nadaraya (1964) cites Parzen (1962) and Watson (1964) cites Rosenblatt (1956). Kernel methods were previously used in spectral density estimation. This connection is discussed in Subsec. 2.2.3.

We give some examples of kernel functions for $\mathcal{X} \subset \mathbb{R}$ taken from Benedetti (1977). There are obvious extensions to $\mathbb{R}^d$.

**Examples:**

| | | |
|---|---|---|
| 1 | Uniform | $K(v) = \frac{1}{2}1_{\|v\|\leq 1}$ |
| 2 | Triangular | $K(v) = (1 - \|v\|)^+$ |
| 3 | Quadratic | $K(v) = \frac{3}{4}(1 - \|v\|^2)^+$ |
| 4 | Exponential | $K(v) = \frac{1}{2}e^{-\|v\|}$ |
| 5 | Gaussian | $K(v) = \frac{1}{\sqrt{2\pi}}e^{-v^2/2}$ |
| 6 | Cauchy | $K(v) = \frac{1}{\pi}\frac{1}{1+v^2}$ |
| 7 | Fejer | $K(v) = \frac{1}{2\pi}\frac{\sin(v/2)}{v/2}$ |

The quadratic kernel is often referred to as the Epanechnikov kernel. Epanechnikov (1969) argues that it is the optimal shape for estimating densities in any dimension so

long as the true density is sufficiently smooth. (It is optimal within the class of bounded symmetric probabilty densities for which all moments are finite. An integrated squared error criterion is used.)

Kernels with negative sidelobes (for instance the Fejer kernel) are used to reduce bias. See Watson (1964) for an example.

### 2.2.2 Nearest Neighbor Smoothers

Nearest neighbor methods originated with Fix and Hodges (1951) in the context of nonparametric discrimination. They were first used in density estimation by Loftsgaarden and Quesenberry (1965). The first general discussion in the regression context seems to be Royall (1966), though Watson (1964) mentions uniform nearest neighbors.

Let $c_{in}$, $1 \leq i \leq n < \infty$ be a triangular array of real numbers. If there are no ties among the first $n$ $X$'s then the nearest neighbor weights are

$$W_i = W_{in} = c_{r(i)n} \tag{2}$$

where $r(i)$ is the rank of $\|X_i - x\|$ among the first $n$ observations. If there are ties in the $X$'s break them arbitrarily, for example by using the index $i$, and assign weights from (2). Then for each set $I$ of indices corresponding to tied $X$'s let

$$W_I = \frac{1}{|I|}\sum_{i \in I} W_i \tag{3}$$

and set $W_i = W_I, \ \forall i \in I$.

Nearest neighbor weights are called $k$ nearest neighbor weights (k-NN) when for some $k = k(n) = o(n)$ $i > k$ implies $c_{in} = 0$. The following are examples of k-NN weights.

**Examples:**

| | | |
|---|---|---|
| 1 | Uniform | $c_{in} = \frac{1}{k} 1_{\|i\| \leq k}$ |
| 2 | Triangular | $c_{in} = 2(k - i + 1)^+ / (k(k-1))$ |
| 3 | Quadratic | $c_{in} = 6\left((k - i + 1)^+\right)^2 / (k(k-1)(2k-1))$ |

Nearest neighbor weights analogous to kernel functions with unbounded support may also be of interest.

### 2.2.3 Symmetric and One-Sided Nearest Neighbors

When $\mathcal{X} = \mathbb{R}$, a family of symmetric nearest neighbor methods are available that generalize the familiar running average. At first assume there are no ties in the $X$'s, and only consider target points that correspond to observations: $x = x_j$ for $j \leq n$. Also assume without loss of generality that the first $n$ observations are ordered $x_1 < x_2 < \ldots < x_n$ to avoid complicating the subscripts.

Let $c_{in}$, $0 \leq n < \infty$ be a triangular array of weights. Then a symmetric nearest neighbor scheme at the target point $x_j$ has

$$W_i \propto c_{|i-j|n}. \tag{4}$$

The constant of proportionality in (4) is usually chosen so that the weights sum to 1. Uniform, triangular, quadratic, etc. versions of symmetric nearest neighbor weights are easily defined.

If $x$ is not an observation, but $x_j < x < x_{j+1}$, some convenient convention can be used for the weights. Natural examples are $W_x = W_{x_j}$ and $W_x = \lambda W_{x_j} + (1-\lambda)W_{x_{j+1}}$ where $\lambda = (x_{j+1} - x)/(x_{j+1} - x_j)$. In practice it is likely to be even more convenient to simply compute $T(\hat{F}_{x_i})$ for $i = 1, \ldots, n$ and linearly interpolate values of $T$.

Ties can be broken as outlined above for nearest neighbor weights, although ties at the target point are more troublesome. The following prescription for tie breaking generalizes the one for nearest neighbors while preserving some symmettry between the right and left sides. If there are an odd number $2j+1$ of observations at $x$, then an arbitrary choice can be made to assign them weights proportional to

$$\{c_{jn}, \ldots, c_{1n}, c_{0n}, c_{1n}, \ldots, c_{jn}\}$$

and to assign weight proportional to $c_{in}$ for $i \geq j$ to the $i$'th observations on each side of the target. If there are an even number $2j$ of observations tied at $x$ then $2j-1$ of them can be assigned as above and one of them can get weight proportional to $c_{jn}$. The $i$'th point

to each side of the tied points then gets weight proportional to $(c_{(j+i-1)n} + c_{(j+i)n})/2$. After such an assignment, the weights are equalized over sets of tied $x_i$'s as before.

A technique of Yang (1981) can be used to express the most commonly considered symmetric nearest neighbor weights in terms of a symmetric kernel function $K(\cdot)$ and $F_n$, the empirical distribution function of $X$:

$$W_i \propto K\left(\frac{F_n(X_i) - F_n(x)}{b_n}\right). \tag{5}$$

The function $F_n$ is also defined when the $X_i$ are obtained from a design. The constant of proportionality is chosen to make the weights sum to 1. (The exact form of (5) is from Stute (1984).) For $x \in (x_j, x_{j+1})$ this formula implicitly sets $W_x = W_{x_j}$.

One advantange of symmetric nearest neighbor weights over kernel weights is that the set of values $W_i$ is fixed in the former and random in the latter. The kernel method must be modified to handle the case where the kernel function is zero for all the observations, but this never happens with symmetric nearest neighbors. Such an event can have positive probability for kernels with bounded support. The probability is generally small enough to ignore in practice, but may pose difficulty in theoretical calculations. An advantage of symmetric nearest neighbor weights over nearest neighbor weights is that they are balanced with respect to the target point—except at the ends there is the same amount of weight on the left as on the right of the target. With nearest neighbors the amount of weight on a given side of $x$ is random and could be zero, even when $x$ is not at the ends of the data. An advantage of symmetric nearest neighbor weights over both kernel weights and nearest neighbor weights is that computation can be much faster. In the case of the uniform weights, the weight function at $x_{j+1}$ differs from that at $x_j$ in at most two weights $W_{j+1+k}$ and $W_{j-k}$. The regression function can be computed quickly by updating a sum of $Y$'s counter and a number of points counter. To compute the regression at all data points requires only $2n$ additions, $2n$ subtractions and $n$ divisions. Triangular, quadratic and higher order symmetric nearest neighbor regressions can be obtained by repeated application of uniform symmetric nearest neighbors.

Now consider (5) with an asymmetric function $K(\cdot)$. An extreme departure from symmetry involves taking in (5) kernels

$$R(v) = 2K(v)1_{v\geq 0}$$

and

$$L(v) = 2K(v)1_{v\leq 0}$$

which define right and left sided nearest neighbor weights. If $F_\bullet$ is piecewise continuous with a discontinuity near $x$, then the one-sided weights from the side opposite the discontinuity may provide a better estimate than a symmetric weights. A comparison of left and right sided estimates of $F_x$ or $T(F_x)$ might provide a means of detecting discontinuities. One sided neighborhoods are used to estimate regressions in McDonald and Owen (1986). Note that left sided weights are not available for the leftmost observation and are based on few points for observations near the left end (and the same comments apply to right sided weights at the right of the data).

The symmetric versions of uniform, triangular, and quadratic nearest neighbor weights are related to the truncated, the modified Bartlett and one of the Parzen estimators of spectral density respectively (Anderson 1971, Chapter 9). The relationship is as follows: the estimate of the spectral density at frequency $\omega$ is a weighted sum of $c_r \cos(\omega r)$ for $|r| \leq k$, where $c_r$ is the sample autocovariance at lag $r$ and the weights are proportional to $1_{r\leq k}$, $(1-|r|/k)1_{r\leq k}$ and $(1-(|r|/k)^2)1_{r\leq k}$ respectively. Anderson also discusses several other spectral density estimators, which could also be turned into k-NN weight functions.

In forecasting, one-sided exponential nearest neighbor weights are used in what is called exponential smoothing (Chatfield 1980). In that application a time series is observed at equally spaced points (so ranks correspond naturally to actual time elapsed) and the weights are placed on the present and past to forecast the future. These weights have the advantage of providing easily updatable regression functions. In the one-sided case, after some startup, the regression estimate at $x_i$ is almost exactly $m(x_i) = \rho Y_i + (1-\rho)m(x_{i-1})$.

In the two-sided case the regression estimate is obtained by taking a weighted average of the left and right sided exponential smooths.

### 2.2.4 Local Linear Regression Weights

An important class of weighting schemes are the linear regression weights. When $\mathcal{Y}$ is one dimensional the regression function at $x$ may be estimated by a linear regression on the points in the neighbor set of $x$. The estimate of the regression is a linear combination of the $Y$ values in the neighborhood, and the weights of the linear combination may be thought of as generating an estimate $\hat{F}_x$ of $F_x$. When $W_i$ are probability weights and $\mathcal{X}$ is $\mathbb{R}$, the weights obtained from a $W_i$-weighted regression of $Y$ on $X$ are

$$\widetilde{W}_i = W_i \left(1 + \frac{(x - \bar{x})}{s} \frac{(X_i - \bar{x})}{s}\right) \tag{6}$$

where $\bar{x} = \sum W_i X_i$ and $s^2 = \sum W_i (X_i - \bar{x})^2$. (If $s = 0$ take $\widetilde{W}_i = W_i$.) When the $W_i$ are uniform ($1/k$ for $k$ points, 0 for the others) the $\widetilde{W}_i$ resemble a kernel with a trapezoidal shape, the height and slope of which depend on $\bar{x}$, $s$ and $k$. The $\widetilde{W}_i$ sum to 1 but can include some negative weights when $x$ is not near the mean of $W_x$ as must happen for $x$ near the ends of the data. For other shapes the linear regression "kernel" is the product of the original weight function and a trapezoidal function that depends on the $X$'s and the original weights through $\bar{x}$ and $s$.

The motivation for linear regression weights is that they preserve linear structure in the data. This is especially valuable at the ends of the observed sample where simple weighted averages flatten out any trend. Friedman and Stuetzle (1983) use regressions over symmetric uniform nearest neighbors for several different $k$ to arrive at an estimate of the regression. See also Friedman (1984). McDonald and Owen (1986) use uniform nearest neighbor linear regressions from several different $k$ values for left, right and symmetric windows. Linear regression weights with uniform symmetric nearest neighbors are updatable and hence can be computed in $O(n)$ operations assuming the data are sorted on $x$.

Stone (1977) states that the local linear weights are not necessarily consistent and shows how to "trim" them to achieve consistency. The trimming tends to remove their utility at the ends of the data and in the middle of the data there is usually not much difference between linear regression weights $\widetilde{W}_i$ and the weights $W_i$ on which they are based (at least in the symmetric uniform case). Marhoul and Owen (1985) study some of the asymptotics of regression estimates based on linear regression weights on symmetric and one-sided neighborhoods. The balance implicit in symmetric nearest neighbor sets is exploited in their proof of the mean square consistency of running linear fits over such sets; the proof would not go through for linear fits over ordinary nearest neighbor sets. The mean square consistency holds for one-sided windows that contain $k-1$ points from one side of the target and 1 point that is either at the target or on the other side.

Stone (1977) gives the generalization of (6) for linear regression weights when $\mathcal{X} = \mathbb{R}^d$.

Linear regressions from symmetric and one-sided uniform nearest neighbor weights are updatable and linear regression versions of exponential smoothing are also updatable.

### 2.2.5 Smoothing Splines

When $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, the smoothing spline estimator of the regression of $Y$ on $X$ is that function $g(\cdot)$ that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(x))^2 + \lambda \int_{\mathcal{X}} g''(x)^2 dx \tag{7}$$

where $\lambda > 0$ is given. The solution $g(x)$ is a cubic spline with knots at the observations by a variational argument of Reinsch (1967) and moreover can be written as a linear combination (Wahba 1975) of $Y$'s

$$g(x) = \sum_{i=1}^{n} G(x,i)Y_i \tag{8}$$

where for each $i$, $G$ provides a function on $\mathcal{X}$ and for each $x$, $G$ provides a vector of weights. The smoothing spline fits into the framework of equation 2.1.4 by setting $W_i = G(x,i)$. In principle this gives spline estimates of $F_x$, although the $G(x,i)$ are difficult to compute.

Silverman (1984) developes an asymptotic approximation to $G$ in terms of a variable kernel:

$$G(x,i) \doteq \frac{1}{n}\frac{1}{f(x_i)}\frac{1}{h(x_i)}\kappa\left(\frac{x-x_i}{h(x_i)}\right) \tag{9}$$

where

$$h(x_i) = \lambda^{1/4} f(x_i)^{-1/4}$$

and

$$\kappa(v) = \tfrac{1}{2}\exp\left(-|v|/\sqrt{2}\right)\sin\left(|v|/\sqrt{2}+\pi/4\right)$$

and $f$ is the (well-behaved) density of $X$.

For a summary of spline smoothing see Silverman (1985) and Wegman and Wright (1983).

### 2.2.6 Other Weights

A variation on kernel weights due to Priestley and Chao (1972) uses weights proportional to

$$\frac{x_i - x_{i-1}}{b_n}K\left(\frac{x-x_i}{b_n}\right) \tag{10}$$

where the observations are arranged so that the sequence $(x_i)$ is nondecreasing. The Priestley-Chao weights modify the Nadaraya-Watson weights so that closely spaced points get relatively less weight and more widely separated points get relatively more weight. Gasser and Muller (1977) show that the weights in (10) have a smaller asymptotic mean square error than do ordinary kernel weights in the case of equidistant and nearly equidistant designs.

The kriging technique, popular in geostatistics, estimates a regression function (usually over two or three dimensions) by a weighted combination of observations, the weights depending on proximity to the target point and upon an assumed covariance structure for the observations. Therefore at least superficially it can be expressed via equation 2.1.4 and the weights used to estimate conditional distribution functions. For a discussion of

kriging see Ripley (1981) or Yakowitz and Szidarovszky (1985) who compare it to kernel nonparametric regression. Watson (1984) shows that spline regression estimation can be obtained as a special case of kriging.

The regression trees of Brieman et. al. (1984) could be used to estimate $F_x$ by putting equal weight on all the observations in each node. That estimate would be used for all the predictor values that lead to the node. Since the splits made by the recursive partitioning algorithm depend in a complicated way on the $Y$ values, so do the resulting weights. For this reason they do not fit into the framework considered here.

Another smoothing technique that does not fit into the present context is the iterative application of running medians in Tukey (1977). A single running median may be interpreted as the conditional median function when uniform symmetric k-NN weights are used, but iterative application of such an algorithm would be quite unnatural if not impossible to interpret as a functional applied to an estimate of the conditional distribution.

Wandering schematic plots (Tukey, 1977) are in the spirit of this work, however. They are formed by partitioning the $X$–axis into bins and computing sample statistics for the $Y$ values that appear in each bin. The resulting values are plotted above the bin medians.

### 2.2.7 Bandwidth Selection

In all of the above weighting schemes there is a parameter $k$ or $b_n$ or $\lambda$ that governs the rate at which the weight drops off as the distance from $X_i$ to $x$ increases. In each case larger values of the parameter result in more spread out weights and the corresponding regression estimates are smoother looking. We use the term bandwidth to refer to any of these quantities. Smaller bandwidths give rise to regression curves that pass closer to the data. In general a regression estimated with a small bandwidth is subject to less bias and more variance than when a large bandwidth is used. The bandwidth to be used can be selected by plotting the results for a few choices and selecting the one that seems best.

For reasons expressed well in Silverman (1985 Sec. 4) it is desirable to have available

an automatic technique for bandwidth selection. The cross-validation method of Stone (1974) is commonly used for this. The idea is to choose the bandwidth that minimizes cross-validated squared error. See Friedman and Stuetzle (1983) who use crossvalidation to select $k$ for a linear regression over a uniform symmetric k-NN neighborhood, Hall (1984) who studies asymptotics for the cross-validated kernel regression, and Wahba and Wold (1975) for cross-validation in smoothing splines. Craven and Wahba (1979) provide a faster alternative to cross-validation, known as generalized cross-validation. Friedman and Stuetzle perform a local cross-validation so that the bias-variance tradeoff implicit in a choice of $k$ can be made for each $x$.

Titterington (1985) surveys smoothing techniques in statistics including image processing and mentions some alternatives to cross-validation including minimum risk and Bayesian methods. In minimum risk strategies, the mimimizing bandwidth for a risk function is obtained or approximated by a closed form expression. Typically such an expression would involve the underlying curve and an approximation to that curve would be substituted.

Bandwidth selection techniques do not usually fit into equation 2.1.4 since the $Y$ values are used to select the bandwidth. When the dependence is very simple however as in the case of a choice of bandwidth from a finite set of consistent bandwidths the results of Chapters 3 and 4 are easy to apply.

If a bandwidth choice is made and used to obtain $W_x$ and then all functionals of interest are computed with the estimate $\hat{F}_x$ then many natural relationships between different functionals will hold for the estimates. For example

$$\hat{\mathcal{E}}\left(g(Y)+h(Y)\,|\,X=x\right)=\hat{\mathcal{E}}\left(g(Y)\,|\,X=x\right)+\hat{\mathcal{E}}\left(h(Y)\,|\,X=x\right)$$

and

$$Var(\hat{F}_x)=\int (y-m(\hat{F}_x))^2\hat{F}_x(dy)$$

and

$$Var(\hat{F}_x)=\hat{\mathcal{E}}\left(Y^2\,|\,X=x\right)-\hat{\mathcal{E}}\left(Y\,|\,X=x\right)^2$$

will hold. For probability weights $W_x$ the estimated quantiles are properly ordered (in particular quantile regressions will not "cross")

$$\hat{F}_x \left\{ |Y - m(\hat{F}_x)| > k\sqrt{Var(\hat{F}_x)} \right\} \leq 1/k^2,$$

so that a pointwise Chebychev's inequality will hold and so on. Such self consistency properties of the estimates are desirable though they may entail some cost: the best bandwidths, in squared error terms say, may differ from functional to functional. For example one might do better with larger bandwidths for variances and extreme quantiles than for means and moderate quantiles respectively. In practice it should often be reasonable to pick the bandwidth to estimate a particular functional such as the mean and then use those weights for all other functionals.

## 2.3 Statistical Functionals

A statistical functional is a mapping defined on a space of distribution functions. Usually the image space is $\mathbb{R}$ but it could also be a set of categories or a higher dimensional Euclidean space. The domain usually includes all empirical distribution functions and the hypothetical true distribution. Statistical functionals are a convenient abstraction; they apply in most statistical situations and allow the use of concepts and techniques from analysis.

Many quantities of interest to statisticians can be expressed as statistical functionals $T(F)$ where $F$ is the distribution of the data. The natural estimate of $T(F)$ is often $T(F_n)$ where $F_n$ is the sample distribution function. For example, the sample mean is $m(F_n)$.

Most calculations that statisticians perform on a set of data can be expressed as statistical functionals on $F_n$. Any function of $n$ i.i.d. observations is a function of a list of the observed values (sorted for example) and a permutation that labels them. Most statistical computations make no use of the labelling of the observations (except perhaps to check independence or identity of distribution) and hence depend only on the list of observations. The list of observations is determined by $F_n$ and $n$. The sample size $n$

cannot be determined from $F_n$. Statistical computations tend to depend more on $F_n$ than on $n$. Many statistics do not depend on $n$ at all. For example the variance is

$$V(F) = \int (y - m(F))^2 \, dF(y),$$

the median is

$$Q_{.5}(F) = \inf\{q : \int_{-\infty}^{q} dF(y) \geq .5\}$$

and an M-estimate $M(F)$ may be obtained as a solution $M$ of

$$0 = \int \psi(y - M) \, dF(y).$$

The most commonly cited statistic that depends on $n$ is the unbiased sample variance:

$$\begin{aligned} s^2 &= \tfrac{1}{n-1} \sum (Y_i - \overline{Y})^2 \\ &= \tfrac{n-1}{n} V(F). \end{aligned}$$

In this and similar cases an auxiliary parameter may be introduced for the sample size. The functional is then defined on $U \times [0, \infty]$ where $U$ is a space of distributions. The sample value is $T(F_n, n)$ and the population value is $T(F, \infty)$. The analytic properties of such sequences of functionals can be considered on this augmented space. For more on auxiliary parameters see Reeds (1976, Sec. 1.6). In particular Reeds considers bivariate Taylor series expansions of functionals whose first argument is a distribution and whose second argument is an auxiliary parameter.

Many important properties of statistics may be expressed in terms of analytic properties of statistical functionals. A statistical functional $T(F_n)$ is robust at $F$ according to Hampel (1971) if $\mathcal{L}(T(F_n))$ as a function of the distribution $F$ of a single observation is a continuous function at $F$ when the Prohorov metric is used in the spaces for both $F$ and $\mathcal{L}(T)$. The augmented statistical functional $T(F_n, n)$ is robust if the continuity is uniform in the auxiliary parameter. Hampel shows that if $T(\cdot)$ itself is continuous at $F$ then it is robust at $F$. His definition of continuity of an augmented functional is essentially that of bivariate continuity at $(F, \infty)$ although to avoid assuming the existence of $T(F, \infty)$ he uses

a version of the Cauchy criterion. It is important to note that robustness, like continuity, depends on both the functional and the point in the domain under consideration. The mean is not continuous at any $F$. The median is not continuous at $F$ if $F^{-1}\{1/2\}$ is an interval of positive length, and hence is not robust at that $F$ either.

The influence curve is a form of derivative of a functional. The use of Taylor expansions of statistical functionals to prove asymptotic normality is known as Von Mises' method. See Sec. 2.6 for a discussion.

If one can obtain results based only on analytic properties of the functionals used then they may apply easily to as yet unknown statistical methods. For example, in Chapter 3 some consistency results for running functionals require only Prohorov continuity of the functional. They therefore apply to any robust functional.

Another advantage of functionals is that there is often a natural extension to spaces that contain more that just distribution functions. The space of all finite signed measures is such an extension as are $C[0,1]$, $D[0,1]$ and $L^p[0,1]$. Such spaces are vector spaces and hence are easier to work with, in the same way that it is easier to work in Euclidean space than in a simplex. The functionals for the mean, median, variance and the M-estimators can be extended meaningfully to larger spaces. Estimators of $F_x$ that put a small amount of negative weight on some observations, perhaps to reduce bias, can be handled naturally by extending the domain of the functionals.

## 2.4 Statistical Metrics

This section presents some of the more useful statistical metrics and discusses their properties. A familiarity with metrics, norms, the topologies they induce and the associated definitions of continuity and convergence is assumed. These concepts are readily found in introductory books on topology, such as Willard (1970).

Throughout this section, $U$ is a space of distributions. They are defined as probability measures on a measure space $(\Omega, \mathcal{M})$, with the important special case $(I\!R, \mathcal{B})$, where $\mathcal{B}$ is

the Borel $\sigma$-field. Sometimes it is convenient to extend $U$ to include finite signed measures or to restrict to measures satisfying a moment condition. $F, G, H, F_n$ and $G_n$ are elements of $U$. $F$ will be a bone fide probability and $F_n$ will denote the empirical probability from a sample of size $n$ from $F$. $G$ and $H$ are general members of $U$ and $G_n$ is a sequence in $U$. On $\mathbb{R}$ the letter used to denote the measure will also be used for the distribution function so that for example $F(x) = F((-\infty, x])$.

If a statistical functional $T$ is continuous at $G$ when a metric $\rho$ is used on $U$ and if $\rho(G_n, G) \to 0$ then $T(G_n) \to T(G)$. The same is true if both $\to$'s are replaced by almost sure convergence or by weak convergence. (This is proved in Lemma 3.1.1. It is not true for $L^p$ convergence.) Therefore consistency of $G_n$ for $G$ implies consistency for a potentially large class of statistical functionals.

Recall that a metric $\rho_1$ on $U$ is stronger than $\rho_2$ (also on $U$) if every open $\rho_2$–ball around a point in $U$ contains an open $\rho_1$–ball around the same point. A sequence that converges in the $\rho_1$ metric converges in the $\rho_2$ metric. Any continuous function on the metric space $(U, \rho_2)$ is continuous on $(U, \rho_1)$. Any continuous function with range $(U, \rho_1)$ is continuous with range $(U, \rho_2)$.

### 2.4.1 Prohorov Metric

Let $\Omega$ be a complete separable metric space with Borel sigma field $\mathcal{M}$ and metric $d$. The most important case is $\Omega = \mathcal{Y} = \mathbb{R}$, $\mathcal{M} = \mathcal{B}$ and $d(x, y) = |x - y|$. For $\epsilon > 0$ and $A \subset \mathcal{Y}$ define

$$A^\epsilon = \{y \in \Omega : d(y, A) < \epsilon\} \tag{1}$$

where $d(y, A) = \inf_{x \in A} d(y, z)$. Let $G$ and $H$ be finite measures on $(\Omega, \mathcal{M})$ and define the distance from $G$ to $H$:

$$\pi(G, H) = \inf\{\epsilon > 0 : G(A) < H(A^\epsilon) + \epsilon, \ \forall A \in \mathcal{M}\}. \tag{2}$$

Now define

$$Proh(G, H) = \max\{\pi(G, H), \pi(H, G)\}. \tag{3}$$

This definition is the one given by Prohorov (1956) except that in (2) Prohorov uses only closed sets $A$. The definitions are equivalent because for each Borel set $A$ and $\eta > 0$ there is a closed set $B \subset A$ with $G(B \backslash A) < \eta$. Prohorov (1956) shows that the space of finite measures on $(\Omega, \mathcal{M})$ with the distance function $Proh$ is itself a complete separable metric space and that $Proh(G_n, G) \to 0$ iff $G_n \to G$ in the sense of weak convergence. That is

$$Proh(G_n, G) \to 0$$

iff for every bounded continuous function $\varphi$ from $\Omega$ to $\mathbb{R}$

$$\int \varphi(y) dG_n(y) \to \int \varphi(y) dG(y).$$

The Prohorov metric is prominent in the robustness literature. It is usually defined there on probability measures. For measures of equal total mass $\pi$ is a metric and metrizes weak convergence. In particular $\pi$ is symmetric so $Proh = \pi$ on probability measures. See Huber (1981).

When two measures have almost the same mass $\pi$ is almost symmetric as the following lemma shows.

**Lemma 2.4.1.** Let $G$ and $H$ be measures on $(\Omega, \mathcal{M})$ with $G(\Omega) \geq H(\Omega)$. Then

$$\pi(H, G) \leq \pi(G, H) \leq \pi(H, G) + G(\Omega) - H(\Omega).$$

PROOF. Argue as Huber (1981, p.27) does in the special case of probability measures. Let $\pi(H, G) = \epsilon$ and let $\epsilon' > \epsilon$. Consider $A = B^{\epsilon' c}$ in the definition of $\pi(H, G)$, where a superscript $c$ denotes complementation. One obtains

$$H(\Omega) - H(B^{\epsilon'}) < G(\Omega) - G(B^{\epsilon' c \epsilon c}) + \epsilon$$

so that

$$G(B^{\epsilon' c \epsilon c}) < \cdot H(B^{\epsilon'}) + \epsilon + G(\Omega) - H(\Omega).$$

Because $B \subset B^{\epsilon' c \epsilon c}$,

$$G(B) < H(B^{\epsilon'}) + \epsilon + G(\Omega) - H(\Omega).$$

Letting $\epsilon' \downarrow \epsilon$

$$G(B) \leq H(B^{\epsilon}) + \epsilon + G(\Omega) - H(\Omega). \tag{4}$$

From (4) $\pi(G,H) \leq \pi(H,G) + G(\Omega) - F(\Omega)$. Equation (4) was derived without using $G(\Omega) \geq F(\Omega)$ so it still holds when the roles of $G$ and $H$ are reversed. From this $\pi(H,G) \leq \pi(G,H)$ and the lemma is proved. ∎

**Corollary.** If $G_n(\Omega) \to G(\Omega)$ then the following are equivalent:

(i) $\pi(G_n, G) \to 0$

(ii) $\pi(G, G_n) \to 0$

(iii) $Proh(G, G_n) \to 0$

PROOF. Immediate from the lemma and (3). ∎

For probability measures $G_n$ and $G$ Billingsley (1971) gives these equivalent characterizations of weak convergence of $G_n$ to $G$:

a) $\limsup G_n(A) \leq G(A) \ \ \forall$ closed $A$

b) $\liminf G_n(A) \geq G(A) \ \ \forall$ open $A$

c) $\lim G_n(A) = G(A) \ \ \forall\, A$ with $G(\partial A) = 0$

For finite measures the above are all equivalent to $Proh(G_n, G) \to 0$ (Prohorov 1956, Sec. 1.3) if the condition $\lim G_n(\Omega) = G(\Omega)$ is adjoined to a) and b).

Hampel (1971) uses the Prohorov metric to define robustness of a statistical functional. His definition is that the map from the distribution of the data to the distribution of the functional is continuous (uniformly in $n$) when the Prohorov metric is used on both spaces. Hampel's theorem for a statistical functional is that it is robust if and only if it is a continuous mapping from the space of distributions to $I\!R$ where the Prohorov metric is used on the space of distributions.

Any quantile is a Prohorov continuous functional at any distribution that has positive mass in all open intervals about the quantile. An M-estimate with a bounded and strictly

monotone $\psi$ function is Prohorov continuous at every distribution. The functional

$$T_y(F) = F(y)$$

for fixed $y$ is Prohorov continous at every $F$ for which $y$ is a continuity point. The $\alpha$-trimmed mean with $0 < \alpha < 1/2$ is Prohorov continuous at every distribution. More generally an L-estimate

$$T(F) = \int_0^1 F^{-1}(u)M(du)$$

where $M$ is a finite signed measure with support in $[\alpha, 1-\alpha]$ is Prohorov continuous at any $F$ for which no discontinuity point of $F^{-1}$ is a point of mass of $M$.

Many important functionals are not Prohorov continuous. That is to say they are not robust. In particular the mean is not continuous at any distribution function. Higher moments and related quantities such as the standard deviation, correlation and coefficient of variation are not continuous anywhere. Similarly $F(y) - F(y-)$, the jump of $F$ at $y$ is not Prohorov continuous for any $F$ with an atom at $y$.

The mean can be made continuous by considering a smaller space $U$. For example, on a space of distributions with uniformly bounded support, all moments are Prohorov continuous. If for $1 \leq p < q$ the distributions in $U$ have a uniformly bounded $q$'th moment, then the $p$'th moment is Prohorov continuous. (Chung, 1974, Theorem 4.5.2.)

In Chapter 3, one of the conditions used is that $F_\bullet$ as a map from $\mathcal{X}$ to $U$ is Prohorov continuous. In other words as $x' \to x$ the distribution of $Y$ given $X = x'$ converges weakly to the distribution of $Y$ given $X = x$. This sort of continuity assumption would seem to be very mild in practice.

In order to study weight sequences with some negative weights it would be useful to have a metric for weak convergence of finite signed measures. Unfortunately no metrization of weak convergence exists for signed measures, except in trivial cases. See Choquet (1969, Vol. I, Sec. 12 and Theorem 16.9). (It is possible to metrize weak convergence of signed measures on some spaces without compact sets of infinite cardinality. )

Recall that a finite signed measure $G$ can be written $G = G^+ - G^-$ where $G^+$ and $G^-$ are mutually singular measures called, respectively, the positive and negative parts of $G$. The measure $|G| = G^+ + G^-$ is the total variation of $G$ (This is the Jordan decomposition, and it is unique.)

The quantity *Proh* defined by (3) is peculiar, on finite signed measures. It is almost a metric, but sometimes $Proh(G, G) > 0$. Furthermore the triangle inequality might not hold if the space $\Omega$ is ill-behaved. (The triangle inequality holds if $(B^a)^b = B^{a+b}$ for all $B \in \mathcal{M}$ and $a, b > 0$.) Convergence of (3) need not imply weak convergence:

**Example 1.** Let $G = 0$ and $G_n = n^2\delta_{1/n} - n^2\delta_{-1/n}$. Then

$$\max\{\pi(G_n, G), \pi(G, G_n)\} = 2/n \to 0$$

but $G_n$ does not converge weakly to $G$. Consider $\varphi(x) = 1 \wedge (x+1)^+$. $\int \varphi(x)dG_n(x) = n$ and $\int \varphi(x)dG(x) = 0$.

Convergence of (3) combined with

$$\limsup |G_n| < B < \infty$$

can be shown to imply weak convergence—first establish convergence for the signed measures of closed sets and then extend to bounded continuous functions as in Pollard (1984, Exercise IV-12).

We can define a metric that is stronger than weak convergence. For finite signed measures $G$ and $H$ on $(\Omega, \mathcal{M})$ define

$$Proh(G, H) = Proh(G^+, H^+) + Proh(G^-, H^-). \tag{5}$$

*Proh* as defined by (5) is still a metric and $Proh(G_n, G) \to 0$ implies weak convergence of $G_n$ to $G$.

It is possible for $G_n$ to converge weakly to $G$ without $Proh(G_n, G)$ (as defined in (5)) converging to zero.

**Example 2.** Let $G = \delta_0$ and $G_n = 2\delta_0 - \delta_{1/n}$. Then $G_n$ converges weakly to $G$ but

$$Proh(G_n, G) = 2.$$

### 2.4.2 Kolmogorov-Smirnov Metric

The Kolmogorov-Smirnov metric for distributions on $\mathbb{R}$ is

$$KS(G, H) = \sup_y |G(y) - H(y)|, \tag{6}$$

the sup norm of $G - H$. It takes its name from the Kolmogorov-Smirnov statistic $KS(F_n, F)$. The space $U$ can be any set of functions on $\mathbb{R}$. This makes it a convenient metric to use when considering distribution functions corresponding to finite signed measures.

The Glivenko-Cantelli theorem states that $KS(F, F_n) \to 0$ a.s. as $n \to \infty$. In Chapter 3 sufficient conditions are given for $KS(F_x, \hat{F}_x) \to 0$ a.s.

The metric $KS$ is stronger than $Proh$. That is

$$KS(G_n, G) \to 0 \Rightarrow Proh(G_n, G) \to 0,$$

and there are sequences for which $Proh(G_n, G) \to 0$ but $KS(G_n, G)$ does not converge to 0. If $KS(G_n, G) \to 0$ the distribution functions $G_n$ are converging uniformly to $G$ whereas if $Proh(G_n, G) \to 0$ the convergence is pointwise at continuity points of $G$. If $G_n$ is a point-mass at $1/n$ and $G$ is a point-mass at 0, Prohorov but not Kolmogorov-Smirnov convergence takes place.

All the functionals that are continuous under the Prohorov metric are continuous under the Kolmogorov-Smirnov metric. Under this stronger metric, the jump functional

$$J_y(F) = F(y) - F(y-)$$

is continuous everywhere. The mean is nowhere continuous.

Suppose that the map $F_\bullet$ from $\mathcal{X}$ to $U$ is KS continuous. Then the function in the $xy-$plane given by $F_x(y)$ is continuous (uniformly in $y$) when traversed parallel to the $x$ axis, but need not be continuous at all when traversed parallel to the $y$ axis.

**Example 1.** If $\lambda(x) > 0$ is a continuous function and $F_x$ is the Poisson distribution with parameter $\lambda(x)$ then $F_\bullet$ is KS continuous. If $Y/\lambda(x)$ has the Poisson distribution with parameter 1 then $F_\bullet$ is not KS continuous unless $\lambda$ is constant.

A KS continuous $F_\bullet$ can have atoms of fixed location in $\mathcal{Y}$ whose size varies continuously with $x$ but cannot have atoms of fixed size whose locations vary continuously.

The weight function $W_x$ will not usually converge in the KS metric to $\delta_x$. For a symmetric kernel and i.i.d. $X_i$ from a distribution without an atom at $x$, $KS(W_x, \delta_x) \doteq .5$ except for end effects.

### 2.4.3 Vasserstein Metrics

These metrics are described in Bickel and Freedman (1981, Section 8). This section is based on their account. Let $B$ be a separable Banach space with norm $\|\cdot\|$. (This is the space $\mathcal{Y}$ which the reader might assume is $\mathbb{R}$.) For $1 \le p < \infty$ let $U = U_p$ be the space of probability measures $F$ on the Borel $\sigma$-field of $B$ for which $\int \|y\|^p F(dy) < \infty$. Then the Vasserstein metric is the infimum of $\mathcal{E}\left( \|X - Y\|^p \right)^{1/p}$ over all pairs of random variables $X$ and $Y$ with $X \sim F$ and $Y \sim G$. Bickel and Freedman's Lemma 8.1 establishes that $V_p$ is a metric and that the infimum is attained.

The Vasserstein metric provides a way of metrizing $L^p$ convergence. Bickel and Freedman's Lemma 8.3 states that $V_p(G_n, G) \to 0$ if and only if

$$G_n \to G \text{ weakly, and } \int \|y\|^p G_n(dy) \to \int \|y\|^p G(dy).$$

Clearly $V_p$ convergence implies Prohorov convergence. In fact $Proh(F, G) \le \sqrt{V_1(F, G)}$, a result due to Dobrushin (1970). Also for distributions in $U_p$ where $p > p' \ge 1$, $V_{p'}(F, G) \le V_p(F, G)$.

For $B = \mathbb{R}$ with norm $|\cdot|$ there is a convenient formula for $V_p$ due to Major (1978):

$$V_p(F,G) = \left(\int_0^1 |F^{-1}(u) - G^{-1}(u)|^p du\right)^{\frac{1}{p}} \tag{7}$$

so that $V_p$ is a "sideways $L^p$" metric. In particular $V_1(F,G)$ is the area between the d.f.s $F$ and $G$ and hence may also be written:

$$V_1(F,G) = \int_{-\infty}^{\infty} |F(y) - G(y)| dy.$$

The metric $V_2$ is also known as the Mallows metric. Mallows (1972) used the form (7) and established that convergence in the Mallows metric is equivalent to combined weak and $L^2$ convergence.

It is natural to adjoin a $V_\infty$ metric based on essential suprema. Define

$$\operatorname{ess\,sup} F = \sup\{B > 0 : F\{\|Y\| > B\} > 0\}$$

and let $U_\infty$ be the set of probability measures with finite essential suprema. Then define for $F, G \in U_\infty$

$$V_\infty(F,G) = \inf \operatorname{ess\,sup} \|X - Y\|$$

where the infimum is taken over pairs $X \sim F$ and $Y \sim G$. In the case of $B = \mathbb{R}$,

$$V_\infty(F,G) = \sup_{0<u<1} |F^{-1}(u) - G^{-1}(u)|. \tag{8}$$

It is clearly a metric since it is the sup norm of $F^{-1} - G^{-1}$. Also $V_\infty$ convergence implies $V_p$ convergence for all finite $p \geq 1$. The form (8) will be used to define a $V_\infty$ metric on the set of all probability distribution functions, not just those with bounded support. The resulting metric may take infinite values.

Convergence of $V_\infty(G_n, G)$ to 0 implies that $Proh(G_n, G) \to 0$ and $\operatorname{ess\,sup} G_n \to \operatorname{ess\,sup} G$. The converse does not hold as the next example illustrates.

**Example 1.** Let $G_n$ be uniform on the set $[0, 1+1/n] \cup [2+1/n, 3]$ and $G$ be uniform on $[0,1] \cup [2,3]$. Then $G_n \to G$ weakly and the essential suprema converge but $V_\infty(G_n, G) = 1$.

KS convergence and $V_p$ convergence (for $B = \mathbb{R}$) are not comparable. (One is tempted to say they are orthogonal.) $KS$ convergence does not imply $V_1$ convergence and $V_\infty$ convergence does not imply $KS$ convergence.

**Example 2.** Take $B = \mathbb{R}$, $G = \delta_0$ and $G_n = (1 - 1/n)\delta_0 + 1/n\delta_n$. Then

$$KS(G_n, G) = 1/n \to 0$$

but $V_1(G_n, G) = 1$.

**Example 3.** Take $B = \mathbb{R}$, $G = \delta_0$ and $G_n = \delta_{1/n}$. Then

$$V_\infty(G_n, G) = 1/n \to 0$$

but $KS(G_n, G) = 1$.

Vasserstein metrics are useful in describing the distance of $W_x$ from $\delta_x$, when $W_i$ are probability weights. For example $V_1(W_x, \delta_x) = \sum W_i \|X_i - x\|$, the weighted average distance of the observations from the target point. Similarly $V_\infty(W_x, \delta_x)$ is the greatest distance from $x$ of any point used in $\hat{F}_x$. When a nonnegative kernel has bounded support and the bandwidth tends to zero, the resulting vector of weights converges in $V_\infty$ to $x$. The same is generally true of nearest neighbor schemes in which all but a vanishingly small proportion of the observations are given 0 weight.

When $W_x$ is not a probability, it is still convenient to use the Vasserstein distance as a shorthand notation for the distance between $W_x$ and $\delta_x$. Therefore for $1 \leq p < \infty$ define

$$V_p(W_x, \delta_x) = \left( \sum_{i=1}^{n} |W_i| \, \|X_i - x\|^p \right)^{1/p}$$

and

$$V_\infty(W_x, \delta_x) = \sup_{W_i \neq 0} \|X_i - x\| \, .$$

The Vasserstein metrics are also useful in manipulating the quantity $|Y_i - Y_i^x|$, the difference between $Y_i$ and "the value it would have taken had $X_i$ been $x$". To whit:

$$\mathcal{E}\left( |Y_i - Y_i^x|^p \mid X_i = x_i \right) = \int_0^1 |F_{x_i}^{-1}(u) - F_x^{-1}(u)|^p du = V_p(F_x, F_{x_i})^p.$$

Therefore if, as is reasonable, $x_i$ close to $x$ implies $V_p(F_x, F_{x_i})$ is small, the bias due to using an observation from $X = x_i$ instead of $x$ should be small.

The following lemma from Bickel and Freedman is of interest:

**Lemma 2.4.2.** Let $Y_i$ be independent; likewise for $Z_i$; assume their distributions are in $U_p$, $1 \le p < \infty$. Then

$$V_p\left(\mathcal{L}\left(\sum_{i=1}^m Y_i\right), \mathcal{L}\left(\sum_{i=1}^m Z_i\right)\right) \le \sum_{i=1}^m V_p(\mathcal{L}(Y_i), \mathcal{L}(Z_i)).$$

PROOF. Bickel and Freedman (1981,Lemma 8.6).

When $B$ is a Hilbert space, Bickel and Freedman (1981) obtain some sharper results for the Mallows metric $V_2$.

### 2.4.4 Other Metrics

The three metrics considered above are the ones that will be used in Chapters 3 and 4. This section rounds out the discussion of statistical metrics with some other metrics in common usage.

The Levy metric for distributions on the real line is

$$Levy(F,G) = \inf\{\epsilon > 0 : F(x-\epsilon) - \epsilon \le G(x) \le F(x+\epsilon) + \epsilon \ \forall x\}.$$

This metric also metrizes weak convergence. It has a geometric interpretation as $1/\sqrt{2}$ times the maximum distance between the distribution functions taken in the northwest to southeast direction. On the space of subprobability measures $G_n$ converges to $G$ in the Levy metric if and only if $G_n$ converges weakly to $G$ and the total mass $G_n(\mathbb{R})$ converges to $G(\mathbb{R})$ (Chung, 1974, p.94).

The bounded Lipschitz metric (Huber, 1980) also metrizes weak convergence on complete separable metric spaces. Assume the metric is bounded by 1. If necessary replace the metric $d(\cdot,\cdot)$ by the topologically equivalent $d(\cdot,\cdot)/(1+d(\cdot,\cdot))$. Then the bounded Lipschitz metric is

$$BLip(F,G) = \sup\left|\int \phi(y)dF(y) - \int \phi(y)dG(y)\right|$$

where the supremum is taken over functions $\phi$ that satisfy $|\phi(y_1) - \phi(y_0)| \leq d(y_1, y_0)$. Huber (1980, Ch.2) shows that

$$Proh(F,G)^2 \leq BLip(F,G) \leq 2Proh(F,G).$$

The KS metric can be generalized. Rewriting it as

$$KS(F,G) = \sup_y |F(-\infty, y] - G(-\infty, y]|$$

suggests generalizations in which the supremum is taken over different classes of sets. Taking the supremum over all measureable sets yields the total variation metric:

$$TV(F,G) = \sup_{A \in \mathcal{B}} |F(A) - G(A)|$$

a very strong metric. This metric is so strong that $F_n$ does not converge to $F$ in total variation when $F$ has a continuous part. On the other hand it is not strong enough to force $V_1$ convergence (see Example 2.4.2). There are many ways to extend the $KS$ metric to higher dimensional spaces. In finite dimensional Euclidean spaces the most straightforward is to take suprema over lower left orthants. Suprema over half spaces or closed balls may also be considered. For convergence of $F_n$ to $F$ to hold for all $F$ in one of these metrics requires that the class of sets over which the supremum is taken not be too rich. A further generalization is to extend suprema over probabilities of sets to suprema over expectations of functions. For a discussion see Pollard (1984, Ch. 2).

Bickel and Freedman (1981) show that $V_p(F,G) = \epsilon$ if and only if there exist random variables $X \sim F$ and $Y \sim G$ such such that

$$\mathcal{E}\left( \|X - Y\|^p \right)^{1/p} = \epsilon.$$

Similar coupling results hold for some other metrics: $Proh(F,G) \leq \epsilon$ iff some such $X$ and $Y$ satisfy

$$P\left(d(X,Y) \leq \epsilon\right) \geq 1 - \epsilon$$

where $d$ is the metric on $\Omega$, $BLip(F,G) \leq \epsilon$ iff some such $X$ and $Y$ satisfy

$$\mathcal{E}\ (\ d(X,Y)\ ) \leq \epsilon$$

where $d$ is the bounded metric used to define $BLip$ and finally $TV(F,G) \leq \epsilon$ iff some such $X$ and $Y$ satisfy

$$P(X \neq Y) \leq \epsilon.$$

The first and third of these follow from Strassen's theorem (Huber, 1980) and the second from Huber's (1980) generalization of a theorem of Kantorovic and Rubinstein.

## 2.5 Models for $F_\bullet$

As indicated in Sec. 2.1 the $X$'s are obtained either by sampling or by design, and then the $Y$'s are conditionally independent with the corresponding distributions. Given $X_i = x_i$ the distribution of $Y_i$ is $F_{x_i}$. All the results in Chapters 3 and 4 are obtained after imposing some structure (or model) on the set of $F_x$'s.

A very weak model is that $F_\bullet$ is Prohorov continuous. That is

$$x_i \to x \Rightarrow Proh(F_{x_i}, F_x) \to 0,$$

so $Y_i \sim F_{x_i}$ converges to $Y \sim F_x$ in distribution. This is a very reasonable assumption for many applications. It says that values of $x_i$ close to $x$ tend to have $Y$ distributions close to the one at $x$. Absent such an assumption, one would hardly use smoothing techniques. Not much is changed by assuming piecewise Prohorov continuity. For pointwise considerations all that is needed is that $F_\bullet$ is Prohorov continuous at $x$.

A stronger model is that $F_\bullet$ is a location-scale family with location $\mu(x)$ and scale $\sigma(x) \geq 0$. That is

$$F_x^{-1}(u) = \mu(x) + \sigma(x)G^{-1}(u) \tag{1}$$

for some distribution function $G(u)$. $G$ may be normalized to have location 0 and scale 1, for some location and scale functionals. The model (1) is still fairly general and will

be used below to give conditions on $F_\bullet$ a more concrete appearance. When $\sigma(x) > 0$ the location-scale model may also be written

$$F_x(y) = G\left(\frac{y - \mu(x)}{\sigma(x)}\right).$$

It is interesting to note that explicit continuity assumptions need not be made when estimating conditional moments. Stone (1977) assumes that $(X, Y)$ has a distribution for which $\mathcal{E}\ (\ |Y^r|\ ) < \infty$ for some $r > 1$ and obtains global $L^r$ consistency for the regression function. Stone (1977, p.641) explains that continuity assumptions are not needed because the regression function, as a function in $L^r$ can be approximated in $L^r$ norm by a continuous function with bounded support to within any $\epsilon > 0$. Devroye (1981) obtains pointwise strong and weak consistency using the moment condition on $Y$.

Neither Prohorov continuity of $F_\bullet$ nor the existence of a moment of $Y$ is empirically verifiable. Both seem to be mild assumptions.

The main benefit of the continuity assumption on the conditional distributions is that it becomes easier to handle non-random $X$'s. The same theorems will cover the random and the design case. A second minor benefit, is that it is possible to consistently estimate a conditional expectation in some cases where $\mathcal{E}\ (\ |Y|\ )$ does not exist. As a trivial example suppose that the $X_i$ are independent standard Cauchy random variables and that $Y_i = X_i$. Then a uniform nearest neighbor scheme with $k = \sqrt{n}$ provides pointwise consistent estimates of the regression. (We could even have added some well-behaved noise.)

Continuity of $F_\bullet$ will also be considered in other metrics, such as the $KS$ metric and the $V_p$ metrics. Some long range conditions are also imposed on $F_\bullet$. Examples are $\rho(F_x, F_{x_i}) \leq B$ for all $x_i$ and $\rho(F_x, F_{x_i}) \leq M_x|x - x_i|$, where $\rho$ is a statistical metric. The latter is a local ($M$ depends on $x$) Lipschitz condition and also imposes a short range constraint on $F_\bullet$.

**Lemma 2.5.1.** Suppose the location-scale model (1) holds and $\mu$ and $\sigma$ are continuous at $x_0$. Then $F_\bullet$ is Prohorov continuous at $x_0$. If $\mathcal{Y} = I\!R$, $\sigma(x_0) > 0$ and $G$ is continuous

then $F_\bullet$ is KS continuous at $x_0$. If $\mathcal{Y} = \mathbb{R}$ and $G$ has a finite $p$'th moment, $p \geq 1$ then $F_\bullet$ is $V_p$ continuous at $x_0$. If $\mathcal{Y} = \mathbb{R}$ and $G$ is bounded then $F_\bullet$ is $V_\infty$ continuous at $x_0$.

PROOF. Let $Z$ be a random variable with distribution $G$ and characteristic function $g$. Let $x_n \to x_0$ and denote $\mu(x_i)$ by $\mu_i$, $\sigma(x_i)$ by $\sigma_i$. Then

$$\mathcal{E}\, e^{it(\mu_n + \sigma_n Z)} = e^{it\mu_n} g(t\sigma_n) \to e^{it\mu} g(t\sigma) = \mathcal{E}\, e^{it(\mu_0 + \sigma_0 Z)}$$

by continuity of $g$. This establishes the point-wise convergence of the characteristic function of $F_{x_n}$ to that of $F_{x_0}$ which implies Prohorov convergence of $F_{x_n}$ to $F_{x_0}$ and hence Prohorov continuity of $F_\bullet$ at $x_0$.

Suppose $G$ is continuous, $\sigma_0 > 0$ and let $y \in \mathbb{R}$. Then

$$F_{x_n}(y) = G\left(\frac{y - \mu_n}{\sigma_n}\right) \to G\left(\frac{y - \mu_0}{\sigma_0}\right) = F_{x_0}(y)$$

since $G$ is continuous and $1/\sigma(\cdot)$ is continuous at $x_0$. This establishes pointwise convergence of $F_{x_n}$ to $F_{x_0}$. Monotonicity and boundedness of $F_{x_n}$ and $F_{x_0}$ combine to strengthen the result to uniform convergence by a lemma of Chung (1974, p. 133) which is restated in Sec 3.2. (That lemma also requires convergence of all the jumps, but $G$ has none.) Therefore $F_\bullet$ is $KS$ continuous at $x_0$.

Suppose $G$ has a finite $p$'th absolute moment. By the Minkowski inequality

$$\begin{aligned} V_p(F_{x_n}, F_{x_0}) &= \left(\int_0^1 |\mu_n + \sigma_n G^{-1}(u) - \mu_0 - \sigma_0 G^{-1}(u)|^p du\right)^{1/p} \\ &\leq \left(\int_0^1 |\mu_n - \mu_0|^p du\right)^{1/p} + \left(\int_0^1 |\sigma_n - \sigma_0|^p |G^{-1}|^p du\right)^{1/p} \\ &= |\mu_n - \mu_0| + |\sigma_n - \sigma_0| (\mathcal{E}\,|Z|^p)^{1/p}. \end{aligned} \tag{4}$$

Therefore $F_\bullet$ is $V_p$ continuous at $x_0$.

If $G$ is bounded

$$\sup_{0<u<1} |\mu_n + \sigma_n G^{-1}(u) - \mu_0 + \sigma_0 G^{-1}(u)| \leq |\mu_n - \mu_0| + |\sigma_n - \sigma_0| \operatorname{ess\,sup} Z$$

so $F_\bullet$ is $V_\infty$ continuous at $x_0$. ∎

In view of (4) above, a long range condition on $V_p$ is achieved by imposing similar conditions on $\mu$ and $\sigma$ in the location scale family. Some authors implicitly control long range behavior by working in $[0,1]$ and imposing continuity on the regression. This implies uniform continuity and also boundedness. Similarly, in the location scale family a Lipshitz condition on $\mu$ and $\sigma$ implies one on $V_p$.

The Lipschitz condition is a fairly weak short range condition. Most results in the literature assume one or two continuous derivatives of $\mu$ exist. Sharper short range conditions such as the existence of derivatives of $F_\bullet$ at $x$ will not be considered here.

## 2.6 Compact Differentiablity and von Mises' Method

This section provides a brief outline of compact or Hadamard differentiability and of von Mises' method for proving asymptotic normality of statistical functionals. It will be used in Chapter 4 to prove asymptotic normality for a class of running statistical functionals. The material in this section is adapted from Fernholz (1983).

Suppose $T$ is a statistical functional defined on a convex set of distribution functions that contains all empirical distributions and a distribution $F$, from which a sample will be obtained. Let $G$ be a member of the convex set. The von Mises derivative $T'_F$ of $T$ at $F$ is defined by

$$T'_F(G-F) = \frac{d}{dt}T\left(F + t(G-F)\right)|_{t=0}$$

so long as there exists a real function $\phi_F(x)$ (not depending on $G$) such that

$$T'_F(G-F) = \int \phi_F(x)d(G-F)(x).$$

This defines $\phi$ up to an additive constant. The derivative is normalized by taking

$$0 = \int \phi_F(x)dF(x).$$

The function $\phi_F(x)$ is better known to statisticians as the influence function:

$$\phi_F(x) = IC(x;,F,T) = \frac{d}{dt}T\left(F + t(\delta_x - F)\right)|_{t=0}.$$

The quantity $T'_F(G-F)$ is a linear approximation to $T(F)-T(G)$. When $G=F_n$

$$\begin{aligned} T(F_n)-T(F) &\doteq T'_F(F_n-F) \\ &= \int \phi_F(x)dF_n(x) \\ &= \tfrac{1}{n}\sum IC(X_i;F,T). \end{aligned} \tag{1}$$

Since (1) is an average of $n$ i.i.d. random variables it (times $\sqrt{n}$) will have a normal limit provided the variance of $IC(X_i;F,T)$ is finite. Von Mises' method consists of establishing the normality of the linear term and the convergence to zero in probability of the remainder:

$$\sqrt{n}Rem(F_n-F)=\sqrt{n}\left(T(F_n)-T(F)-T'_F(F_n-F)\right).$$

Strictly, *Rem* should be $Rem_F$.

Now we define the compact or Hadamard derivative. For von Mises' method, the set $V$ below is the space of distributions, and $W$ is usually $\mathbb{R}$.

**Definition.** Let $V$ and $W$ be topological vector spaces. A function $T$ from $V$ to $W$ is *compactly differentiable* if there is a continuous linear transformation $T'_F$ from $V$ to $W$ such that for any compact set $K\subset V$

$$\lim_{t\to 0}\frac{T(F+tH)-T(F)-T'_F(tH)}{t}=0$$

uniformly for $H\in K$. The linear transformation $T'_F$ is the *compact derivative* of $T$ at $F$.

When the limit is required to hold uniformly on any bounded set the stronger notion of Frechet differentiability results. When the limit is only required to hold pointwise, the weaker concept of Gateaux differentiability emerges. The Gateaux derivative is very similar to von Mises' derivative. Whenever the compact derivative exists it coincides with the Gateaux. Frechet differentiability is strong enough that the remainder term $\sqrt{n}Rem(F_n-F)\to 0$ in pr., if $T$ has a Frechet derivative at $F$. Unfortunately, Frechet differentiability is too strong to be applicable to most statistical functionals. For example the median is not Frechet differentiable at the uniform distribution on (0,1). The Gateaux

derivative is weak enough that most statistical functionals of interest are differentiable. Gateaux differentiability is not enough to gaurantee that the remainder term converges to 0. The compact derivative was shown by Reeds (1976) to be strong enough, that its existence forces the remainder term to 0 in probability. It is also weak enough that it applies to many statistical functionals. For examples see Reeds (1976) and Fernholz (1983).

In Chapter 4 von Mises' method is used for conditionally estimated statistical functionals $T(F_x)$. It is shown there that existence of the compact derivative together with a Brownian limit for the empirical process $\sqrt{n_x}(\hat{F}_x - F_x)$ and a further mild condition on the weights is sufficient for the remainder term $\sqrt{n_x}Rem(\hat{F}_x - F_x)$ to converge in probability to zero.

# 3 Consistency

This chapter considers consistency of $\hat{F}_\bullet$ for $F_\bullet$ and of $T(\hat{F}_\bullet)$ for $T(F_\bullet)$. We will consider pointwise consistency, i.e. the convergence of $\hat{F}_x$ to $F_x$ for fixed $x \in \mathcal{X}$. Pointwise consistency of $T(\hat{F}_\bullet)$ for $T(F_\bullet)$ follows for continuous $T$. Prohorov (weak), Kolmogorov-Smirnov and Vasserstein consistency of $\hat{F}_x$ are treated.

## 3.1 Introduction and Definitions

Consistency of $\hat{F}_x$ for $F_x$ has two aspects to it: how the distance between $F_x$ and $\hat{F}_x$ is to be measured and the nature of the convergence of the (random) distance so measured, to zero. Possibility for confusion arises because common ways of expressing the distance between two distributions have probabilistic interpretations in terms of variables with those distributions. For example convergence of $\mathcal{L}(Z_n)$ to $\mathcal{L}(Z)$ in the Prohorov metric is equivalent to weak convergence of $Z_n$ to $Z$. When the distance itself is studied as a random variable it may be converging weakly, or strongly or in $L^p$. If the metric converges weakly then its probability law is converging in the Prohorov metric to that of a point-mass at zero. For clarity, the metric interpretation will be used for the distance between $\hat{F}_x$ and $F_x$ and the usual probabilistic concepts will be used for the distance between the metric and 0.

Let $U$ be a metric space containing $G$ and the sequence $G_n$, and let $\rho$ be its metric.

**Definition** $G_n$ is *strongly $U$-consistent* for $G$ if $\rho(G_n, G) \to 0$ a.s. as $n \to \infty$.

**Definition** $G_n$ is *weakly $U$-consistent* for $G$ if $\rho(G_n, G) \to 0$ in pr. as $n \to \infty$.

**Definition** $G_n$ is *$U$-consistent in $L^p$*, $p \geq 1$ for $G$ if $\mathcal{E}\ (\ \rho(G_n, G)^p\ ) \to 0$ as $n \to \infty$.

Either strong or $L^p$ $U$-consistency implies weak $U$-consistency. Neither strong nor $L^p$ $U$-consistency implies the other without added conditions such as boundedness of the metric.

When the set of distributions in $U$ is understood $U$-consistency may be referred to as $\rho$-consistency where $\rho$ is the metric. Write $G_n \to G$ $\rho$ in pr., $G_n \to G$ $\rho$ a.s., and $G_n \to G\rho$ $L^p$ for weak strong and $L^p$ consistency of the sequence $G_n$ for $G$.

We will obtain strong and weak $\rho$ consistency of $\hat{F}_x$ for $F_x$ where $x \in \mathcal{X}$ is fixed. Such consistency is called pointwise consistency.

Two alternatives to pointwise consistency are global consistency and uniform consistency. Global consistency is the convergence of $\rho(\hat{F}_X, F_X)$ to zero where $X$ is a random variable independent of the data and $X, X_1, X_2, \ldots$ are i.i.d. Global $L^p$ consistency was considered by Stone (1977) for several functionals with $\hat{F}_x$ obtained by nearest neighbor methods. In his discussion of Stone's paper, Bickel (1977) remarks that the pointwise notions of convergence would seem to be more important from a practical point of view. Weak or strong pointwise consistency established at almost all $x \in \mathcal{X}$ implies global consistency of the corresponding type. The implication does not hold for pointwise $L^p$ consistency without some other condition such as a bound for the pointwise $L^p$ errors that can be integrated with respect to $\mathcal{L}(X)$. Global consistency does not apply to the design case.

Uniform consistency is said to hold when for any compact $K \subset \mathcal{X}$

$$\sup_{x \in K} \rho(\hat{F}_x, F_x)$$

converges to 0. Weak or strong uniform consistency is of course stronger than the corresponding pointwise concept.

Several pointwise consistency results are proved below for $\hat{F}_x$. Weak and strong pointwise consistency is inherited by continuous functionals.

**Lemma 3.1.1** Let $T$ be a function from the metric space $U$ to the metric space $V$

that is continuous at $F_x \in U$. If $\hat{F}_x$ is strongly (weakly) $U-$ consistent for $F_x$ then $T(\hat{F}_x) \to T(F_x)$ a.s. (in pr.).

PROOF. For strong consistency the proof follows by using the continuity of $T$ on the set of probability 1 for which $\hat{F}_x$ converges to $F_x$. Let $\epsilon > 0$. For weak consistency the probability that $T(\hat{F}_x)$ is within an $\epsilon$–ball of $T(F_x)$ is no less than the probability that $\hat{F}_x$ is within some $\delta$–ball of $F_x$ by the continuity of $T$ and the latter probability converges to 1 by the consistency of $\hat{F}_x$. ■

For weak consistency, Lemma 3.1.1 is a special case of the continuous mapping theorem (see Billingsley (1968) or Pollard (1984)). The general result has convergence in distribution where the above has convergence in probability to a constant. The general version of continuity at the limit is continuity with probability 1 at the (random) limit.

$L^p$ consistency of $\hat{F}_x$ and continuity of $T$ at $F_x$ does not imply $L^p$ consistency of $T$. A further condition, such as Lipschitz continuity of $T$, is needed.

Lemma 3.1.1 asserts the pointwise consistency of $T(\hat{F}_x)$ for $T(F_x)$. Its two conditions are consistency of $\hat{F}_x$ and continuity of $T$. Continuity of statistical functionals with respect to statistical metrics is discussed in Sec. 2.4. The next three sections give sufficient conditions for the consistency of $\hat{F}_x$ in the Prohorov, Kolmogorov-Smirnov, and Vasserstein metrics, in that order. The conditions are expressed in terms of the nature of the continuity of $F_\bullet$, the convergence of the weight measure $W_x$ to $\delta_x$ in an appropriate metric and the rate at which the effective local sample size $n_x$ becomes infinite.

## 3.2 Prohorov Consistency of $\hat{F}_x$

In this section Prohorov continuity of $F_\bullet$ and some regularity conditions on the set of weights are used to establish pointwise weak and strong Prohorov consistency of $\hat{F}_x$.

The Prohorov metric for finite measures is given in Sec. 2.4. Convergence of this metric is equivalent to weak convergence of the measures. Weak convergence of finite signed measures is, except for trivial exceptions, not metrizable. Sec. 2.4 defines a metric

*Proh* that is stronger than weak convergence, on finite signed measures. In this section the weights regarded as a finite signed measure on $\mathcal{X}$ are required to converge in the metric *Proh* to a pointmass at a target point $x$. This is a shorthand way of saying that the sum of the negative weights converges to zero and that for any open set in $\mathcal{X}$ the sum of the positive weights attached to that set converges to 1 or 0 according to whether $x$ is or is not in the set. At the end of this section, more general conditions are given that imply weak convergence of $\hat{F}_x$ to $F_x$, in pr. and a.s. Under these more general conditions, the weight functions can have a nonzero limiting sum of negative weights.

Throughout this section $\mathcal{X}$ and $\mathcal{Y}$ are complete separable metric spaces. The next theorem does most of the work for Prohorov consistency of $\hat{F}_x$.

**Theorem 3.2.1** Let $\varphi$ be a bounded measureable function that is continuous on a set of $F_x$ probability 1. Then under conditions i) and ii) below

$$\int \varphi(y)d\hat{F}_x(y) \to \int \varphi(y)dF_x(y) \text{ in pr.}$$

and under conditions i) and iii) below

$$\int \varphi(y)d\hat{F}_x(y) \to \int \varphi(y)dF_x(y) \text{ a.s.}$$

*i*) $F_\bullet$ is Prohorov continuous at $x$

*ii*) $W_x \to \delta_x$ *Proh* in pr. *and* $n_x \to \infty$ in pr.

*iii*) $W_x \to \delta_x$ *Proh* a.s. *and* $n_x/\log n \to \infty$ a.s.

PROOF. Define

$$\overline{\varphi} = \int \varphi(y)dF_x(y) \quad \text{and} \quad \overline{\varphi}_i = \int \varphi(y)dF_{x_i}(y),$$

let $B = \sup_y |\varphi(y)|$ and $\epsilon > 0$. Then

$$\int \varphi(y)d\hat{F}_x(y) - \int \varphi(y)dF_x(y) = \sum_{i=1}^{n} W_i(\varphi(Y_i) - \overline{\varphi}) - \overline{\varphi}(1 - \sum_{i=1}^{n} W_i). \tag{1}$$

The second term in (1) converges to 0 weakly under ii) and strongly under iii).

By the continuous mapping theorem (Billingsley 1968, Sec. 1.5) there is an open set $\Delta \subset \mathcal{X}$, with $x \in \Delta$ such that $x_i \in \Delta$ implies $|\overline{\varphi}_i - \overline{\varphi}| < \epsilon$. The first term from (1) may now be written

$$\sum W_i(\varphi(Y_i) - \overline{\varphi}) = \sum_{x_i \in \Delta} W_i(\varphi(Y_i) - \overline{\varphi}) + \sum_{x_i \notin \Delta} W_i(\varphi(Y_i) - \overline{\varphi}). \tag{2}$$

The second term in (2) converges to 0 weakly under ii) and strongly under iii) because

$$|\varphi(Y_i) - \overline{\varphi}| \leq 2B.$$

Let $|W| = \sum |W_i|$. Conditionally on the $X$'s the first term in (2) has expectation bounded in absolute value by $2B|W|\epsilon$ and variance bounded by $4B^2/n_x$. If $|W| < 2$ and $n_x > 4B^2/\epsilon^3$ then by Chebychev's inequality the conditional probability that

$$\left| \sum_{x_i \in \Delta} W_i(\varphi(Y_i) - \overline{\varphi}) \right| > 3\epsilon$$

is less than

$$\frac{4B^2/(4B^2/\epsilon^3)}{(3\epsilon - 2\epsilon)^2} = \epsilon.$$

It follows that the unconditional probability

$$P\Big(|\sum_{x_i \in \Delta} W_i(\varphi(Y_i) - \overline{\varphi})| > 3\epsilon\Big) < P(n_x \leq 4B^2/\epsilon^3) + P(|W| \geq 2) + \epsilon \to \epsilon$$

by ii). This establishes the first result of the theorem.

Turning to strong convergence, condition on a sequence of $X$ values satisfying

$$n_x / \log n \to \infty \quad \text{and} \quad Proh(W_x, \delta_x) \to 0. \tag{3}$$

Such sequences have probability 1 under iii). Conditionally on the $X$'s the quantities

$$W_i\,(\varphi(Y_i) - \overline{\varphi}_i)$$

are independent, have expectation 0 and are bounded in absolute value by $|W_i|B$. Using Hoeffding's inequality (see for example Pollard (1984, Appendix B)),

$$\begin{aligned}
&P(\sum_{x_i \in \Delta} |W_i(\varphi(Y_i) - \overline{\varphi})| > (1 + |W|)\epsilon) \\
\leq &P(\sum_{x_i \in \Delta} |W_i(\varphi(Y_i) - \overline{\varphi}_i)| + \sum_{x_i \in \Delta} |W_i(\overline{\varphi}_i - \overline{\varphi})| > (1 + |W|)\epsilon) \\
\leq &P(\sum_{x_i \in \Delta} |W_i(\varphi(Y_i) - \overline{\varphi}_i)| > \epsilon) \\
\leq &\exp(-2\epsilon^2 / 4B^2 \sum W_i^2) \\
= &\exp(-n_x \epsilon^2 B^{-2}/2) \\
\leq &n^{-\epsilon^2 B^{-2} n_x / 2 \log n} \\
< &n^{-2} \qquad (4)
\end{aligned}$$

for large enough $n$ by (3).

Because (4) sums we conclude that the conditional probability

$$P(\sum_{x_i \in \Delta} |W_i(\varphi(Y_i) - \overline{\varphi})| > (1 + |W|)\epsilon \; i.o. \mid X) = 0 \qquad (5)$$

by the Borel-Cantelli lemma. Since (3) implies $|W| \to 1$ a.s. by iii), we may replace $(1 + |W|)\epsilon$ by $3\epsilon$ in (5). Since (5) holds for a set of sequences $X$ with probability 1, by Fubini's theorem

$$P(\sum_{x_i \in \Delta} |W_i(\varphi(Y_i) - \overline{\varphi})| > 3\epsilon \; i.o.) = 0. \qquad \blacksquare$$

Theorem 3.2.1 holds for any complete separable metric spaces $\mathcal{X}$ and $\mathcal{Y}$. The main applications are to Euclidean spaces, but also covered are the unit circle and sphere (for periodic or directional data) the space of continuous functions on a compact interval with metric induced by the sup norm, and the space of infinite real sequences with metric induced by the sup norm.

The condition $n_x / \log n \to \infty$ a.s. can be replaced by the slightly sharper, but less evocative

$$\sum \exp(-n_x \epsilon) \to 0 \text{ a.s.} \qquad \forall \epsilon > 0.$$

Theorem 3.2.1 is enough to prove consistency for many functionals that can be analyzed in terms of a finite number of $\varphi(\cdot)$'s. For example, an M estimate of location generated by a bounded continuous monotone $\psi$ function, with a unique value at $F_x$ must be consistent because the root based on the positive part of $\hat{F}_x$ is consistent and the negative part becomes too small to change the root by much. Note that for signed measures $\hat{F}_x$ the M estimate will not necessarily have a unique value, but the smallest and largest values will be consistent.

**Theorem 3.2.2** Under conditions i) and ii) of Theorem 3.2.1

$$Proh(\hat{F}_x, F_x) \to 0 \text{ in pr.}$$

and under conditions i) and iii) of Theorem 3.2.1

$$Proh(\hat{F}_x, F_x) \to 0 \text{ a.s.}$$

PROOF. Let 0 represent the zero measure. Since $F_x$ is a probability measure

$$\begin{aligned} Proh(\hat{F}_x, F_x) &= Proh(\hat{F}_x^+, F_x) + Proh(\hat{F}_x^-, 0) \\ &= Proh(\hat{F}_x^+, F_x) + \hat{F}_x^-(\mathcal{Y}) \\ &= Proh(\hat{F}_x^+, F_x) + W_x^-(\mathcal{X}) \\ &\to Proh(\hat{F}_x^+, F_x) \end{aligned}$$

weakly under i) and ii) and strongly under i) and iii).

Also

$$\hat{F}_x^+(\mathcal{Y}) = W_x^+(\mathcal{X}) \to 1$$

weakly under i) and ii) and strongly under i) and iii) so by Lemma 2.4.1 it suffices to prove convergence for $\pi(\hat{F}_x^+, F_x)$.

Let $\epsilon > 0$. Because $\mathcal{Y}$ is a complete separable metric space, and $F_x$ is a probability measure there are disjoint sets $B_0, B_1, \ldots, B_r$ with $F_x(\partial B_j) = 0$, $F_x(B_0) < \epsilon/4$ and for $j \geq 1$ $B_j$ has diameter less than $\epsilon$, that is $B_j \subset \{y\}^\epsilon$ whenever $y \in B_j$. Note that

the indicators of the sets $B_j$ are $F_x$–a.e. continuous bounded measureable functions so Theorem 3.2.1 applies to them.

Suppose that $\pi(\hat{F}_x^+, F_x) > \epsilon$. Then there is a set $A \subset \mathcal{Y}$ such that

$$\hat{F}_x^+(A) > F_x(A^\epsilon) + \epsilon \tag{6}$$

where $A^\epsilon$ is defined by equation 2.4.2. Let $A_j = A \cap B_j$ be a partition of $A$. The inequality (6) only happens when either

$$\hat{F}_x^+(A_0) - F_x(A_0^\epsilon) > \epsilon/2 \tag{7}$$

or for some $j \geq 1$

$$\hat{F}_x^+(A_j) - F_x(A_j^\epsilon) > \epsilon/2r. \tag{8}$$

But $\hat{F}_x^+(A_0) \leq \hat{F}_x^+(B_0) \to F_x(B_0) < \epsilon/4$ with weak convergence under i) and ii) and strong convergence under i) and iii). Therefore the probability of the event (7) converges to 0 under i) and ii) and the probability that (7) happens infinitely often is 0 under i) and iii). As for (8)

$$\begin{aligned}\hat{F}_x^+(A_j) - F_x(A_j^\epsilon) &\leq \hat{F}_x^+(B_j) - F_x(A_j^\epsilon) \\ &\leq \hat{F}_x^+(B_j) - F_x(B_j),\end{aligned}$$

so (8) can occur only if

$$\hat{F}_x^+(B_j) - F_x(B_j) > \epsilon/2r$$

and as before this event has probability tending to 0 under i) and ii), and zero probability of infinite occurrence under i) and iii). ∎

For $\mathcal{Y} = \mathbb{R}$ the strong result above can be obtained from strong convergence of the $\hat{F}_x$ probabilities for an appropriate countable set of intervals to the corresponding $F_x$ probabilities.

**Corollary** If $T$ is a statistical functional that is robust at $F_x$ and the $W_i$ are probability weights, then under i) and ii) of Theorem 3.2.1

$$T(\hat{F}_x) \to T(F_x) \text{ in pr.}$$

and under i) and iii) of Theorem 3.2.1

$$T(\hat{F}_x) \to T(F_x) \text{ a.s.}$$

PROOF. Because $T$ is robust at $F_x$, it is continuous at $F_x$ on the space of probability distributions on $\mathcal{Y}$ under the Prohorov metric, by Hampel's theorem. The result then follows from Theorem 3.2.2 and Lemma 3.1.1.

To obtain consistency of running robust functionals when negative weights are used it suffices to show that the functionals are still continuous when extended to finite signed measures.

The mean is not a Prohorov continuous functional, so the Prohorov consistency theorem does not yield a consistency proof for the regression. The mean *is* Prohorov continuous on the space of distribution functions that satisfy $\int |Y|^{1+\delta} dF(Y) < B$ for some $\delta > 0, B < \infty$. This follows for example from Theorem 4.5.2 of Chung (1974). Assuming that $\sup_x \int |Y|^{1+\delta} dF_x(Y) < B$ is not quite enough, since a bound has to hold on the sequence $\hat{F}_x$.

Under the assumption that $|Y| \leq B < \infty$, consistency of the regression function is now easy to obtain.

**Theorem 3.2.3** Let $m(F) = \int y dF(y)$ and assume $|Y| \leq B < \infty$. If conditions i) and ii) of Theorem 3.2.1 hold then

$$m(\hat{F}_x) \to m(F_x) \text{ in pr.}$$

and under conditions i) and iii)

$$m(\hat{F}_x) \to m(F_x) \text{ a.s.}$$

PROOF. Use $\varphi(Y_i) = Y_i$. Because $|Y| \leq B$ Theorem 3.2.1 applies. ∎

Devroye (1981) obtains strong pointwise consistency for the regression function assuming bounded $Y$. His conditions on the weights are slightly stronger than those above,

(he uses probability weights and imposes a stronger condition on the largest of them) but he does not place the Prohorov continuity condition on the conditional distribution of $Y$. He obtains weak pointwise convergence without using bounded $Y$, for nearest neigbor weights (that are exactly 0 for all but a vanishingly small fraction of the observations) and for a restricted class of kernel estimates. Devroye (1982) extends the regression consistency results and obtains some sufficient conditions under the bounded $Y$ assumption.

The theorems above need slight modification to apply to weight schemes, including many kernel methods, that have asymptotically non-negligible negative weights. For $Proh(W_x, \delta_x)$ to vanish, the sum of the negative weights has to go to 0. More typically there is a constant $b \in (0, \infty)$ such that

$$Proh\left(W_x^-, b\delta_x\right) \to 0 \tag{9a}$$

and

$$Proh\left(W_x^+, (1+b)\delta_x\right) \to 0. \tag{9b}$$

Then $Proh(W_x, \delta_x) \to 2b > 0$. The conclusions of Theorem 3.2.1 still hold when in pr. and a.s. versions of (9ab) are used. Theorem 3.2.2 won't hold because $\hat{F}_x^-(y) \to b$. The essence of Theorem 3.2.2 is that $\hat{F}_x \to F_x$ in the sense of weak convergence, and that result can be generalized.

Let $\mathcal{O}$ be the set of open sets in the topology of weak convergence.

**Definition** $\hat{F}_x \to F_x$ *weakly* in pr. if $F_x \in O \in \mathcal{O}$ implies

$$\lim_{n\to\infty} P\left(\hat{F}_x \in O\right) = 1.$$

**Definition** $\hat{F}_x \to F_x$ *weakly* a.s. if $F_x \in O \in \mathcal{O}$ implies

$$P\left(\lim_{n\to\infty} \hat{F}_x \in O\right) = 1.$$

The following theorem employs a sequence of nonnegative random variables

$$b_n = b_n(X_1, \ldots, X_n).$$

Useful possibilities are $b_n = |W_x^-|$ and $b_n = b = \int K^-(v)dv$ for a kernel function $K$.

**Theorem 3.2.4** Let $\varphi$ be a bounded measureable function that is continuous on a set of $F_x$ probability 1. Then under conditions i) and ii) below

$$\int \varphi(y)d\hat{F}_x(y) \to \int \varphi(y)dF_x(y) \text{ in pr.} \quad \text{and} \quad \hat{F}_x \to F_x \textit{ weakly} \text{ in pr.}$$

Under conditions i) and iii) below

$$\int \varphi(y)d\hat{F}_x(y) \to \int \varphi(y)dF_x(y) \text{ a.s.} \quad \text{and} \quad \hat{F}_x \to F_x \textit{ weakly} \text{ a.s.}$$

*i)* $F_\bullet$ is Prohorov continuous at $x$

*ii)* There exist nonnegative r.v.s $b_n(X_1, \ldots, X_n)$ such that:

$Proh\,(W_x^+, (1+b_n)\delta_x) \to 0$ in pr.

$Proh\,(W_x^-, b_n\delta_x) \to 0$ in pr.

$\forall \epsilon > 0\ \exists B_\epsilon < \infty$ with $\limsup P(b_n \geq B_\epsilon) < \epsilon$

$n_x \to \infty$ in pr.

*iii)* There exist nonnegative r.v.s $b_n(X_1, \ldots, X_n)$ such that:

$Proh\,(W_x^+, (1+b_n)\delta_x) \to 0$ a.s.

$Proh\,(W_x^-, b_n\delta_x) \to 0$ a.s.

$\exists B < \infty$ with $P(\limsup b_n \geq B) = 0$

$n_x \to \infty$ a.s.

PROOF. For $\varphi$ bounded, measureable and continuous a.e. $[F_x]$ write

$$\begin{aligned} &|\int \varphi(y)d\hat{F}_x(y) - \int \varphi(y)dF_x(y)| \\ &\leq |\int \varphi(y)d\hat{F}_x^+(y) - (1+b_n)\int \varphi(y)dF_x(y)| + |\int \varphi(y)d\hat{F}_x^-(y) - b_n \int \varphi(y)dF_x(y)|. \quad (10) \end{aligned}$$

The proof of Theorem 3.2.1 can be adapted to show that both terms in (10) converge to 0, in pr. under i) and ii) and a.s. under i) and iii). The bounding conditions on $b_n$ are used in the Chebychev and Hoeffding arguments applied to the first term in (2).

Let $O \in \mathcal{O}$ contain $F_x$. Then there is a basic open set $\mathcal{N}$ such that $F_x \in \mathcal{N} \subset O$. The neighborhood base at $F_x$ in the topology of weak convergence consists of sets of the form

$$\cap_{j=1}^{k}\{G : |\int \varphi_j(y)dG(y) - \int \varphi_j(y)dF_x(y)| < \epsilon \quad j = 1, \ldots, k\}$$

for nonnegative integers $k$, positive $\epsilon$ and bounded continuous functions $\varphi_j$. It follows from the convergence of (10) to 0 for any finite set of bounded continuous functions $\varphi_j$ that $\hat{F}_x \to F_x$ weakly in pr. under i) and ii) and weakly a.s. under i) and iii). ∎

Perhaps the conditions above can be further weakened to weak convergence of $W_x$ to $\delta_x$, in pr. and a.s. Such generality is not needed in most smoothing applications. Consider the following example: Let $r_i$ be an enumeration of a countable dense subset of $\mathcal{X}$. Let

$$W_x^+ = \delta_x + \sum_i 2^{-i}\delta_{r_i} \quad \text{and} \quad W_x^- = \sum_i 2^{-i}\delta_{t_{in}}$$

where $d(r_i, t_{in}) \leq 1/n$. Then $W_x \to \delta_x$ weakly, but does not satisfy the conditions of Theorem 3.2.4. For applications, it is reasonable to assume that $|W_x| \to 0$ on the complement of any open set containing $x$. This was used to handle the second term in (2).

## 3.3 KS Consistency of $\hat{F}_x$

This section provides sufficient conditions for $KS(\hat{F}_x, F_x)$ to vanish. The result is similar to that for the Prohorov metric except that the Prohorov continuity condition on $F_\bullet$ is strengthened to KS continuity. Fortunately it is not necessary to strengthen the Prohorov convergence of $W_x$ to KS convergence, since the latter only holds together with $n_x \to \infty$ when the number of $x_i$ equal to $x$ grows without bound. The Kolmogorov-Smirnov metric is stronger than the Prohorov metric so that convergence in the former implies convergence in the latter.

Any functional $T$ that is continuous when the Prohorov metric is used on its domain is also continuous when the KS metric is used. Functionals such as $J_y(F) = F(y) - F(y-)$ are continuous when the KS metric is used on the distributions, but may not be when the

Prohorov metric is used. Therefore the KS consistency results of this section are useful in situations where there are atoms in the distribution of $Y$.

First we note for later use:

**Lemma 3.3.1** Let $F$ be a distribution function on $\mathbb{R}$. Let $J$ be the set of points of jump of $F$ and let $Q$ be the set of rational numbers. If

$$F_n(y) \to F(y) \quad \forall y \in Q$$

and

$$F_n(y) - F_n(y-) \to F(y) - F(y-) \quad \forall y \in J$$

then $KS(F_n, F) \to 0$.

PROOF. This is proved in Chung (1974, p.133).

**Lemma 3.3.2** Let $y_0 \in \mathcal{Y} = \mathbb{R}$. Under conditions i) and ii) below

$$\hat{F}_x(y_0) \to F_x(y_0) \text{ in pr.}$$

and under conditions i) and iii) below

$$\hat{F}_x(y_0) \to F_x(y_0) \text{ a.s.}$$

*i)* $F_\bullet$ $KS$ continuous at $x$

*ii)* $W_x \to \delta_x$ *Proh* in pr. and $n_x \to \infty$ in pr.

*iii)* $W_x \to \delta_x$ *Proh* a.s. and $n_x / \log n \to \infty$ in pr.

**Definition** A sequence will be said to *converge appropriately* if it converges weakly under conditions i) and ii) and strongly under conditions i) and iii).

PROOF. Let $\varphi(Y_i) = 1_{Y_i \le y_0}$. If $y_0$ is a continuity point of $F_x$ then $\varphi$ satisfies the conditions of Theorem 3.2.1 and so $\hat{F}_x(y_0)$ converges appropriately to $F_x(y_0)$. If $y_0$ is not a continuity point of $F_x$ then $\varphi(\cdot)$ though bounded and measureable, fails to be continuous a.e. $[F_x]$.

In the proof of Theorem 3.2.1 the a.s. continuity of $\varphi(\cdot)$ was only used to establish the existence of an open set $\Delta \ni x$ such that $x_i \in \Delta$ implies $|\int \varphi(y)dF_{x_i}(y) - \int \varphi(y)dF_x(y)| < \epsilon$. But $KS$ continuity of $F_\bullet$ at $x$ gaurantees the existence of such a set and so with this modification we can establish the appropriate convergence of $\hat{F}_x(y_0)$ to $F_x(y_0)$ as in Theorem 3.2.1. ∎

**Lemma 3.3.3** Let $y_0 \in \mathcal{Y} = I\!R$. Under conditions i) and ii) of Lemma 3.3.2

$$\hat{F}_x(y_0) - \hat{F}_x(y_0-) \to F_x(y_0) - F_x(y_0-) \text{ in pr.}$$

and under conditions i) and iii) of Lemma 3.3.2

$$\hat{F}_x(y_0) - \hat{F}_x(y_0-) \to F_x(y_0) - F_x(y_0-) \text{ a.s.}$$

PROOF. Proceed as in the proof of Lemma 3.3.2. Let $\varphi(Y_i) = 1_{Y_i=y_0}$. If $y_0$ is not an atom of $F_x$ then the proof follows from Theorem 3.2.1. If $y_0$ is an atom, $KS$ continuity of $F_\bullet$ implies that the open set $\Delta$ required in Theorem 3.2.1 for $\varphi(\cdot)$ exists. In either case $\hat{F}_x(y_0) - \hat{F}_x(y_0-)$ converges appropriately to $F_x(y_0) - F_x(y_0-)$.

**Theorem 3.3.1** Let $\mathcal{Y} = I\!R$. Under conditions i) and ii) of Lemma 3.3.2

$$\hat{F}_x \to F_x \quad KS \text{ in pr.}$$

and under conditions i) and iii) of Lemma 3.3.2

$$\hat{F}_x \to F_x \quad KS \text{ a.s.}$$

PROOF. Define

$$\widetilde{F}_x(y) = \frac{\hat{F}_x^+(y)}{\sup_y \hat{F}_x^+(y)} = \frac{\hat{F}_x^+(y)}{W_x^+(\mathcal{X})}$$

Now

$$\begin{aligned} KS(\hat{F}_x, F_x) &\le KS(\hat{F}_x^+, F_x) + KS(\hat{F}_x^-, 0) \\ &\le KS(\widetilde{F}_x, F_x) + KS(\hat{F}_x^+, \widetilde{F}_x) + KS(\hat{F}_x^-, 0). \end{aligned} \tag{1}$$

The third term in (1) is bounded by $|W_x^-(\mathcal{X})|$ and the second term is bounded by

$$W_x^+(\mathcal{X})\left|\frac{1}{W_x^+(\mathcal{X})} - 1\right|$$

both of which converge appropriately to 0.

Write

$$\widetilde{F}_x(y) = \sum \widetilde{W}_i 1_{Y_i \le y}$$

where $\widetilde{W}_i = W_i^+ / \sum_j W_j^+$. The weights $\widetilde{W}_i$ satisfy condition ii) when the $W_i$ do and similarly for condition iii).

For strong convergence, apply Lemma 3.3.2 with weights $\widetilde{W}_i$ at points in the set $Q$ of rational numbers and Lemma 3.3.3 to all the points in $J$, the set of points of jump of $F_x$. (The set $Q \cup J$ is countable.) Then except on the union of a countable number of null sets (which is again a null set)

$$\widetilde{F}_x(y) \to F_x(y) \quad \forall y \in Q$$

and

$$\widetilde{F}_x(y) - \widetilde{F}_x(y-) \to F_x(y) - F_x(y-) \quad \forall y \in J.$$

Therefore with probability 1

$$KS(\widetilde{F}_x, F_x) \to 0$$

by Lemma 3.3.1.

For weak convergence, let $\epsilon > 0$. Select a finite grid

$$-\infty = y_0 < y_1 < \ldots y_{r-1} < y_r = \infty$$

such that the $F_x$ probability of each open interval $(y_j, y_{j+1})$, $0 \le j < r$ is less than $\epsilon$. (The grid contains any atoms of $F_x$ that are greater than $\epsilon$.) By Lemmas 3.3.2 and 3.3.3 $\widetilde{F}_x(y_j) \to F_x(y_j)$ in pr. and $\widetilde{F}_x(y_j) - \widetilde{F}_x(y_j-) \to F_x(y_j) - F_x(y_j-)$ in pr. at each $y_j$. Therefore $KS(\widetilde{F}_x, F_x) \to 0$ in pr. by a standard multi-$\epsilon$ argument that uses the monotonicity of $\widetilde{F}_x$ and $F_x$. ∎

Although Theorem 3.3.1 assumes the $KS$ continuity of $F_\bullet$ at $x$, it was only used to get the continuity of the running probabilities $F_\bullet(y_0)$ and jumps $F_\bullet(y_0) - F_\bullet(y_0-)$ at $x$. Nowhere was the uniform continuity of these quantities that $KS$ continuity imposes explicitly used. Yet it follows from Lemma 3.3.1 that the continuity of the jumps and probabilities at $x$ implies the $KS$ continuity of $F_\bullet$ at $x$.

The results of this section were obtained under the assumption that $Proh(W_x, \delta_x) \to 0$. This can be weakened to accomodate weights that have asymptotically nonnegligible negative components. The conditions of Theorem 3.2.4 are adequate. The proof of Theorem 3.3.1 must be modified slightly: consider $KS(\hat{F}_x^+, (1+b_n)F_x)$ and $KS(\hat{F}_x^-, b_n F_x)$.

It should be possible to extend the results of this section to Glivenko-Cantelli classes of sets in $\mathbb{R}^d$.

## 3.4 Vasserstein Consistency of $\hat{F}_x$

Recall that $\hat{F}_x \to F_x$ in the Vasserstein metric $V_p$ iff $\hat{F}_x \to F_x$ in the Prohorov metric and $\int |Y|^p d\hat{F}_x \to \int |Y|^p dF_x$. The main reason to consider these metrics is to study the corresponding moments, particularly the regression function. The main advantages to using the Vasserstein metric instead of a direct method, is that with the bias-variance decomposition based on the $Y_i^x$, the bias term is conveniently handled. The triangle inequality for metrics can be used to split the problem into bias and variance parts and some conditions on the weights can be expressed naturally in terms of Vasserstein distances.

The results for strong convergence are not as sharp as those obtainable by direct arguments based on the regression function. The sharpest available results appear to be those of Zhao and Fang (1985) and Zhao and Bai (1984). The paper by Zhao and Fang considers what are essentially uniform kernels and obtains strong global consistency. The paper by Zhao and Bai considers a very general family of nearest neighbor methods and obtains strong pointwise consistency. They exploit an asymptotic equitability constraint

that in the language of Chapter 2 is

$$\sup_n \max_i n_x W_i < \infty$$

for their probability weights. If $n_x$ is the effective sample size, then no observation should get too great a multiple of the "fair share" $1/n_x$. (Of course most observations get an infinitesimal fraction of $1/n_x$.) Both papers consider the sampling case and assume $\mathcal{E}\,|Y|^p < \infty$ for some $p > 1$. Most other published works use at least a finite second moment for $Y$. The indirect results given here use "off the shelf" laws of large numbers for triangular arrays. For strong convergence one can do better by exploiting relationships between the rows of the arrays.

It follows from straightforward analysis that $V_p(\hat{F}_x, F_x) \to 0$ in pr. iff

$$Proh(\hat{F}_x, F_x) \to 0 \text{ in pr.} \quad \text{and} \quad \int |Y|^p d\hat{F}_x \to \int |Y|^p dF_x \text{ in pr.}$$

By considering the fixed points in the sample space, $V_p(\hat{F}_x, F_x) \to 0$ a.s. iff

$$Proh(\hat{F}_x, F_x) \quad \text{and} \quad \int |Y|^p d\hat{F}_x - \int |Y|^p dF_x \to 0 \text{ a.s.}$$

For Vasserstein consistency, the conditions for Prohorov consistency are strengthened. Assuming Prohorov consistency, the weak or strong consistency of $V_p$ is equivalent to the weak or strong consistency of the $p$'th absolute moment.

The bias-variance split is

$$V_p(\hat{F}_x, F_x) \le V_p(\hat{F}_x, \hat{F}_x^x) + V_p(\hat{F}_x^x, F_x)$$

where

$$\hat{F}_x^x = \sum W_i 1_{Y_i^x \le y}. \tag{1}$$

The bias term will be handled by direct consideration of $V_p(\hat{F}_x, \hat{F}_x^x)$. For the variance term it is easier to work with $\int |Y|^p d\hat{F}_x - \int |Y|^p dF_x$.

We consider first the variance term. For strong convergence of the variance term, we need strong convergence for certain row sums of random variables in a triangular

array. This is more difficult to obtain than strong convergence of sample means and more moments are assumed. The source for most of the results on strong convergence including parts (i) through (iv) of the next lemma is Stout (1969). That reference has very sharp laws of large numbers for triangular arrays, under conditions that are much more general than required here.

**Lemma 3.4.1** Let $W_{nk}$ be fixed real numbers for $1 \leq k \leq n < \infty$ and let $D_k$ be i.i.d. random variables with $\mathcal{E}(D_k) = 0$. Set $n_x^{-1} = \sum_{k=1}^{n} W_{nk}^2$ and $T_n = \sum_{k=1}^{n} W_{nk} D_k$. Then $T_n \to 0$ a.s. if any of the following sets of conditions holds:

(i) $|W_{nk}| \leq Bn^{-\alpha}$, and $n_x/\log n \to \infty$, and $\mathcal{E}|D_k|^{2/\alpha} < \infty$

(ii) $|W_{nk}| \leq Bn^{-\alpha}$, and $n_x/\log n \to \infty$, and $\mathcal{E}|D_k|^{2+1/\alpha} < \infty$

(iii) $|W_{nk}| \leq Bn^{-\alpha}$, and $n_x \geq Bn^{1-\alpha\lambda}$, and $\mathcal{E}|D_k|^{2+\lambda} < \infty$

(iv) $|W_{nk}| \leq Bk^{-1/2}$, and $n_x \geq Bn^{\alpha}$, and $\mathcal{E}|D_k|^{2} < \infty$,

where $B > 0$, $t > 0$, $\lambda > 0$ and $\alpha \in (0,1)$ are constants.

PROOF. Note that $\frac{n_x}{\log n} \to \infty$ implies $\sum \exp(-tn_x) < \infty, \forall t > 0$. Stout (1969) uses the latter condition.

Part (i) follows from Stout's Corollary 1, which is derived from his Theorem 1(i) with $\beta = 1 - \alpha$. Part (ii) follows from Stout's Theorem 1(i) with $\beta = \alpha$. Part (iii) follows from Stout's Theorem 1(i) with $\beta = \alpha(1+\lambda) - 1$. Part (iv) follows from Stout's Theorem 2. ∎

Conditions (i) and (ii) place the mildest restrictions on the growth of $n_x$. For $\alpha > 1/2$ (i) is preferred to (ii) and the reverse holds for $\alpha < 1/2$. When stronger conditions are placed on the growth of $n_x$ a better tradeoff between the bound on $|W_i|$ and the number of moments required of $D_k$ can be obtained via (iii). Part (iv) is unusual in that the bound is not on the maximum weight in a row, but in the maximum weight ever placed on a given $D_k$. It allows a milder moment condition.

**Definition** A sequence $Z_i$ *converges completely* to 0 if $\forall \epsilon > 0$

$$\sum_{i=1}^{\infty} P(|Z_i| > \epsilon) < \infty.$$

Complete convergence implies a.s. convergence by the Borel-Cantelli lemma and is in fact strictly stronger that a.s. convergence. Under conditions (i) through (iii) complete convergence to 0 is obtained.

If the moment conditions in Lemma 3.4.1 are suitably strengthened, the $D_k$ do not need to be identically distributed.

**Lemma 3.4.1'** Let $W_{nk}$ be fixed real numbers for $1 \leq k \leq n < \infty$ and let $D_k$ be independent random variables with $\mathcal{E}(D_k) = 0$. Set $n_x^{-1} = \sum_{k=1}^{n} W_{nk}^2$ and $T_n = \sum_{k=1}^{n} W_{nk} D_k$. Then $T_n \to 0$ a.s. if any of the following sets of conditions holds:

(i) $|W_{nk}| \leq Bn^{-\alpha}$, and $n_x/\log n \to \infty$, and $\mathcal{E}\left( |D_k|^{2/\alpha} (\log^+ |D_k|)^{1+\eta} \right) < B$

(ii) $|W_{nk}| \leq Bn^{-\alpha}$, and $n_x/\log n \to \infty$, and $\mathcal{E}\left( |D_k|^{2+1/\alpha} (\log^+ |D_k|)^{1+\eta} \right) < B$

(iii) $|W_{nk}| \leq Bn^{-\alpha}$, and $n_x \geq Bn^{1-\alpha\lambda}$, and $\mathcal{E}\left( |D_k|^{2+\lambda} (\log^+ |D_k|)^{1+\eta} \right) < B$

where $B > 0$, $\lambda > 0$, $\eta > 0$, and $\alpha \in (0, 1)$ are constants.

PROOF. Items i-iii follow from Stout's Theorem 4 in the same way that the corresponding parts of Lemma 3.4.1 do from Stout's Theorem 3. ∎

Stout does not provide a version of his Theorem 2 for the non identically distributed case, so there is no Lemma 3.4.1'(iv).

The following technical lemma from Chung (1974) is used for weak convergence of the variance term.

**Lemma 3.4.2** Let $\{\theta_{nj}, 1 \leq j \leq k_n\}$ be a double array of complex numbers such that as $n \to \infty$:

$$\max_{1 \leq j \leq k_n} |\theta_{nj}| \to 0 \tag{2a}$$

$$\sum_{1=1}^{k_n} |\theta_{nj}| \leq M < \infty \tag{2b}$$

$$\sum_{1=1}^{k_n} \theta_{nj} \to \theta \tag{2c}$$

where $\theta$ is a finite complex number. Then

$$\prod_{i=1}^{k_n}(1+\theta_{nj}) \to e^{\theta}.$$

PROOF. Chung (1974, p. 199)

**Corollary** Let $\{\theta_{nj}, 1 \le j \le k_n\}$ be a double array of complex random variables such that as $n \to \infty$:

$$\max_{1\le j\le k_n} |\theta_{nj}| \to 0 \text{ in pr.} \tag{3a}$$

$$P(\sum_{1=1}^{k_n} |\theta_{nj}| \le M) \to 1 \tag{3b}$$

$$\sum_{1=1}^{k_n} \theta_{nj} \to \theta \text{ in pr.} \tag{3c}$$

where $\theta$ is a finite complex number and $M < \infty$ is a constant. Then

$$\prod_{i=1}^{k_n}(1+\theta_{nj}) \to e^{\theta} \text{ in pr.}$$

PROOF. Let $\epsilon > 0$. By Lemma 3.4.2 there exists $\delta > 0$ such that $\max_{1\le j\le k_n} |\theta_{nj}| < \delta$ and $\sum_{1=1}^{k_n} |\theta_{nj}| \le M$ and $|\sum_{1=1}^{k_n} \theta_{nj} - \theta| < \delta$ together imply

$$\left|\prod_{i=1}^{k_n}(1+\theta_{nj}) - e^{\theta}\right| < \epsilon.$$

Therefore

$$P\left(\left|\prod_{i=1}^{k_n}(1+\theta_{nj}) - e^{\theta}\right| > \epsilon\right) \to 0. \qquad \blacksquare$$

The next two lemmas establish weak and strong convergence of the variance term in (1), assuming the weak and strong (respectively) convergence of $Proh(\hat{F}_x^x, F_x)$.

**Lemma 3.4.3** For $p \ge 1$ suppose that

$$\mu_p(x) = \int |y|^p dF_x(y) < \infty$$

and let $W_i$ be weights satisfying

$$n_x \to \infty \text{ in pr.} \tag{4a}$$

$$\sum_{i=1}^{n} W_i \to 1 \text{ in pr.} \tag{4b}$$

$$P(\sum_{i=1}^{n} |W_i| \le B) \to 1 \tag{4c}$$

for some fixed $B < \infty$. Then

$$\sum_{i=1}^{n} W_i \, |Y_i^x|^p \to \mu_p(x) \text{ in pr.}$$

PROOF. Let

$$Z_i = |Y_i^x|^p - \mu_p(x).$$

Then

$$\sum_{i=1}^{n} W_i \, |Y_i^x|^p - \mu_p(x) = \sum_{i=1}^{n} W_i Z_i - \mu_p(x)\Big(1 - \sum_{i=1}^{n} W_i\Big). \tag{5}$$

The second term in (5) converges to 0 in pr. by (4b).

Let $g$ be the characteristic function of the $Z_i$. Then

$$\mathcal{E}\, e^{it\sum_j W_j Z_j} = \mathcal{E} \prod_j g(tW_j) \tag{6}$$

and it suffices to show that (6) converges to 1. In fact, because the integrand in (6) is bounded, it suffices to show

$$\prod_j g(tW_j) \to 1 \text{ in pr.} \tag{7}$$

For $t = 0$, (7) is trivial; suppose $t \neq 0$. Because $\mathcal{E}(Z_i) = 0$, and all the $Z_i$ have the same characteristic function $g$,

$$g(tW_i) = 1 + \theta_{nj}$$

where for any $\epsilon > 0$, there is a $\delta > 0$ such that

$$|\theta_{nj}| < \epsilon \, |tW_j| \quad \text{whenever} \quad \max_{1 \le j \le n} |W_j| < \delta. \tag{8}$$

Also $n_x \to \infty$ in pr. implies that $\max_j |W_j| \to 0$ in pr. and hence

$$\max_j |\theta_{nj}| \to 0 \text{ in pr.} \tag{9}$$

From (8) and (9), with $M > B/|t|$

$$P(\sum_j |\theta_{nj}| \leq M) \to 1 \tag{10}$$

and finally (10) and (9) and (8) imply

$$\sum_j \theta_{nj} \to 0 \text{ in pr.} \tag{11}$$

By the Corollary to Lemma 3.4.2, with $\theta = 0$, (7) follows from (9), (10) and (11). ∎

**Lemma 3.4.4** For $p \geq 1$ and i.i.d. $Y_i^x \sim F_x$, suppose that $|Y_i^x|^p$ satisfies one of the moment conditions in Lemma 3.4.1(i-iv) and that $W_i$ satisfy the corresponding condition a.s. Suppose also that the $W_i$ are independent of the $|Y_i^x|^p$ and satisfy the further condition

$$\sum W_i \to 1 \text{ a.s.} \tag{12}$$

Then

$$\sum W_i |Y_i^x|^p \to \mu_p(x) = \int |y|^p dF_x(y) \text{ a.s.}$$

PROOF. Let $D_i = |Y_i^x|^p - \int |y|^p dF_x(y)$. Then

$$\sum W_i |Y_i^x|^p = \sum W_i D_i - \mu_p(x)(1 - \sum W_i). \tag{13}$$

The second term in (13) converges to 0 a.s. by (12). Whichever moment condition from Lemma 3.4.1 is satisfied by $|Y_i^x|^p$, it is also satisfied by $D_i$. The $D_i$ are i.i.d. with mean zero. This also holds conditionally on the $W_i$, by independence.

Condition on $W_i = w_i$ that satisfy the requirements of Lemma 3.4.1. Then $\sum w_i D_i \to 0$ a.s. By Fubini's theorem, we can remove the conditioning and so $\sum W_i D_i \to 0$ a.s. ∎

The bias term is $V_p(\hat{F}_x, \hat{F}_x^x) = (\sum W_i |Y_i - Y_i^x|^p)^{1/p}$. Conditionally on $X$, the mean of $V_p^p$ is

$$\sum W_i V_p(F_{x_i}, F_x)^p$$

and the variance is bounded by

$$\sum W_i^2 V_{2p}(F_{x_i}, F_x)^{2p}.$$

To control the bias term, conditions governing the behavior of $V_p(F_{x_i}, F_x)$ as a function of $x_i$ can be traded off against conditions governing how the weight measure $W_x$ converges to $\delta_x$. If $\mathcal{X}$ is compact and $V_p(F_{x_i}, F_x)$ is a continuous function of $x_i$ then it is bounded. The boundedness of $V_p(F_{x_i}, F_x)$ allows relatively weak conditions to be imposed on $W_x$. At the other end of the spectrum, $V_\infty$ convergence of $W_x$ allows weak conditions to be placed on $V_p(F_{x_i}, F_x)$.

Mack and Silverman (1982) assume a uniform (in $x$) bound on $\int |y|^2 dF_x(y)$, which they describe as a mild condition. (They establish uniform convergence of the regression over suitable bounded intervals.) This is weaker than the boundedness of $Y$ that Devroye (1981) uses which as they point out does not even allow the usual normal linear model. Their condition does not allow $(X, Y)$ to be bivariate normal with nonzero correlation. A uniform bound on $\int |y|^2 dF_x(y)$ implies a uniform bound on $V_2(F_{x_i}, F_x)$.

Lemma 3.4.6 places conditions on $V_p(F_{x_i}, F_x)$, such as

$$V_p(F_{x_i}, F_x) \leq M_x(|x_i - x| + |x_i - x|^a)$$

for $a \geq 1$. The first term dominates for $x_i$ near $x$, where most of the observations are asymptotically, and the second regulates the long range behavior of the model $F_\bullet$. Recall (Sec. 2.5) that for a location-scale family

$$F_x^{-1}(u) = \mu(x) + \sigma(x) F^{-1}(u), \; u \in (0,1)$$

the following bound holds:

$$V_p(F_{x_i}, F_x) \leq |\mu(x_i) - \mu(x)| + |\sigma(x_i) - \sigma(x)| \left( \int |F^{-1}(u)|^p du \right)^{1/p}.$$

It follows that in a location-scale family conditions on the conditional location and scale imply similar conditions on $V_p$. A range of conditions relating $V_p(F_{x'}, F_x)$ to $\|x' - x\|$

is considered, and the weaker the $V_p(F_{x'}, F_x)$ condition is, the stronger the condition imposed on $W_x$ must be.

The next lemma is used in the proof of Lemma 3.4.6(ii) and is used several times in Chapter 4.

**Lemma 3.4.5** Let $X$ and $Y$ be random variables with $\mathcal{E}\,(\,|Y|\,|\,X\,) < \infty$ a.s. and let $\epsilon > 0$. Then

$$P(|Y| > \epsilon) \leq \epsilon + P\left(\mathcal{E}\,(\,|Y|\,|\,X\,) > \epsilon^2\right).$$

PROOF.

$$\begin{aligned} P(|Y| > \epsilon) &= \mathcal{E}\,(\,P(|Y| > \epsilon \mid X)\,) \\ &\leq \mathcal{E}\,(\,\epsilon 1_{P(|Y|>\epsilon|X)\leq\epsilon} + 1_{P(|Y|>\epsilon|X)>\epsilon}\,) \\ &= \epsilon + P(P(|Y| > \epsilon \mid X) > \epsilon) \\ &\leq \epsilon + P\left(\frac{1}{\epsilon}\mathcal{E}\,(\,|Y|\,|\,X\,) > \epsilon\right) \\ &\leq \epsilon + P\left(\mathcal{E}\,(\,|Y|\,|\,X\,) > \epsilon^2\right) \end{aligned}$$

∎

**Lemma 3.4.6** Let $W_x$ be probability weights. Assume that $F_\bullet$ is $V_p$ continuous at $x$ and that $W_x \to \delta_x$ *Proh* in pr. Then

$$V_p(\hat{F}_x^x, \hat{F}_x) \to 0 \text{ in pr.}$$

if any of the conditions below hold:

(i) $V_p(F_{x_i}, F_x) < B$

(ii) $V_p(F_{x_i}, F_x) \leq M_x \max\{\|x - x_i\|, \|x - x_i\|^a\}$ and $W_x \to \delta_x$ $V_{ap}$ in pr.

(iii) $V_p(F_{x_i}, F_x) \leq \phi(x_i - x)$ and $\sum W_i \phi(x_i - x)^p \to 0$ in pr.

(iv) $V_\infty(W_x, \delta_x) \to 0$ in pr.

where $B > 0$, $p \geq 1$, $a \geq 1$, $M_x$ and $\alpha \in (0, 1)$ are constants.

PROOF. For any $\epsilon > 0$ there is a radius $\delta > 0$ such that

$$\|x_i - x\| < \delta \Rightarrow V_p(F_{x_i}, F_x) < \epsilon.$$

Let $S_\delta = \{v \in \mathcal{X} : \|v - x\| < \delta\}$. Denote by $W_x(S_\delta)$ the sum of the weights corresponding to $X_i \in S_\delta$. $\mathcal{E}\ (\ W_x(S_\delta)\ ) \to 1$ since $W_x(S_\delta)$ is uniformly bounded by 1 and converges to 1 in probability.

For case (i)

$$\begin{aligned}
\mathcal{E}\left(V_p(\hat{F}_x^x, \hat{F}_x)^p\right) &= \mathcal{E}\left(\sum W_i |Y_i^x - Y_i|^p\right) \\
&= \mathcal{E}\left(\sum W_i V_p(F_{X_i}, F_x)^p\right) \\
&\le \mathcal{E}\left(W_x(S_\delta)\epsilon^p + W_x(S_\delta^c)B^p\right) \\
&\to \epsilon^p
\end{aligned}$$

Therefore $V_p(\hat{F}_x^x, \hat{F}_x) \to 0$ in $L^p$ and hence also in pr.

For case (ii)

$$\begin{aligned}
P\left(V_p(\hat{F}_x^x, \hat{F}_x)^p > \epsilon\right) &\le \epsilon + P\left(\mathcal{E}\left(V_p(\hat{F}_x^x, \hat{F}_x)^p \mid X\right) > \epsilon^2\right) \\
&= \epsilon + P\left(\sum W_i \mathcal{E}\left(|Y_i^x - Y_i|^p \mid X\right) > \epsilon^2\right) \\
&= \epsilon + P\left(\sum W_i V_p(F_{X_i}, F_x)^p > \epsilon^2\right) \\
&\le \epsilon + P\left(\sum W_i M_x^p(\|X_i - x\|^p + \|X_i - x\|^{ap}) > \epsilon^2\right) \\
&= \epsilon + P\left(M_x^p(V_p(W_x, \delta_x)^p + V_{ap}(W_x, \delta_x)^{ap}) > \epsilon^2\right) \\
&\to \epsilon.
\end{aligned}$$

The proof of (iii) is essentially the same as the one for (ii).

For case (iv)

$$\begin{aligned}
P\left(V_p(\hat{F}_x^x, \hat{F}_x)^p > \epsilon\right) &\le P\Big(\sum_{x_i \in S_\delta} W_i |Y_i^x - Y_i|^p > \epsilon/2\Big) + P\Big(\sum_{x_i \notin S_\delta} W_i |Y_i^x - Y_i|^p > \epsilon/2\Big) \\
&\le P\Big(\sum_{x_i \in S_\delta} W_i |Y_i^x - Y_i|^p > \epsilon/2\Big) + P\Big(V_\infty(W_x, \delta_x) > \delta\Big) \\
&\le \frac{2}{\epsilon}\mathcal{E}\Big(\sum_{x_i \in S_\delta} W_i |Y_i^x - Y_i|^p\Big) + P\Big(V_\infty(W_x, \delta_x) > \delta\Big) \\
&\le \frac{2}{\epsilon}\epsilon^p + P\Big(V_\infty(W_x, \delta_x) > \delta\Big) \\
&\to 2\epsilon^{p-1}
\end{aligned}$$

Therefore $V_p(\hat{F}_x^x, \hat{F}_x) \to 0$ in pr. ∎

For strong convergence of the bias term there is the possibility of combining any of the strong laws for non-identically distributed random variables with any of the tradeoffs between regularity of $F_\bullet$ and convergence of $W_x$. Instead of producing a lemma with some twelve parts, we select part (iii) of Lemma 3.4.2′, and strong versions of parts (i) and (iv) of Lemma 3.4.6.

**Lemma 3.4.7** Let $W_x$ be probability weights. Assume that $F_\bullet$ is $V_p$ continuous, that $W_x \to \delta_x$ *Proh* a.s., $n_x \geq Bn^{1-\alpha\lambda}$ a.s. and $\max|W_i| \leq Bn^{-\alpha}$ a.s..

Then

$$V_p(\hat{F}_x^x, \hat{F}_x) \to 0 \text{ a.s.}$$

if either of the following hold:

(i) $V_{p\gamma}(F_{x_i}, F_x) \leq B$

(ii) $F_\bullet$ is $V_{p\gamma}$ continuous at $x$ and $V_\infty(W_x, \delta_x) \to 0$ a.s.

where $B > 0$, $p \geq 1$, $\lambda > 0$, $\gamma > 2 + \lambda$ and $\alpha \in (0,1)$ are constants.

**Remark** The variable $\gamma$ is introduced to simplify the exposition. A uniform bound on $\mathcal{E}\left(|D_k|^{p\gamma}\right)$ implies a uniform bound on $\mathcal{E}\left((|D_k|^p)^{2+\lambda}\log^+(|D_k|^p)^{1+\eta}\right)$ for any $\eta > 0$. The latter condition is the one used in Lemma 3.4.1′.

PROOF. For any $\epsilon > 0$ there is a radius $\delta > 0$ such that

$$\|x_i - x\| < \delta \Rightarrow V_p(F_{x_i}, F_x) < \epsilon. \tag{14}$$

When $F_\bullet$ is $V_{p\gamma}$ continuous at $x$ there is a radius $\delta$ such that

$$\|x_i - x\| < \delta \Rightarrow V_{p\gamma}(F_{x_i}, F_x) < \epsilon. \tag{15}$$

Let $S_\delta = \{v \in \mathcal{X} : \|v - x\| \leq \delta\}$. Denote by $W_x(S_\delta)$ the sum of the weights corresponding to $X_i \in S_\delta$. $W_x(S_\delta) \to 1$ a.s.

In either case condition on $X$ values that satisfy the a.s. conditions on the $W_i$. Strong conditional convergence is sufficient by Fubini's theorem.

For case (i), pick $\delta > 0$ to satisfy (14). Then

$$\begin{aligned} V_p^p(\hat{F}_x^x, \hat{F}_x) &= \sum W_i |Y_i^x - Y_i|^p \\ &= \sum W_i (|Y_i^x - Y_i|^p - V_p^p(F_{x_i}, F_x)) \\ &+ \sum_{x_i \in S_\delta} W_i V_p^p(F_{x_i}, F_x) \\ &+ \sum_{x_i \notin S_\delta} W_i V_p^p(F_{x_i}, F_x) \end{aligned}$$

the first term of which coverges to zero a.s. by Lemma 3.4.1′(iii). The second term is bounded by $\epsilon^p$ and the third by $B^p W_x(S_\delta^c) \to 0$.

In (ii) pick $\delta$ to satisfy (15). Then

$$\begin{aligned} V_p^p(\hat{F}_x^x, \hat{F}_x) &= \sum_{x_i \in S_\delta} W_i (|Y_i^x - Y_i|^p - V_p^p(F_{x_i}, F_x)) \\ &+ \sum_{x_i \in S_\delta} W_i V_p^p(F_{x_i}, F_x) \\ &+ \sum_{x_i \notin S_\delta} W_i |Y_i^x - Y_i|^p \end{aligned}$$

the last term of which is eventually zero with probability 1. The second term is bounded by $\epsilon^p$, and the first term satisfies the conditions of Lemma 3.4.1′(iii). ∎

The results for weak convergence may be summarized as follows:

**Theorem 3.4.1** Suppose for some finite $p \geq 1$, that $F_\bullet$ is $V_p$ continuous at $x$ and that $W_x$ is obtained from probability weights with

$$W_x \to \delta_x Proh \text{ in pr.} \quad \text{and} \quad n_x \to \infty \text{ in pr.}$$

Then

$$\hat{F}_x \to F_x \; V_p \text{ in pr.}$$

under any of the conditions below:

(i) $V_p(F_{x_i}, F_x) < B$

(ii) $V_p(F_{x_i}, F_x) \leq M_x \max\{\|x - x_i\|, \|x - x_i\|^a\}$ and $W_x \to \delta_x \; V_{ap}$ in pr.

(iii) $V_p(F_{x_i}, F_x) \leq \phi(x_i - x)$ and $\sum W_i \phi(x_i - x)^p \to 0$ in pr.

(iv) $V_\infty(W_x, \delta_x) \to 0$ in pr.

where $B > 0$, $M_x > 0$ and $a \geq 1$ are constants and $\phi$ is a nonnegative real function.

PROOF. By the triangle inequality

$$V_p(\hat{F}_x, F_x) \leq V_p(\hat{F}_x, \hat{F}_x^x) + V_p(\hat{F}_x^x, F_x) \tag{16}$$

where $\hat{F}_x^x$ is defined by (1). By Lemma 3.4.6, $V_p(\hat{F}_x^x, \hat{F}_x) \to 0$ in pr. and from $n_x \to \infty$ follows $Proh(\hat{F}_x^x, F_x) \to 0$ in pr. (Apply Theorem 3.2.2 with every $x_i = x$.) Also by Lemma 3.4.3, $\sum W_i |Y_i^x|^p \to \mu_p(x)$ in pr. Therefore $V_p(\hat{F}_x^x, F_x) \to 0$ in pr. ∎

The results for strong convergence may be summarized as follows:

**Theorem 3.4.2** Suppose for some finite $p \geq 1$, that $F_\bullet$ is $V_p$ continuous at $x$, that $F_x$ has a finite $2+\lambda$'th absolute moment for some $\lambda > 0$ and that $W_x$ is obtained from probability weights with

$$W_x \to \delta_x \; Proh \text{ a.s.} \quad \text{and} \quad n_x \geq Bn^{1-\alpha\lambda} \text{ a.s.} |W_i| \leq Bn^{-\alpha} \text{ a.s.}$$

for $\alpha \in (0, 1)$. Then

$$\hat{F}_x \to F_x \; V_p \text{ a.s.}$$

under either of the conditions below:

(i) $V_{p\gamma}(F_{x_i}, F_x) \leq B$

(ii) $F_\bullet$ is $V_{p\gamma}$ continuous at $x$ and $V_\infty(W_x, \delta_x) \to 0$ a.s.

where $\gamma > 2 + \lambda$.

PROOF. Decompose $V_p(\hat{F}_x, F_x)$ into bias and variance components as in Theorem 3.4.1. The bias term $V_p(\hat{F}_x^x, \hat{F}_x) \to 0$ a.s. by Lemma 3.4.7. By Lemma 3.4.4 $\sum W_i |Y_i^x|^p \to \mu_p(x)$ a.s. using condition (iii) of Lemma 3.4.1 and the independence of the $Y_i^x$ and the $W_i$. Also $Proh(\hat{F}_x^x, F_x) \to 0$ a.s. by Theorem 3.2.2, so that the variance term $V_p(\hat{F}_x^x, F_x) \to 0$ a.s. ∎

# 4 Asymptotic Normality

## 4.1 Introduction

In Chapter 3, weak and strong consistency of running functionals was obtained. In this chapter, many running functionals turn out to be asymptotically normal. As for the estimate $\hat{F}_x$, it converged to $F_x$ weakly or strongly (depending on the strength of the conditions) in several metrics, in Chapter 3. In this chapter conditions are given under which the normalized difference $\sqrt{n_x}(\hat{F}_x - F_x)$ converges weakly to a Brownian bridge. Unifying features of the two chapters are that the same bias-variance split is used and the effective sample size $n_x$ plays a role analogous to that played by $n$ in the i.i.d. setup. The result is to refine the notion that estimation at $x$ is like that based on a biased sample of size $n_x$ from $F_x$.

The development is as follows: The estimated regression function is split into bias and variance terms. Sec. 4.2 develops necessary and sufficient conditions for the variance term to have a normal limit. A multivariate central limit theorem follows immediately by the Cramer-Wold device. Sec. 4.3 provides conditions under which the bias term goes to zero fast enough that the regression itself is asymptotically normal. The variance term of $\hat{F}_x$ converges weakly to a Brownian bridge under conditions given in Sec. 4.4, and under further conditions the bias term converges to zero. Von Mises method and the theory of compact differentiability prove asymptotic normality for a class of running functionals in Sec. 4.5.

The bias variance split for $\hat{F}_x$ is

$$\hat{F}_x - F_x = (\hat{F}_x^x - F_x) + (\hat{F}_x - \hat{F}_x^x) \tag{1}$$

where $\hat{F}_x^x$ is obtained by substituting $Y_i^x$ for $Y_i$ in $\hat{F}_x$ and the split for a functional $T(\cdot)$ is

$$T(\hat{F}_x) - T(F_x) = (T(\hat{F}_x^x) - T(F_x)) + (T(\hat{F}_x) - T(\hat{F}_x^x)) \tag{2}$$

which for the conditional expectation becomes

$$\hat{m}(x) - m(x) = \sum W_i(Y_i^x - m(x)) + \sum W_i(Y_i - Y_i^x). \tag{3}$$

The second term is named after the bias because it is nonzero due to the discrepancy between $F_x$ and $F_{X_i}$ and the first term is named after the variance because it is nonzero due to sampling variation from $F_x$.

## 4.2 Asymptotic Normality of the Regression Variance

The variance term in 4.1.3 is a weighted sum of centered $Y_i^x$s. The quantities $Y_i^x - m(x)$ are i.i.d. with mean 0, and we will assume, a finite variance. There is no essential difference in the treatment of $Y_i^x$ and $h(Y_i^x)$ provided $h(Y_i^x)$ satisfies the moment conditions. Therefore it will make the notation clearer to replace $Y_i^x$ or $h(Y_i^x)$ by $V_i$ where the $V_i$ are i.i.d. and have first and second moments. By construction (Sec. 2.1) the $Y_i^x$'s are independent of the $X_i$'s and hence of the $W_i$'s.

**Lemma 4.2.1** Let $W_{ni}$, $1 \le i \le n < \infty$ be a triangular array of real constants with

$$\sum_{i=1}^{n} W_{ni} = 1,$$

and set

$$n_x = n_x(n) = \left( \sum_{i=1}^{n} W_{ni}^2 \right)^{-1}.$$

Let $V_i$ be i.i.d. from a distribution $F$ with mean $\mu$ and positive variance $\sigma^2 < \infty$. Then

$$Z_n = \sqrt{n_x} \left( \sum_{i=1}^{n} W_{ni} V_i - \mu \right) \xrightarrow{D} N(0, \sigma^2)$$

for any such $F$ iff

$$n_x \to \infty \tag{1}$$

and

$$\max_{1\le i\le n}\sqrt{n_x}|W_{ni}| \to 0. \tag{2}$$

PROOF. Necessity of (1) and of (2) is trivial. For sufficiency there is no loss of generality in taking $\mu = 0$ and $\sigma^2 = 1$. To conform with our usual notation, abbreviate $W_{ni}$ to $W_i$.

The proof begins by applying the Lindeberg theorem (Billingsley 1979, Theorem 27.2) to the double array with $n, i$ element $\sqrt{n_x}W_iV_i$. We need only establish Lindeberg's condition which here amounts to showing

$$\sum_{i=1}^{n}\int_{|\sqrt{n_x}W_iV_i|>\eta} n_xW_i^2V_i^2dF \to 0 \tag{3}$$

for any $\eta > 0$.

Put $W = \max|W_i|$ in each row of the table. Then the sum in (3) does not exceed

$$\begin{aligned}&\sum_{i=1}^{n} n_xW_i^2\int_{|\sqrt{n_x}WV_i|>\eta} V_i^2dF\\ &=\int_{|\sqrt{n_x}WV_1|>\eta} V_1^2dF\\ &\le\int_{|V_1|>\eta\sqrt{\lfloor(W^2n_x)^{-1}\rfloor}} V_1^2dF\end{aligned} \tag{4}$$

where $\lfloor z\rfloor$ denotes the largest integer less than or equal to $z$.

The sequence in (4) tends to zero in pr. if

$$\int_{|V_1|>\eta\sqrt{n}} V_1^2dF \to 0. \tag{5}$$

Note that (5) is the Lindeberg condition for $\sqrt{n}$ times the sample average of $n$ i.i.d. $V_i$ which has a normal limit. Since $W\sqrt{n_x} \to 0$

$$\lim_{n\to\infty}\max_{1\le i\le n} P(|\sqrt{n_x}W_iV_i| > \epsilon) = 0 \tag{6}$$

for any $\epsilon > 0$. Then (6) and Feller's theorem (Billingsley, 1979, Theorem 27.4) together imply (5). ∎

Lemma 4.2.1 can also be proved using characteristic functions. Because the $V_i$ are i.i.d. the "little o" terms all come from the same Taylor approximation and so their sum is easy to manage.

**Corollary** The condition $\sum_{i=1}^{n} W_{ni} = 1$ may be replaced by

$$\sqrt{n_x}(1 - \sum_{i=1}^{n} W_{ni}) \to 0 \tag{7}$$

in Lemma 4.2.1.

PROOF. Immediate.

Since $W_x$ is random, it is essential to extend the conditions of Lemm 4.2.1.

**Lemma 4.2.2** Let $W_{ni}$, $1 \leq i \leq n < \infty$ be a triangular array of real random variables, and set

$$n_x = n_x(n) = \left( \sum_{i=1}^{n} W_{ni}^2 \right)^{-1}.$$

Let $V_i$ be i.i.d. from a distribution $F$ with mean $\mu$ and positive variance $\sigma^2 < \infty$. Also assume that the $V_i$ are independent of the $W_{ni}$. Then

$$Z_n = \sqrt{n_x} \left( \sum_{i=1}^{n} W_{ni} V_i - \mu \right) \xrightarrow{\mathrm{D}} N(0, \sigma^2)$$

if all of the following hold:

$$n_x \to \infty \text{ in pr.} \tag{8a}$$

$$\max_{1 \leq i \leq n} \sqrt{n_x} |W_{ni}| \to 0 \text{ in pr.} \tag{8b}$$

$$\sqrt{n_x}(1 - \sum_{i=1}^{n} W_{ni}) \to 0 \text{ in pr.} \tag{8c}$$

**Remark** For probability weights (8a) implies (8bc).

PROOF. As before abbreviate $W_{ni}$ by $W_i$. Make the split

$$Z_n = \sqrt{n_x} \sum W_i (V_i - \mu) - \sqrt{n_x}(1 - \sum W_i)\mu. \tag{9}$$

The second term in (9) tends to zero in pr. by (8c). We may assume that $\sum W_i > 0$ by (8ac) and so dividing each $W_i$ by $\sum W_i$ yields weights that sum to 1 without changing the first term in (9). Therefore we may assume that $\sum W_i = 1$.

Let $z \in I\!R$ and $\epsilon > 0$. If the $W_i$ were fixed, then by Lemma 4.2.1 there would exist $\delta > 0$ such that $n_x > 1/\delta$ and $\sqrt{n_x}\max|W_i| < \delta$ imply that $|P(Z_n < z) - \Phi(z)| < \epsilon$ where $\Phi$ is the standard normal distribution function. But by independence of the $W_i$ and $V_i$, the conditional distribution of $Z_n$ given values of the $W_i$'s is exactly what it would be for fixed $W_i$'s taking those values. Therefore

$$\begin{aligned}|P(Z_n < z) - \Phi(z)| &\le \mathcal{E}\,|P(Z_n < z \,|\, W_1, \ldots, W_n) - \Phi(z)| \\ &\le \epsilon + P(n_x \le 1/\delta) + P(\sqrt{n_x}\max|W_i| \ge \delta) \\ &\to \epsilon.\end{aligned}$$

Therefore $P(Z_n < z) \to \Phi(z)$. ∎

Lemma 4.2.2 extends to a multivariate central limit theorem as follows:

**Lemma 4.2.3** Let $V_i$ be i.i.d. random vectors of length $p$ with mean $\mu$ and variance-covariance matrix $\Sigma$. Let $W_{ni}$ satisfy (8abc). Assume that the $V_i$ are independent of the $W_i$. Then

$$Z_n \stackrel{\text{def}}{=} \sqrt{n_x}\Big(\sum W_i V_i - \mu\Big) \xrightarrow{\text{D}} N_p(0, \Sigma).$$

PROOF. Let $l$ be any fixed p-vector. The asymptotic distribution of $l \cdot Z_n$ is normal with limiting first two moments 0 and $l\Sigma l'$ by Lemma 4.4.2. Since this holds for any $l$ the asymptotic distribution of $Z_n$ is multivariate normal with mean 0 and variance-covariance $\Sigma$. (See Rao 1973, 2c.5iv). ∎

## 4.3 Asymptotic Negligibility of the Regression Bias

In this section we provide conditions under which

$$\sqrt{n_x}\sum W_i(Y_i - Y_i^x) \xrightarrow{\text{D}} 0.$$

With the factor $\sqrt{n_x}$, the variance term converges to a normal distribution with mean 0 and variance $\mathcal{E}\,(y - m(x))^2$. To make the bias converge, we require $W_x$ to converge to $\delta_x$

in some sense. For $W_x \to \delta_x$ to imply that the bias disappears it is necessary to suppose that when $x_i$ is close to $x$, that $F_{x_i}$ is suitably close to $F_x$. Typically one assumes that the regression curve admits so many continuous derivatives and applies a Taylor expansion. Here, that condition is replaced by an assumption that

$$V_1(F_{x_i}, F_x) \leq M_x \|x_i - x\|$$

at least for $x_i$ close enough to $x$. In the presence of Prohorov continuity of $F_\bullet$ the condition above is weaker than the existence of a derivative of $m(x)$. To make the normalized bias converge, it will be necessary to have $W_x$ converge to $\delta_x$ *faster* in some sense than $n_x$ is going to infinity. In practice, one usually tolerates some asymptotic bias, in order to obtain a lower mean square error.

**Lemma 4.3.1** If $F_\bullet$ satisfies

$$V_1(F_{x_i}, F_x) \leq M_x \|x_i - x\| \tag{1}$$

and

$$\mathcal{E}\left(\sqrt{n_x} V_1(W_x, \delta_x)\right) \to 0 \tag{2}$$

then

$$\sqrt{n_x} \sum W_i (Y_i - Y_i^x) \to 0 \ L^1.$$

PROOF.

$$\begin{aligned}
\mathcal{E}\left(\left|\sqrt{n_x} \sum W_i (Y_i - Y_i^x)\right|\right) &\leq \mathcal{E}\left(\sqrt{n_x} \sum |W_i| |Y_i - Y_i^x|\right) \\
&= \mathcal{E}\left(\sqrt{n_x} \sum |W_i| V_1(F_{x_i}, F_x)\right) \\
&\leq \mathcal{E}\left(\sqrt{n_x} \sum |W_i| M_x \|x_i - x\|\right) \\
&= \mathcal{E}\left(\sqrt{n_x} M_x V_1(W_x, \delta_x)\right) \\
&\to 0.
\end{aligned}$$

∎

Condition (2) says that the weighted average absolute distance of the observations used to estimate the regression from the target point must go to zero faster than the

reciprocal of the square root of the effective sample size. For $k$-NN, the $k$'th neighbor should be at distance $o(1/k)$ from $x$. In the sampling case the $k$'th neighbor is usually at distance $O_p(k/n)$ from the target point. Because condition (2) involves the expectation of $V_1(W_x, \delta_x)$ it may be awkward when the $X_i$ are sampled from a long-tailed distribution. For the next lemma (2) is weakened to convergence in pr., and the conclusion is correspondingly weaker, but is enough to give the regression an asymptotically normal distribution.

**Lemma 4.3.2** If $F_\bullet$ satisfies

$$V_1(F_{x_i}, F_x) \leq M_x \|x_i - x\|$$

and if

$$\sqrt{n_x} V_1(W_x, \delta_x) \to 0 \text{ in pr.} \tag{3}$$

then

$$\sqrt{n_x} \sum W_i (Y_i - Y_i^x) \to 0 \text{ in pr.}$$

PROOF. Let $\epsilon > 0$, and put

$$B = |\sqrt{n_x} \sum W_i (Y_i - Y_i^x)|.$$

Then, using $X$ to denote the sequence of $X_i$'s, and recalling Lemma 3.4.5:

$$\begin{aligned}
P(B > \epsilon) &\leq \epsilon + P\left(\mathcal{E}(B \mid X) > \epsilon^2\right) \\
&= \epsilon + P\left(\mathcal{E}(\sqrt{n_x} \sum |W_i| |Y_i - Y_i^x| \mid X) > \epsilon^2\right) \\
&= \epsilon + P\left(\sqrt{n_x} \sum |W_i| \mathcal{E}(|Y_i - Y_i^x| \mid X) > \epsilon^2\right) \\
&= \epsilon + P\left(\sqrt{n_x} \sum |W_i| V_1(F_{X_i}, F_x) > \epsilon^2\right) \\
&\leq \epsilon + P\left(\sqrt{n_x} \sum |W_i| M_x \|X_i - x\| > \epsilon^2\right) \\
&= \epsilon + P\left(\sqrt{n_x} M_x V_1(W_x, \delta_x) > \epsilon^2\right) \\
&\to \epsilon
\end{aligned}$$

∎

Condition (2) is that the area between the distribution curves is locally Lipschitz. This is a mild short range condition, but it does have long range consequences. Most authors handle the long range problem by either working in a compact set, or by using a $W_x$ that has $V_\infty$ convergence to $\delta_x$. (Examples are kernels with bounded support, and k-NN schemes.) With $V_\infty$ convergence of $W_x$ it is only necessary to assume that $V_1(F_{x_i}, F_x) \le M_x \|x_i - x\|$ for sufficiently small $\|x_i - x\|$. With compact $\mathcal{X}$ and Prohorov continuous $F_\bullet$, continuity of $m(\cdot)$ implies condition (2).

**Lemma 4.3.3** For some positive $D < \infty$, suppose $F_\bullet$ satisfies

$$V_1(F_{x_i}, F_x) \le M_x \|x_i - x\| \quad \text{whenever} \quad \|x_i - x\| \le D.$$

Assume that $P(n_x \ge 1) \to 1$ and

$$\sqrt{n_x} V_\infty(W_x, \delta_x) \to 0 \text{ in pr.}, \tag{3}$$

and for some positive $E < \infty$

$$P\left(\sum |W_i| \ge E\right) \to 0. \tag{4}$$

Then

$$\sqrt{n_x} \sum W_i (Y_i - Y_i^x) \xrightarrow{D} 0.$$

PROOF. Let $\epsilon > 0$ and define

$$H = \{n_x \ge 1\} \cap \{\sum |W_i| < E\} \cap \{V_\infty(W_x, \delta_x) < D\}$$

and

$$B = |\sqrt{n_x} \sum W_i (Y_i - Y_i^x)|.$$

Then

$$\begin{aligned} P(B > \epsilon) &\le P(B 1_H > \epsilon) + P(H^c) \\ &\le \epsilon + P\left(\mathcal{E}(B 1_H \mid X) > \epsilon^2\right) + P(H^c) \end{aligned}$$

$$\leq \epsilon + P\left(\sqrt{n_x} M_x E V_\infty(W_x, \delta_x) > \epsilon^2\right) + P(H^c)$$

$$\to \epsilon + P(H^c)$$

$$\leq \epsilon + P(\sum |W_i| \geq E) + P(n_x < 1 \text{ or } V_\infty(W_x, \delta_x) > D)$$

$$\to \epsilon + P(n_x < 1 \text{ or } V_\infty(W_x, \delta_x) > D)$$

$$\leq \epsilon + P(n_x < 1) + P(V_\infty(W_x, \delta_x) > D \ \& \ n_x \geq 1)$$

$$\to \epsilon + P(V_\infty(W_x, \delta_x) > D \ \& \ n_x \geq 1)$$

$$\leq \epsilon + P(\sqrt{n_x} V_\infty(W_x, \delta_x) > D \ \& \ n_x \geq 1)$$

$$\leq \epsilon + P(\sqrt{n_x} V_\infty(W_x, \delta_x) > D)$$

$$\to \epsilon \qquad \blacksquare$$

Condition (4) is introduced because of the way $V_\infty(W_x, \delta_x)$ is extended to finite signed measures $W_x$ in Subsec. 2.4.3. For bias elimination, very light conditions are placed on the sequence $n_x$. For example 1-NN schemes, in which the closest neighbor to $x$ gets unit weight and all other observations get 0 weight satisfy the lemmas above. The condition governing $\sum |W_i|$ is important in the bias considerations, but was not needed to handle the variance term in Sec. 4.2.

Now, combining the results of this section and Sec. 4.2:

**Theorem 4.3.1** If for some positive $B < \infty$ $W_x$ satisfies:

$$n_x \to \infty \text{ in pr.} \tag{5a}$$

$$\max_{1 \leq i \leq n} \sqrt{n_x} |W_i| \to 0 \text{ in pr.} \tag{5b}$$

$$\sqrt{n_x}(1 - \sum_{i=1}^{n} W_i) \to 0 \text{ in pr.} \tag{5c}$$

$$\sqrt{n_x} V_1(W_x, \delta_x) \to 0 \text{ in pr.} \tag{5d}$$

$$P(\sum |W_i| < E) \to 1 \tag{5e}$$

and for some positive $M_x < \infty$ $F_\bullet$ satisfies:

$$0 < \sigma_x^2 = \int (y - m(F_x))^2 < \infty \tag{6a}$$

$$V_1(F_{x_i}, F_x) \leq M_x \|x_i - x\| \tag{6b}$$

then

$$\sqrt{n_x}(m(\hat{F}_x) - m(F_x)) \xrightarrow{D} N(0, \sigma_x^2). \tag{7}$$

If (6b) only holds for $\|x_i - x\| \leq D < \infty$ and (5d) is strengthened to

$$\sqrt{n_x} V_\infty(W_x, \delta_x) \to 0 \text{ in pr.} \tag{8}$$

then (7) holds.

PROOF. Write

$$\begin{aligned}\sqrt{n_x}(m(\hat{F}_x) - m(F_x)) &= \sqrt{n_x}\left(\sum W_i Y_i - m(F_x)\right) \\ &= \sqrt{n_x}\left(\sum W_i Y_i^x - m(F_x)\right) + \sqrt{n_x}\sum W_i(Y_i - Y_i^x). \end{aligned} \tag{9}$$

The first term in (9) tends in distribution to $N(0, \sigma_x^2)$ by Lemma 4.2.2, because the $Y_i^x$ are independent of the $W_i$, and because of (5abc). Under (5de) and (6ab) the second term in (9) converges to 0 in pr. by Lemma 4.3.2. If (6b) only holds locally, but (8) holds, then the second term in (9) converges to 0 in pr. by Lemma 4.3.3. (Note that (5a) implies $P(n_x \geq 1) \to 1$.) ∎

Schuster (1972) obtains asymptotic joint normality of the regression function at a finite number of points, for kernel regressions. The regression values at distinct points are asymptotically independent. Royall (1966) obtains asymptotic normality for nearest neighbor methods. Stute (1984) obtains asymptotic normality for symmetric nearest neighbor methods with a bounded kernel. Where Schuster assumes a finite third moment for $Y$, Stute needs only a finite second moment.

## 4.4 Asymptotic Distribution of $\sqrt{n_x}(\hat{F}_x - F_x)$

This section shows that the conditional empirical process $\hat{F}_x - F_x$ has a Brownian limit when normalized by $\sqrt{n_x}$ under very general conditions on the weights.

Start by making the split

$$\hat{F}_x - F_x = (F_x - \hat{F}_x^x) + (\hat{F}_x^x - \hat{F}_x)$$

where $\hat{F}_x^x$ is obtained by replacing each $Y_i$ in $\hat{F}_x$ by $Y_i^*$. Recall that $Y_i^*$ are i.i.d. from $F_x$. The first term above is the variance term and the second is the bias.

The goal is to make the variance term normalized by $\sqrt{n_x}$ converge to a Gaussian process and to make the normalized bias term converge weakly to zero. The Gaussian process is supposed to be the Brownian bridge when $F_x$ is absolutely continuous. In the absolutely continuous case it is sufficient to consider $F_x = U[0,1]$. That is assume

$$F_x(t) \stackrel{\text{def}}{=} P(Y \leq t | X = x) = t.$$

The value of the variance process at $t$ is then

$$\begin{aligned} Z_n(t) &\stackrel{\text{def}}{=} \sqrt{n_x} \sum W_i (1_{Y_i^x \leq t} - t) \\ &= \sqrt{n_x} \sum W_i (1_{U_i \leq t} - t) \end{aligned}$$

where $U_i$ are i.i.d. $U[0,1]$. To accomodate conflicting conventions the sign of the variance term has been reversed.

Let $0 \leq t_1 < \ldots < t_k \leq 1$ for some finite $k$. The vector $V_n = (Z_n(t_1), \ldots, Z_n(t_k))'$ has mean zero and for $i < j$ the $i,j$ element of its variance covariance matrix is $t_i(1 - t_j)$. Thus it has the same first two moments as the Brownian bridge process.

**Lemma 4.3.1** If $F_x$ is uniform[0,1] and $\hat{F}_x$ is obtained by weights satisfying (4.2.8abc) then the finite dimensional distributions of $\sqrt{n_x}(\hat{F}_x^x - F_x)$ converge to those of the Brownian bridge.

PROOF. Apply Lemma 4.2.3.

In addition to the convergence of the finite dimensional distributions to those of the Brownian bridge, it is also necessary to govern the behavior of the process over small intervals. This is usually done by proving uniform tightness of the sequence of processes. We will instead use a similar approach from Pollard (1984, Chapter V). Consider $Z_n$ as a member of $D[0,1]$, the space of real valued functions defined on $[0,1]$ that are continuous from the right and have limits from the left. Such functions are sometimes called cadlag

functions, from the French: ***continue a droit, limites a gauche.*** Equip the space $D[0,1]$ with the uniform metric $d(Z,W) = \sup_{0\le x\le 1}|Z(x) - W(x)|$ and the projection $\sigma$-field. The projection $\sigma$-field differs from the usual Borel $\sigma$-field in that empirical distribution functions are measureable. The former is generated by all closed balls, the latter by all closed subsets. The trace of the projection $\sigma$-field on $C[0,1]$, the space of continuous functions on $[0,1]$, coincides with the Borel $\sigma$-field of $C[0,1]$. For a detailed discussion of this approach see Pollard (1984). His Theorem V.3 is the main result. It is:

**Theorem 4.4.1** Let $Z, Z_1, Z_2, \ldots$ be random elements of $D[0,1]$ under the uniform metric and the projection $\sigma$-field. Suppose $P\{Z \in C\} = 1$ for some separable subset $C$ of $D[0,1]$. The necessary and sufficient conditions for $\{Z_n\}$ to converge in distribution to $Z$ are:

(i) the finite dimensional distributions of $Z_n$ converge to those of $Z$

(ii) to each $\epsilon > 0$ and $\delta > 0$ there corresponds a grid $0 = t_0 < t_1 < \ldots < t_m = 1$ such that

$$\limsup P\{\max_{i=0}^{m-1} \sup_{t\in[t_i,t_{i+1})} |Z_n(t) - Z_n(t_i)| > \delta\} < \epsilon. \tag{1}$$

PROOF. Pollard (1984, pp. 92-3).

When $Z$ is the Brownian bridge $C$ can be taken to be $C[0,1]$.

**Definition** A sequence $Z_n$ of random elements in $D[0,1]$ under the uniform metric and projection $\sigma$-field is *nearly tight* if condition (ii) of Theorem 4.4.1 holds. The property of being nearly tight will be called *near tightness.*

A uniformly tight sequence is nearly tight. A nearly tight sequence need not be uniformly tight. For example Pollard (1984) shows that $F_n$ is nearly tight, but $F_n$ is not uniformly tight. See Fernholz (1983, p. 28) for a characterization of the bounded sets of $D[0,1]$ that have compact closure.

To establish weak convergence of the variance term to the Brownian bridge it only remains to prove near tightness of $\sqrt{n_x}(\hat{F}_x^x - F_x)$.

In certain special cases the convergence to the Brownian bridge is very easy to show. We can borrow from the i.i.d. case in which the Brownian limit is well known, in the same way that Lemma 4.2.1 borrows from the i.i.d. central limit theorem via Feller's theorem. Perhaps the simplest is the following, which includes $k$ nearest neighbor smoothers, symmetric $k$ nearest neighbor smoothers and one sided nearest neighbor smoothers.

**Theorem 4.4.2** If $F_x$ is uniform[0,1] and $\hat{F}_x$ is obtained by weights of which

$$k = k(n) \to \infty \text{ in pr.}$$

are $1/k$ and the rest are 0, then

$$Z_n = \sqrt{n_x}(\hat{F}_x - F_x) \xrightarrow{\mathrm{D}} B$$

where $B$ is the Brownian bridge.

PROOF. We have $\sum W_i = 1$, $n_x = k$ and $\sqrt{n_x}\max|W_i| = 1/\sqrt{k}$ so by Lemma 4.4.1 the finite dimensional distributions of $Z_n$ converge to those of the Brownian bridge.

Let $\epsilon > 0$ and $\delta > 0$. It is well known that the desired weak convergence holds when $k = n$. Let $\tilde{Z}_n$ be the sequence of processes obtained by taking $W_i = 1/n$ in the expression for $Z_n$ and by taking $\sqrt{n}$ for $\sqrt{n_x}$. By the necessity part of Theorem 4.4.1 there is a grid $0 = t_0 < t_1 < \ldots < t_m = 1$ such that (1) holds for $\tilde{Z}_n$.

Now

$$\begin{aligned}
&\limsup_{n\to\infty} P\{\max_{i=1}^{m-1} \sup_{t\in[t_i,t_{i+1})} |Z_n(t) - Z_n(t_i)| > \delta\} \\
&= \limsup_{n\to\infty} P\{\max_{i=1}^{m-1} \sup_{t\in[t_i,t_{i+1})} |\tilde{Z}_{k_n}(t) - \tilde{Z}_{k_n}(t_i)| > \delta\} \\
&\le \limsup_{n\to\infty} P\{\max_{i=1}^{m-1} \sup_{t\in[t_i,t_{i+1})} |\tilde{Z}_n(t) - \tilde{Z}_n(t_i)| > \delta\} \\
&< \epsilon.
\end{aligned}$$

The equality above is due to $Z_n$ having the same distribution as $\tilde{Z}_{k_n}$, the first inequality follows because the limit supremum of a sequence is no less than that of any subsequence, and the last inequality is by construction. ∎

The approach above generalizes to some schemes in which a finite number of weight levels are used. For example, suppose weight $2/(3k)$ is put on each of the $k$ nearest neighbors and weight $1/(3k)$ is put on each of the next $k$ closest neighbors. The process $Z_n$ is then a sum of two processes, one for the nearest neighbors and another for the second group. Each term in the sum, converges to a constant times a Brownian bridge. For each $n$ the terms are independent. It follows that their sum converges to a constant times a Brownian bridge and the normalization $\sqrt{n_x}$ is such that the standard Brownian bridge would be the result. Proceding from 2 to $L$ levels is straightforward and many interesting weight schemes can be approximated this way for large $L$.

A large class of weighting schemes might be shown to have variance terms which converge to the Brownian bridge by arguments based on approximating this way and showing that the differences between the approximate and actual processes are asymptotically negligible. Instead, an argument that parallels the development of the functional central limit theorem in Pollard (1984, Chapter V) is given below.

The key step in the derivation is to bound the probability that the supremum of $|Z_n(t)|$ over a short interval exceeds a constant $\delta$ by a probability based only on the difference between the endpoints of the interval. This is accomplished by the following lemma, which Pollard gives as Lemma V.7:

**Lemma 4.4.2** Let $\{Z(t) : 0 \le t \le 1\}$ be a process with cadlag sample paths taking the value zero at $t = 0$. Suppose $Z(t)$ is $\mathcal{E}_t$-measureable, for some increasing family of $\sigma$-fields $\{\mathcal{E}_t : 0 \le t \le b\}$. If at each point of $\{|Z(t)| > \delta\}$,

$$P\left( |Z(b) - Z(t)| \le \frac{1}{2}|Z(t)| \,\middle|\, \mathcal{E}_t \right) \ge \beta,$$

where $\beta$ is a positive number depending only on $\delta$, then

$$P\left( \sup_{0 \le t \le b} |Z(t)| > \delta \right) \le \beta^{-1} P\left( |Z(b)| > \beta/2 \right).$$

PROOF. Pollard (1984, pp. 94–5).

**Theorem 4.4.3** If $F_x$ is uniform[0,1] and $\hat{F}_x$ is obtained from $W_i$ satisfying (4.2.8abc) then

$$Z_n = \sqrt{n_x}(\hat{F}_x^x - F_x) \xrightarrow{D} B_0$$

where $B_0$ is the Brownian bridge process.

PROOF. By Lemma 4.4.1 the finite dimensional distributions of $Z_n(t)$ converge to those of $B_0$, and so by Theorem 4.4.1 it only remains to establish (1).

It suffices to show near tightness of $Z_n$ for fixed $W_i$ satisfying

$$n_x \to \infty, \quad \sum W_i = 1 \quad \text{and} \quad \sqrt{n_x} \max |W_i| \to 0$$

because then (1) follows for random $W_i$ satisfying (4.2.8abc) by the technique of Lemma 4.2.2.

Let $\epsilon > 0$ and $\delta > 0$. With the $W_i$ fixed,

$$W(t) \stackrel{\text{def}}{=} \sum W_i 1_{U_i \le t}$$

is a nondecreasing process with cadlag sample paths. $W(0) = 0$ and $W(1) = 1$ and $W(\cdot)$ jumps by the fixed amount $W_i$ at the random place $U_i$. Since the $U_i$ are independent uniform[0,1] the process $W(t+s) - W(s)$ on $0 \le t \le b \le 1 - s$ has the same distribution as $W(t)$ on $0 \le t \le b$. Note that $Z_n(t) = \sqrt{n_x}(W(t) - t)$ so it also has this stationarity property.

When $t_i = i/m$, (1) reduces to

$$\begin{aligned}
&\limsup P\{\max_{i=0}^{m-1} \sup_{t \in [i/m,(i+1)/m)} |Z_n(t) - Z_n(t_i)| > \delta\} \\
&\le \limsup \sum_{i=0}^{m-1} P\{\sup_{t \in [i/m,(i+1)/m]} |Z_n(t) - Z_n(t_i)| > \delta\} \\
&= \limsup mP\{\sup_{t \in [0,1/m)} |Z_n(t) - Z_n(t_i)| > \delta\}
\end{aligned}$$

using the stationarity in the last step. The last probability will be replaced by one involving only $Z_n(1/m)$. For notational convenience put $b = 1/m$.

Let $\mathcal{E}_t$ be the $\sigma$-field generated by $Z_n(s)$ on $0 \le s \le t$. It is determined by those $U_i$ which fall in the same interval. Conditionally on $\mathcal{E}_t$,

$$D \stackrel{\text{def}}{=} W_n(b) - W_n(t)$$

is a sum of a random number of $W_i$ which are themselves randomly selected without replacement from the $W_i$ corresponding to $U_i > t$. Given $\mathcal{E}_t$, the number of such $W_i$ has a binomial distribution with parameters $n_t$ and $p_t = (b-t)/(1-t)$, where $n_t$ is the number of $U_i > t$.

Lemma 4.4.3 below establishes the bound

$$\mathcal{E}\left( (D - (b-t))^2 \mid \mathcal{E}_t \right) \le p_t n_x^{-1} + p_t^2 (1 - W(t))^2$$

under the assumption that $n_t > 2$. To assume that $n_t > 2$ is no loss of generality since $n_t > n_b$ and $n_b/n \to b$ a.s. On the set where $|Z_n(t)| > \delta$,

$$\begin{aligned}
&P\left(|Z_n(b) - Z_n(t)| > 1/2|Z_n(t)| \mid \mathcal{E}_t\right) \\
&= P\left(|D - (b-t)| > 1/2|W_n(t) - t| \mid \mathcal{E}_t\right) \\
&\le 4\mathcal{E}\left( (D - (b-t))^2 \mid \mathcal{E}_t \right) / (W_n(t) - t)^2 \\
&\le 4\left(p_t n_x^{-1} + p_t^2 (W_n(t) - t)^2\right) / (W_n(t) - t)^2 \\
&= 4\left(p_t n_x^{-1}\right) / (W_n(t) - t)^2 + 4p_t^2 \\
&\le 4\left(p_t n_x^{-1}\right) / \left(\delta^2 n_x^{-1}\right) + 4p_t^2 \\
&\le 4p_t/\delta^2 + 4p_t^2 \\
&\le 1/2
\end{aligned}$$

for small enough $b$, that is for large enough $m$. (It is easy to see that $p_t < b(1-b)^{-1}$.)

Using Lemma 4.4.2 with $\beta = 1/2$, and the convergence of $Z_n(b)$ to $N(0, b(1-b))$

$$\begin{aligned}
\limsup mP\{\sup_{t\in[0,b]} |Z_n(t)| > \delta\} &\le \limsup 2mP\{|Z_n(b)| > \delta/2\} \\
&= 2mP\{|N(0, b-b^2)| > \delta/2\} \\
&\le 2mP\{|N(0,b)| > \delta/2\} \\
&= 2mP\{|N(0,1)| > \delta/2\sqrt{b}\} \\
&= 2mP\{N(0,1)^4 > \delta^4/16b^2\} \\
&\le 32mb^2\mathcal{E}\left(N(0,1)^4\right)\delta^{-4} \\
&\le 32m^{-1}\mathcal{E}\left(N(0,1)^4\right)\delta^{-4} \\
&< \epsilon
\end{aligned}$$

for large enough $m$. ∎

**Lemma 4.4.3** Let $D = \sum_{i=1}^{R} W_i$ where $R$ has a binomial distribution with parameters $n_t > 2$ and $p_t = (b-t)/(1-t) \stackrel{\text{def}}{=} 1 - q_t$ and given $R = r$, the $W_i$ are sampled without replacement from $n_t$ real numbers that sum to $1 - W(t)$ and whose squares sum to less than $n_x^{-1}$. Then

$$\mathcal{E}\left((D - (b-t))^2\right) \le p_t n_x^{-1} + p_t^2(1 - W(t))^2.$$

PROOF. Suppose $W_1$ and $W_2$ are selected by sampling without replacement from the $n_t$ numbers. Then $\mathcal{E}(W_1) = (1 - W(t))/n_t$ and $\mathcal{E}(W_1^2) \le n_x^{-1}/n_t$ and

$$\mathcal{E}(W_1 W_2) \le (1 - W(t))^2/(n_t - 1)n_t.$$

Using the above and $\mathcal{E}(Q) = \mathcal{E}(\mathcal{E}(Q|R))$ for various $Q$,

$$\begin{aligned}
&\mathcal{E}\left((D - (b-t))^2\right) \\
&= \mathcal{E}(D^2) - 2(b-t)\mathcal{E}(D) + (b-t)^2 \\
&= \mathcal{E}(R)\,\mathcal{E}(W_1^2) + \mathcal{E}(R^2 - R)\,\mathcal{E}(W_1 W_2) - 2(b-t)\mathcal{E}(R)\,\mathcal{E}(W_1) + (b-t)^2 \\
&\le p_t n_x^{-1} + \frac{n_t p_t q_t + (n_t p_t)^2 - n_t p_t}{n_t^2 - n_t}(1 - W(t))^2 - 2(b-t)p_t(1 - W(t)) + (b-t)^2 \\
&= p_t n_x^{-1} + p_t^2(1 - W(t))^2 - 2(b-t)p_t(1 - W(t)) + (b-t)^2
\end{aligned}$$

$$
\begin{aligned}
&= p_t n_x^{-1} + \left(p_t\left(1 - W(t)\right) - (b - t)\right)^2 \\
&= p_t n_x^{-1} + (b - t)^2\left(\left(1 - W(t)\right)/(1 - t) - 1\right)^2 \\
&= p_t n_x^{-1} + (b - t)^2\left(\left(t - W(t)\right)/(1 - t)\right)^2 \\
&= p_t n_x^{-1} + p_t^2\left(t - W(t)\right)^2
\end{aligned}
$$

∎

The bias term is

$$\sqrt{n_x}\sum W_i(1_{Y_i \le t} - 1_{Y_i^x \le t}).$$

To make it converge to zero, it is necessary to have each $Y_i$ close to $Y_i^x$, or to have the corresponding $W_i$ close to zero. The worst case arises when $F_x$ is a stochastic relative extremum of $F_\bullet$. Then all of the $1_{Y_i \le t} - 1_{Y_i^x \le t}$ are of the same sign. Picture a sum of boxcar functions with height $\sqrt{n_x}W_i$ and endpoints $Y_i$ and $Y_i^x$. The wide ones tend to be short and the tall ones thin. This will allow pointwise convergence of the sum to zero. For uniform convergence there is a further subtlety. The $Y_i^x$ endpoints are i.i.d. uniform[0,1], and so they are spread out over the interval. But the $Y_i$ endpoints are not uniform and they can pile up in arbitrarily small intervals. Since the boxcar functions with the largest weights have $x_i$ close to $x$ they also have $Y_i$ close to $Y_i^x$ and so their $Y_i$ are well spread out. The ones that might pile up have smaller weight.

So that $x_i$ close to $x$ implies $Y_i^x$ close to $Y_i$, we impose a condition on $V_\infty(F_{x_i}, F_x)$. Sufficient conditions for that condition will be given later in Lemmas 4.4.4 and 4.4.5.

We also will need a condition to cause the bias term to win the race to zero. The proof of the next theorem employs a truncation of observations for which $||x_i - x|| > \Delta_n$. The sequence $\Delta_n$ has to be small enough to impose good behavior on the truncated term. Then $W_x$ has to approach $\delta_x$ fast enough that the truncation has a negligible impact. If one takes $\Delta_n = 1/(\sqrt{n_x}\log n)$ then it will be necessary for $n_x \log n V_1(W_x, \delta_x) \to 0$ in pr. For $k$-NN this means that the average distance from $x$ must be somewhat smaller than $1/k$.

**Theorem 4.4.4** Suppose that $F_x$ is uniform$[0,1]$, that for some $D > 0$

$$V_\infty(F_{x_i}, F_x) \le M_x \|x_i - x\| \quad \text{whenever} \quad \|x_i - x\| < D, \tag{2}$$

that the probability weights $W_i$ satisfy (4.2.8ab) and that there exists a sequence

$$\Delta = \Delta_n(X_1, \ldots, X_n)$$

such that

$$\sqrt{n_x}\Delta \to 0 \text{ in pr.} \quad \text{and} \quad \sqrt{n_x}\Delta^{-1}V_1(W_x, \delta_x) \to 0 \text{ in pr.}$$

Then

$$\sqrt{n_x}(\hat{F}_x - F_x) \xrightarrow{\mathrm{D}} B_0$$

where $B_0$ is the Brownian bridge.

PROOF. Let

$$Z_n(t) = \sqrt{n_x}\sum W_i(1_{Y_i^x \le t} - t)$$

be the variance process and

$$B_n(t) = \sqrt{n_x}\sum W_i(1_{Y_i \le t} - 1_{Y_i^x \le t})$$

be the bias term. Under the conditions above the variance term converges to the Brownian bridge. It remains to show that the supremum of the absolute value of the bias process converges in probability to 0. This is done by constructing a bounding process that has the desired convergence.

To construct the bounding process, recall that

$$V_\infty(F_{x_i}, F_x) = \sup_{0<u<1} |F_{x_i}^{-1}(u) - F_x^{-1}(u)| \ge |Y_i^x - Y_i|$$

from which

$$|1_{Y_i \le t} - 1_{Y_i^x \le t}| \le 1_{t - V_\infty(F_{x_i}, F_x) < Y_i^x \le t + V_\infty(F_{x_i}, F_x)}.$$

Then

$$\begin{aligned}
|B_n(t)| &\le \sqrt{n_x}\sum W_i 1_{t-V_\infty(F_{x_i},F_x)<Y_i^x\le t+V_\infty(F_{x_i},F_x)} \\
&\le \sqrt{n_x}\sum_{|x_i-x|\le\Delta} W_i 1_{t-V_\infty(F_{x_i},F_x)<Y_i^x\le t+V_\infty(F_{x_i},F_x)} \\
&+ \sqrt{n_x}\sum_{|x_i-x|>\Delta} W_i \\
&\le \sqrt{n_x}\sum_{|x_i-x|\le\Delta} W_i 1_{t-\Delta M_x<Y_i^x\le t+\Delta M_x} \\
&+ \sqrt{n_x}V_1(W_x,\delta_x)/\Delta
\end{aligned}$$

so long as $\Delta \le D$. Since $P(\Delta > D) \to 0$ and $\sqrt{n_x}\Delta^{-1}V_1(W_x,\delta_x) \to 0$ in pr., it suffices to show that

$$G_n(t) = \sqrt{n_x}\sum W_i 1_{t-\Delta M_x<Y_i^x\le t+\Delta M_x}$$

converges uniformly to 0 in probability. At a fixed value of $t$

$$\begin{aligned}
P(|G_n(t)| > \epsilon) &\le \epsilon + P\left(\mathcal{E}\left(|G_n(t)| \mid X\right) > \epsilon^2\right) \\
&\le \epsilon + P\left(\sqrt{n_x}\sum W_i 2\Delta M_x > \epsilon^2\right) \\
&\le \epsilon + P\left(\sqrt{n_x}\Delta M_x > \epsilon^2\right) \\
&\to \epsilon.
\end{aligned}$$

Therefore the finite dimensional limiting distributions of the bounding process $G_n$ are all degenerate at 0. It remains to show that $G_n$ is nearly tight, and that is accomplished by using the near tightness of the variance process.

Notice that the range of $Y_i$ might be larger than that of $Y_i^x$. However, in this case the bias process outside the range assumes its maximum at 0 or 1. It follows that near tightness need only be shown in the interval $[0,1]$.

Pick $\epsilon > 0$ and $\delta > 0$. Pick $m$ so that

$$\limsup P\{\max_{0\le i\le m-1}\sup_{t\in[i/m,(i+1)/m)}|Z_n(t) - Z_n(t_i)| > \delta\} < \epsilon.$$

Such an $m$ was constructed in proof of Theorem 4.4.3. By symmetry it follows that

$$\limsup P\{\max_{1\le i\le m}\sup_{t\in((i-1)/m,i/m]}|Z_n(t) - Z_n(t_i)| > \delta\} < \epsilon.$$

Now for $t_i = i/m$,

$$G_n(t) = Z_n(t + \Delta M_x) - Z_n(t - \Delta M_x) + 2\sqrt{n_x}\Delta M_x$$

so that

$$\begin{aligned} G_n(t) - G_n(t_i) &= Z_n(t + \Delta M_x) - Z_n(t - \Delta M_x) \\ &\quad - Z_n(t_i + \Delta M_x) + Z_n(t_i - \Delta M_x) \\ &= Z_n(t + \Delta M_x) - Z_n(t_{i+1}) \\ &\quad + Z_n(t_{i+1}) - Z_n(t_i + \Delta M_x) \\ &\quad + Z_n(t_i) - Z_n(t - \Delta M_x) \\ &\quad + Z_n(t_i - \Delta M_x) - Z_n(t_i) \end{aligned}$$

Since $P(\Delta M_x \geq 1/m) \to 0$, it may be assumed that $\Delta M_x < 1/m$ so that $t \in [t_i, t_{i+1})$ implies that either $t + \Delta M_x \in [t_i, t_{i+1})$ or $t + \Delta M_x \in [t_{i+1}, t_{i+2})$. Similarly there are two intervals that might possibly contain $t - \Delta M_x$. Using elementary bounds

$$\limsup P\{\max_{1 \leq i \leq m} \sup_{t \in (i/m, (i+1)/m]} |G_n(t) - G_n(t_i)| > 6\delta\} < 6\epsilon.$$

This completes the proof. ∎

Theorem 4.4.4 shows that very general weighting schemes are capable of providing asymptotically Brownian estimates of the uniform distribution. The conclusion is applicable so long as the distributions of the random variables $F_x(Y_i)$ satisfy the $V_\infty$ condition above. This does not follow from a similar $V_\infty$ condition applied to $F_\bullet$. Sufficient conditions are provided by the next lemma:

**Lemma 4.4.4** Suppose that $F_x$ admits a density that is bounded above by $B < \infty$, and that $F_\bullet$ satisfies $V_\infty(F_{x_i}, F_x) \leq A_x\|x_i - x\|$. Then for some $M_x$,

$$V_\infty\left(\mathcal{L}(F_x(Y) \mid X = x_i), \mathcal{L}(F_x(Y) \mid X = x)\right) \leq M_x\|x_i - x\|.$$

PROOF.

$$\begin{aligned} V_\infty\left(\mathcal{L}(F_x(Y) \mid X = x_i), \mathcal{L}(F_x(Y) \mid X = x)\right) &= \sup_{0<u<1} |F_x\left(F_{x_i}^{-1}(u)\right) - u| \\ &\leq B \sup_{0<u<1} |F_{x_i}^{-1}(u) - F_x(u)| \\ &\leq BV_\infty(F_{x_i}, F_x) \\ &\leq= BA_x\|x_i - x\| \end{aligned}$$

so we may take $M_x = BA_x$. ∎

A restriction to distributions with bounded density is unpalatable, since it rules out such distributions as the exponential. Large densities allow $F_x(y)$ to be very different from $F_{x_i}(y_i)$ even when $y$ is close to $y_i$. It is often reasonable to suppose that when $F_x$ has a large density that $F_{x_i}$ does too, when $x_i$ is close to $x$. This motivates the next lemma.

**Lemma 4.4.5** Suppose that $F_\bullet$ satisfies

$$KS(F_{x_i}, F_x) \leq M_x\|x_i - x\|.$$

Then

$$V_\infty\left(\mathcal{L}(F_x(Y) \mid X = x_i), \mathcal{L}(F_x(Y) \mid X = x)\right) \leq M_x\|x_i - x\|.$$

PROOF. Let $U$ be a uniform [0,1] random variable.

$$\begin{aligned} V_\infty\left(\mathcal{L}(F_x(Y) \mid X = x_i), \mathcal{L}(F_x(Y) \mid X = x)\right) &= \sup_{0<u<1} |F_x(F_{x_i}^{-1}(u)) - u| \\ &= \sup_{0<u<1} |F_x(F_{x_i}^{-1}(u)) - F_{x_i}(F_{x_i}^{-1}(u))| \\ &\leq KS(F_x, F_{x_i}) \\ &\leq M_x\|x - x_i\| \end{aligned}$$ ∎

**Theorem 4.4.5** Suppose that $F_x$ is absolutely continuous, and that $W_i$ are probability weights satisfying (4.2.8ab) and that there exists a sequence

$$\Delta = \Delta_n(X_1, \ldots, X_n)$$

such that

$$\sqrt{n_x}\Delta \to 0 \text{ in pr.} \quad \text{and} \quad \sqrt{n_x}\Delta^{-1}V_1(W_x, \delta_x) \to 0 \text{ in pr.}$$

and that for $\|x_i - x\| \leq D \in (0, \infty)$, $F_\bullet$ satisfies either

$$KS(F_{x_i}, F_x) \leq M_x \|x_i - x\|$$

or

$$V_\infty(F_{x_i}, F_x) \leq M_x \|x_i - x\| \quad \text{and} \quad \sup_y \frac{d}{dy} F_x(y) \leq K < \infty.$$

Then

$$\sqrt{n_x}(\hat{F}_x - F_x) \xrightarrow{\mathrm{D}} B$$

where $B$ is a continuous Gaussian process with mean 0 and for $s < t$

$$\mathrm{Cov}(B(s), B(t)) = F_x(s)(1 - F_x(t)).$$

PROOF. Apply Lemmas Lemmas 4.4.4 and 4.4.5 and Theorem 4.4.4. ∎

Stute (1986) obtains a Brownian limit for symmetric nearest neighbor estimates with a bounded kernel function. (See Sec. 2.2 for a kernel based definition of symmetric nearest neighbor methods.) His results are obtained for multivariate $Y$ and univariate $X$. For the variance term, Stute assumes that

$$\sup_{|t-s| \leq \delta} |F_{x'}(t) - F_{x'}(s)| = o\left((\log \delta^{-1})^{-1}\right)$$

as $\delta \to 0$ unformly for $x'$ in a neighborhood of $x$. Stute remarks that this implies equicontinuity of $F_{x'}(y)$, which is refered to as $KS$ continuity here.

## 4.5 Asymptotic Normality of Running Functionals

In this section we apply the results of the earlier sections and the theory of compact differentiability to consider asymptotic normality for a class of statistical functionals. For a brief summary of compact differentiability and von Mises' method see Sec. 2.5. For a complete exposition see Fernholtz (1983) or Reeds (1976).

Suppose that the statistical functional $T$ has a compact derivative $T'_{F_x}$ at $F_x$. Then

$$\begin{aligned}\sqrt{n_x}(T(\hat{F}_x) - T(F_x)) &= \sqrt{n_x} T'_{F_x}(\hat{F}_x - F_x) + \sqrt{n_x} Rem(\hat{F}_x - F_x) \\ &= \sqrt{n_x} \sum W_i IC(Y_i; F_x, T) + \sqrt{n_x} Rem(\hat{F}_x - F_x)\end{aligned}$$

so if the random variables $V_i = IC(Y_i; F_x, T)$ and the weights $W_i$ satisfy the conditions of Theorem 4.3.1 then the lead term is asymptotically normal. If also the remainder term converges to 0 in probability, then $\sqrt{n_x}(T(\hat{F}_x) - T(F_x))$ is asymptotically normal. For each functional, good behavior of the lead term implies a regularity condition on $F_\bullet$. We establish asymptotic negligibility of the remainder term.

Following Fernholz (1983,Ch 4), we assume that $F_x$ is $U[0,1]$ and that the statistical functional is defined on $D[0,1]$. This is only a slight loss of generality. A statistical functional $T$ induces a functional $\tau$ on $D[0,1]$ by $\tau(G) = T(G \circ F_x)$. So long as $F_x$ is increasing, any distribution function can be expressed as $G \circ F_x$ for some $G$. The asymptotic negligibility of the remainder term will be established by an argument that parallels Fernholz's (1983, Secs. 4.1-4.3) which is in turn based on a method of Reeds' (1976, Sec.6.5). There are two important differences. Since $\hat{F}_x$ is measureable in this treatment, there is no need to appeal to inner or outer measures. More seriously, the unequal weighting of observations in $\hat{F}_x$ adds complication. It will be necessary to assume that the weighting is not *too* unequal.

We use the following lemma from Fernholz (1983). The distance between $H \in D[0,1]$ and $K \subset D[0,1]$ will be taken to be $dist(H,K) = \inf_{G \in K} ||H - G||$.

**Lemma 4.5.1** Let $Q : D[0,1] \times \mathbb{R} \to \mathbb{R}$ and suppose for any compact set $K \subset D[0,1]$

$$\lim_{t \to 0} Q(H,t) = 0$$

uniformly in $H \in K$. Let $\epsilon > 0$ and let $\delta_n \downarrow 0$ be a sequence of numbers. Then for any compact $K \subset D[0,1]$, there exists $n_0$ such that for $n > n_0$, and $dist(H,K) \leq \delta_n$,

$$|Q(H, \delta_n)| < \epsilon.$$

PROOF. See Fernholz (1983, Lemma 4.3.1)

To apply Lemma 4.5.1 to a sequence with $\delta_n \to 0$ in pr., the following version is of more direct use.

**Corollary** Let $Q$ be as in Lemma 4.5.1 and let $\epsilon > 0$. Then for any compact $K \subset D[0,1]$ there exists $\eta > 0$ such that $\delta < \eta$ and $dist(H, K) \leq \delta$ implies

$$|Q(H, \delta)| < \epsilon.$$

PROOF. Suppose not. Then there is a compact set $K$ and infinite sequences $\eta_i \downarrow 0$ and $H_i$ such that $dist(H_i, K) \leq \eta_i$ and $Q(H_i, \eta_i) > \epsilon$. But this contradicts Lemma 4.5.2. ∎

The next lemma is used in the proof of the convergence of the remainder term. It is of some interest in its own right since it has weaker conditions on the weights than Theorem 4.5.1. Introduce the process $\widetilde{F}_x$ for which

$$\widetilde{F}_x(Y_i) = \hat{F}_x(Y_i), \quad \widetilde{F}_x(0) = 0, \quad \widetilde{F}_x(1) = 1$$

and $\widetilde{F}_x$ is piecewise linear over the $n+1$ intervals between those points. Assume that the $F_{x_i}$ are continuous distributions so that there are no ties. Then by construction

$$|\widetilde{F}_x(y) - \hat{F}_x(y)| < \max_{1 \leq i \leq n} |W_i|$$

for all $y \in [0,1]$. For the rest of this section

$$W = \max_{1 \leq i \leq n} |W_i|.$$

**Lemma 4.5.2** Suppose that $T$ has compact derivative $T'_U$ at $U = F_x$, all the $F_{x_i}$ are continuous and that

$$\sqrt{n_x}(\hat{F}_x - F_x) \xrightarrow{\mathrm{D}} B_0 \tag{1}$$

where $B_0$ is the Brownian bridge. Then

$$\sqrt{n_x} Rem(\widetilde{F}_x - F_x) \to 0 \text{ in pr.} \tag{2}$$

**Remarks** Sufficient conditions for (1) are given in Theorem 4.4.4. See also Theorem 4.4.5. Note that (1) implies $n_x \to \infty$ in pr. and $\sqrt{n_x} W \to 0$ in pr.

PROOF. Let $\epsilon > 0$. The process $\sqrt{n_x}(\widetilde{F}_x - F_x)$ is within $\sqrt{n_x}W$ of $\sqrt{n_x}(\hat{F}_x - F_x) \xrightarrow{D} B_0$. Since $\sqrt{n_x}W \to 0$ in pr. it follows that $\sqrt{n_x}(\widetilde{F}_x - F_x) \xrightarrow{D} B_0$. Moreover since $\sqrt{n_x}(\widetilde{F}_x - F_x) \in C[0,1]$, a separable metric space, there is by Prohorov's theorem a compact set $K \in C[0,1]$ such that

$$P\left(\sqrt{n_x}(\widetilde{F}_x - F_x) \in K\right) > 1 - \epsilon.$$

$K$ is also compact in $D[0,1]$.

Because $T$ is compactly differentiable, for $H \in K$ and $n_x$ sufficiently large (greater than $n_*$ say)

$$\left|\sqrt{n_x}Rem\left(\frac{1}{\sqrt{n_x}}H\right)\right| < \epsilon.$$

Therefore

$$P\left(\left|\sqrt{n_x}Rem(\widetilde{F}_x - F_x)\right| > \epsilon\right) < \epsilon + P(n_x \le n_*) \to \epsilon,$$

and so $\sqrt{n_x}Rem(\widetilde{F}_x - F_x) \to 0$ in pr. ∎

We see also that if $T$ has a *Frechet* derivative then $\sqrt{n_x}Rem(\hat{F}_x - F_x) \to 0$ in pr. under the conditions of Lemma 4.5.2. This is because the set $\{H : dist(H,K) < \epsilon\}$ is bounded for compact $K$ and the remainder term converges to zero uniformly over bounded sets under Frechet differentiability.

**Theorem 4.5.1** Suppose that $T$ has compact derivative $T'_U$ at $U = F_x$, all the $F_{x_i}$ are continuous, that (2) holds and that $n_x W = O_p(1)$. Then $\sqrt{n_x}Rem(\hat{F}_x - F_x) \to 0$ in pr..

**Remark** The condition that $n_x W = O_p(1)$ is not too restrictive. A "fair share" for a point would be $1/n_x$ and the condition bounds the multiple of that amount that any point can receive. Also note that by consideration of the finite dimensional distribution functions that (2) implies (4.2.8abc), and in particular that $n_x \to \infty$ in pr.

PROOF. Let $\epsilon > 0$. Choose $B < \infty$ so that $\limsup P(n_x W \ge B) < \epsilon$. By the argument of Lemma 4.5.2 there is a compact set $K \subset D[0,1]$ such that

$$P\left(\sqrt{n_x}(\widetilde{F}_x - F_x) \in K\right) > 1 - \epsilon$$

and since

$$||\widetilde{F_x} - \hat{F}_x|| \leq W$$

by construction,

$$P\left(dist(\sqrt{n_x}(\hat{F}_x - F_x), K) > \sqrt{n_x}W\right) < \epsilon.$$

By the compact differentiability of $T$ at $U$, the function

$$Q(H,t) = \frac{Rem(B^{-1}Ht)}{B^{-1}t}$$

satisfies the conditions of Lemma 4.5.1. By the Corollary to Lemma 4.5.1 there exists $\eta > 0$ such that $\delta < \eta$ and $dist(H,K) \leq \delta$ imply $|Q(H,t)| < \epsilon$.

Therefore

$$\begin{aligned}
&P\left(|\sqrt{n_x}Rem(\hat{F}_x - F_x)| > \epsilon\right) \\
=&P\left(|Q(\sqrt{n_x}(\hat{F}_x - F_x), B/\sqrt{n_x})| > \epsilon\right) \\
\leq&P\left(B/\sqrt{n_x} > \eta\right) + P\left(dist(\sqrt{n_x}(\hat{F}_x - F_x), K) > B/\sqrt{n_x}\right) \\
\to&P\left(dist(\sqrt{n_x}(\hat{F}_x - F_x), K) > B/\sqrt{n_x}\right) \\
\leq&\epsilon + P\left(\sqrt{n_x}W > B/\sqrt{n_x}\right) \\
=&\epsilon + P\left(n_x W > B\right)
\end{aligned}$$

so that

$$\limsup P(|\sqrt{n_x}Rem(\hat{F}_x - F_x)| > \epsilon) \leq 2\epsilon. \qquad \blacksquare$$

Which statistical functionals induce functionals on $D[0,1]$ that are compactly differentiable at $U$? Fernholtz (1983) establishes such compact differentiability for $M$ estimates with continuous piecewise differentiable $\psi$, such that $\psi'$ is bounded and vanishes off a bounded interval, when $F$ has a piecewise continuous positive density. She also establishes compact differentiability for some $L$ estimates:

$$\int_0^1 h(F_x^{-1}(u))M(u)du$$

**provided $h$ is continuous and piecewise differentiable with bounded derivative (usually one takes $h(y) = y$) and $M \in L^2[\alpha, 1-\alpha]$ for some $\alpha \in (0, 1/2]$, and $F$ has a positive density. Similar regularity conditions on $R$ estimators make them compactly differentiable. Quantiles get special treatment. She shows that they induce compactly differentiable functionals on $C[0,1]$ when $F$ is well behaved near the quantile in question. The asymptotic negligibility of the remainder term for quantiles then follows by considering continuous versions of the empirical distribution function that are constructed to agree with the empirical at the quantile.**

ACKNOWLEDGEMENTS

I would like to thank Jerome Friedman and Iain Johnstone for their helpful comments.

# References

Anderson, T.W. (1971). *The Statistical Analysis of Time Series.* Wiley, New York.

Benedetti, J.K. (1977). On the Non Parametric Estimation of Regression Functions. *J.R.S.S.* Ser. B, vol. 39, 241–253.

Bickel, P.J. (1977). Discussion of: Consistent Nonparametric Regression by Stone, C.J. *Ann. Statist.* 5 595–645.

Bickel, P.J. and Freedman, D.A. (1981). Some Asymptotic Theory for the Bootstrap. *Ann. Statist.* 9 1196–1217.

Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

Billingsley, P. (1971). *Weak Convergence of Measures.* Reg. Conf. Ser. in App. Math, SIAM, Philadelphia.

Billingsley, P. (1979). *Probability and Measure.* Wiley, New York.

Brieman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, CA.

Brillinger, D.R. (1977). Discussion of: Consistent Nonparametric Regression by Stone, C.J.. *Ann. Statist.* 5 595–645.

Chatfield, C. (1980). *The Analysis of Time Series,* 2nd Ed. Chapman and Hall, London.

Choquet, G. (1969). *Lectures on Analysis.* W.A. Benjamin Inc., Reading, MA.

Chung, K.L. (1974). *A Course in Probability Theory,* 2nd Ed. Academic Press, New York.

Cleveland, W.S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *J.A.S.A.* 74, 828–836.

Collomb, G. (1985). Nonparametric Regression: An Up-To-Date Bibliography. *statistics* 16 309–324.

Craven, P. and Wahba, G. (1979). Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation. *Numer. Math.* 31, 377–403.

Devroye, L.P. (1981). On the Almost Everywhere Convergence of Nonparametric Regression Function Estimates. *Ann. Statist.* 9 1310–1319.

Devroye, L.P. (1982). Necessary and Sufficient Conditions for the Pointwise Convergence of Nearest Neighbor Regression Function Estimates. *Zeit. fur Wahr. und ver. Geb.* 61, 467–481.

Dobrushin, R.L. (1970). Describing a System of Random Variables by Conditional Distributions. *Theory Probab. Appl.* 15 458–486.

Doksum, K.A. and Yandell, B.S. (1983). Properties of Regression Estimates Based on Censored Survival Data. *Festschrift for Erich L. Lehmann,* Bickel, P.J. Doksum, P.J. and Hodges, J.L., Editors 140–156, Wadsworth, Inc., Belmont, CA.

Efron, B. (1967). The Two Sample Problem with Censored Data. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. IV* 831–853, Univ. of Cal. Press, Berkeley CA.

Epanechnikov, V.A. (1969). Nonparametric Estimation of a Multivariate Probability Density. *Theory Probab. Appl.* 14, 153–158.

Fernholz, L.T. (1983). *von Mises Calculus for Statistical Functionals.* Lecture Notes in Statistics, No. 19, Springer-Verlag, New York.

Fix, E. and Hodges, J.L. (1952). "Discriminatory Analysis, Nonparametric Discrimination, Consistency Properties". Randolph Field, Texas, Project 21-49-004, Report No. 4..

Friedman, J.H. (1984). "A Variable Span Smoother". Dept. of Statistics Tech. Rep. LCS 5, Stanford University.

Friedman, J.H. and Stuetzle, W. (1983). "Smoothing of Scatterplots". Dept. of Statistics Tech. Rep. ORION 3, Stanford University.

Gasser, T. and Muller, E.G. (1977). Kernel Estimation of Regression Functions. *Lecture Notes in Mathematics* 757, 23–68, Springer-Verlag, New York.

Hall, P. (1984). Asymptotic Properties of Integrated Square Error and Cross-Validation for Kernel Estimation of a Regression Function. *Zeit. fur Wahr. und ver. Geb.* 63, 175–195.

Hampel, F.R. (1971). A General Qualitative Definition of Robustness. *Ann. Math. Statist.* 42, 1887–1896.

Hardle, W. and Gasser, T. (1984). Robust Nonparametric Function Fitting. *J.R.S.S.* Ser.

B 46, 42–51.

Huber, P.J. (1981). *Robust Statistics.* Wiley, New York.

Loftsgaarden, D.O. and Quesenberry, C.P. (1965). A Nonparametric Estimate of a Multivariate Density Function. *Ann. Math. Statist.* 36, 1045–1051.

Mack, Y.P. and Silverman, B.W. (1982). Weak and Strong Uniform Consistency of Kernel Regression Estimates. *Zeit. fur Wahr. und ver. Geb.* 61, 405–415.

Major, P. (1978). On the Invariance Principle for Sums of Independent, Identically Distributed Random Variables. *Jour. Mult. Anal.* 8, 487–501.

Mallows, C.L. (1972). A Note on Asymptotic Joint Normality. *Ann. Math. Statist.* 43, 508–515.

Marhoul, J.C. and Owen, A.B. (1985). "Consistency of Smoothing with Running Linear Fits". Dept. of Statistics Tech. Report LCS 8, Stanford University.

McDonald, J.A. Owen, A.B. (1986). Smoothing with Split Linear Fits. *Technometrics* 28, 195–208.

Nadaraya, E.A. (1964). On Nonparamteric Estimates of Density Functions and Regression Curves. *Theory Probab. Appl.* 15, 134–137.

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *Ann Math. Statist.* 33, 1065–1076.

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer Series in Statistics, Springer-Verlag, New York.

Priestley, M.B. and Chao, M.T. (1972). Nonparamteric Function Fitting. *J.R.S.S.* Ser. B 34, 385–392.

Prohorov, Yu. V. (1956). Convergence of Random Processes and Limit Theorems in Probability Theory. *Theory Probab. Appl.* 1, 157–214.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications* 2nd Ed. Wiley, New York.

Reeds, J.A. (1976). *On the Definition of von Mises Functionals.* Ph.D. Dissertation, Harvard University.

Reinsch, C.H. (1967). Smoothing by Spline Functions. *Numer. Math.* 10, 177–183.

Ripley, B.D. (1981). *Spatial Statistics.* Wiley, New York.

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *Ann. Math. Statist.* 27, 832–837.

Royall, R.M. (1966). *A Class of Nonparametric Estimates of a Smooth Regression Function.* Ph.D. Dissertation, Stanford University.

Schuster, E.F. (1972). Joint Asymptotic Normality of the Estimated Regression Function at a Finite Number of Distinct Points. *Ann. Math. Statist.* 43, 84–88.

Segal, M.R. (1986). *Regression Trees Based on Rank Statistics.* Ph.D. Dissertation, Stanford University.

Silverman, B.W. (1984). Spline Smoothing: The Equivalent Variable Kernel Method. *Ann. Statist.* 12, 898–916.

Silverman, B.W. (1985). Some Aspects of the Spline Smoothing Approach to Nonparametric Regression Curve Fitting. *J.R.S.S.* Ser. B 47.

Stone, H.M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *J.R.S.S.* Ser. B 36, 111–147.

Stone, C.J. (1977). Consistent Nonparametric Regression. *Ann. Statist.* 5, 595–645.

Stout, W.F. (1969). Some Results on the Complete and Almost Sure Convergence of Linear Combinations of Independent Random Variables and Martingale Differences. *Ann. Math. Statist.* 39, 1549–1562.

Stute, W. (1984). Asymptotic Normality of Nearest Neighbor Regression Function Estimates. *Ann. Statist.* 12, 917–926.

Stute, W. (1986). Conditional Empirical Processes. *Ann. Statist.* 14, 638–647.

Tibshirani, R.J. (1984). *Local Likelihood Estimation.* Ph.D. Dissertation, Stanford University.

Titterington, D.M. (1985). Common Structure of Smoothing Techniques in Statistics. *Int. Statist. Rev.* 53, 141–171.

Tukey, J.W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading MA.

Wahba, G. (1975). Smoothing Noisy Data with Spline Functions. *Numer. Math.* 24, 383–393.

Wahba, G. and Wold, S. (1975). A Completely Automatic French Curve: Fitting Spline Functions by Crossvalidation. *Comm. in Statist.* 4, 1–17.

Watson, G.S. (1964). Smooth Regression Analysis. *Sankhya* Ser. A 26, 359–372.

Watson, G.S. (1984). Smoothing and Interpolation by Kriging and with Splines. *Math. Geol.* 16, 601–615.

Wegman, E.J. and Wright, I.W. (1983). Splines in Statistics. *J.A.S.A.* 78, 351–365.

Willard, S. (1970). *General Topology*. Addison-Wesley, Reading MA.

Yakowitz, S.J. and Szidarovszky, F. (1985). A Comparison of Kriging and Nonparametric Regression Methods. *Jour. Mult. Anal.* 16, 21-53.

Yang, S. (1981). Linear Functions of Concomitants of Order Statistics with Application to Nonparametric Estimation of a Regression Function. *J.A.S.A.* 76, 658–662.

Zhao, L. and Bai, Z. (1984). Strong Consistency of the Nearest Neighbor Estimates of Nonparametric Regression Functions. *Scienta Sinica* Ser. A XXVII, 1027–1034.

Zhao, L. and Fang, Z. (1984). Strong Convergence of Kernel Estimates of Nonparametric Regression Functions. *Chin. Ann. of Math.* Ser. B 6, 147–155.