Resampling Evaluation of Signal Detection and Classification

With Special Reference to Breast Cancer, Computer-Aided Detection and the Free-Response Approach

ANNA BORNEFALK HERMANSSON
Dissertation presented at Uppsala University to be publicly examined in Hörsal 2, Ekonomikum, Kyrkogårdsgatan 10, Uppsala, Friday, February 16, 2007 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

The first part of this thesis is concerned with trend modelling of breast cancer mortality rates. By using an age-period-cohort model, the relative contributions of period and cohort effects are evaluated once the unquestionable existence of the age effect is controlled for. The result of such a modelling gives indications in the search for explanatory factors. While this type of modelling is usually performed with 5-year period intervals, the use of 1-year period data, as in Paper I, may be more appropriate.

The main theme of the thesis is the evaluation of the ability to detect signals in x-ray images of breasts. Early detection is the most important tool to achieve a reduction in breast cancer mortality rates, and computer-aided detection systems can be an aid for the radiologist in the diagnosing process.

The evaluation of computer-aided detection systems includes the estimation of distributions. One way of obtaining estimates of distributions when no assumptions are at hand is kernel density estimation, or the adaptive version thereof that smoothes to a greater extent in the tails of the distribution, thereby reducing spurious effects caused by outliers. The technique is described in the context of econometrics in Paper II and then applied together with the bootstrap in the breast cancer research area in Papers III-V.

Here, estimates of the sampling distributions of different parameters are used in a new model for free-response receiver operating characteristic (FROC) curve analysis. Compared to earlier work in the field, this model benefits from the advantage of not assuming independence of detections in the images, and in particular, from the incorporation of the sampling distribution of the system's operating point.

Confidence intervals obtained from the proposed model with different approaches with respect to the estimation of the distributions and the confidence interval extraction methods are compared in terms of coverage and length of the intervals by simulations of lifelike data.

Keywords: breast cancer, trend modelling, FROC, confidence intervals, threshold independence, bootstrap, kernel density estimation, mammography, computer-aided detection

Anna Bornefalk Hermansson, Department of Information Science, Box 513, Uppsala University, SE-75120 Uppsala, Sweden

© Anna Bornefalk Hermansson 2007

ISSN 1652-9030
ISBN 978-91-554-6783-8
urn:nbn:se:uu:diva-7452 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-7452)
To Jarl and Mattis
List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


V  Bornefalk Hermansson, A. (2006) Evaluation of the bootstrap smoothed conditional (BSC) interval and others in small sample FROC analysis of mammography CAD system performance *Submitted*

Reprints were made with permission from the publishers.
Contents

1 Introduction .......................................... 9
2 Statistics in epidemiology ............................... 11
  2.1 Age-period-cohort analysis ............................ 12
  2.2 A classical approach ................................. 13
  2.3 Using 1-year period data ............................. 13
3 A new approach to the analysis of FROC curves ............... 17
  3.1 Performance evaluation of a CAD system .............. 18
  3.2 Bootstrapping the leave-one-out ...................... 20
  3.3 Smoothing-after-bootstrap with kernel density estimation ... 21
  3.4 Verification with cross-validations .................... 23
  3.5 Simulations with the XFROC model .................... 24
4 Summary of the papers ................................... 25
  4.2 Paper II: The distributions of perceptions and expectations of inflation in Sweden ................................. 25
  4.3 Paper III: On the comparison of FROC curves in mammography CAD systems ........................................ 26
  4.4 Paper IV: Statistical aspects of threshold independent performance assessment of mammography CAD systems .... 26
  4.5 Paper V: Evaluation of the bootstrap smoothed conditional (BSC) interval and others in small sample FROC analysis of mammography CAD system performance .............. 27
5 Bibliography ........................................... 29
1. Introduction

Breast cancer was the cause of death for 67 419 Swedish women between 1952 and 2003, in the age range 30 to 89 years alone. It is the leading cause of cancer deaths for women in Sweden, as in most of the western world, and reduced mortality as well as incidence is of continuous urgency. It has been shown that nation-wide mammography screening programs reduce mortality, since early detection is the most important factor for a positive outcome [1]. Specialized radiologists examine the images and computer-aided methods may be used as a second reader to increase the detection rate. These methods may also increase recall rates above some desired level, without raising the number of detected cancers. The performance of a computer-aided detection (CAD) system is commonly characterized by a free-response receiver operating characteristic (FROC) curve [2, 3]. This curve captures the inherent trade-off between a high detection rate and a low number of false markings per image.

This thesis is concerned with the trendwise development of breast cancer mortality and accompanying modelling aspects, computer-intensive methods in breast cancer research and, in particular, the development, verification and evaluation of a new model for CAD performance assessment. The model includes the uncertainty of the system’s operating point and it does not assume independence of detections.
2. Statistics in epidemiology

The mortality of breast cancer has been quite stable in Sweden during the second half of the 20th century, at around 31 deaths per 100 000 and year [4], when averaging the age specific rates with the mean population (all age groups) of Swedish women in 1970 as weights. Concentrating on women 30-89 years of age, the mean is around 52 deaths per 100 000 and year, while incidence has increased from 142 in 1970 to 231 in 2004 (see figure 2.1). A slight upward trend in mortality was present before the introduction of mammography screening in the 1970s, from when the trend shifted downwards with a moderate intensification. The upward incidence trend intensified in the mid 1980s, reflecting, among other things, the full expansion and high attendance to the nation-wide screening programs, since the earlier detection also assembles those women who would otherwise have died from competing causes before the breast cancer was discovered. The actual increase in incidence depends on several factors, including a modern lifestyle that increases the risk of breast cancer.

![Figure 2.1: Breast cancer incidence (1970-2004) and mortality (1952-2003) for Swedish women in the ages 30-89. Age standardised to the Swedish mean population of women in the ages 30-89 in 1970.](image-url)
2.1 Age-period-cohort analysis

Age-period-cohort analysis is used in cancer epidemiology as a descriptive tool for the analysis of cancer rates. There is an extensive amount of literature on the subject, see e.g. [5, 6, 7, 8, 9]. The analysis clarifies trends and narrows the field of factors that could affect the epidemiology of a disease. It gives important indications and is a proper first step in the search for explaining and prognostic factors. For example, a cohort effect suggests a mortality (or incidence) decrease (or increase) that pertains to individuals born in a specific period, i.e. something happened this period that affects these people for the rest of their lives. This could be the introduction of a vaccine for newborn babies. A period effect, on the other hand, implicates that something happened during a specific period, which would have an impact over the entire population.

An age effect manifests itself as a slope in the age group direction (see figure 2.2), a period effect as the presence of a slope in the time direction and a cohort effect as diagonal ridges or valleys. These data show a clear age effect, as do all-cause mortality. Period and cohort effects are more difficult to see in this figure.

![Figure 2.2: Breast cancer mortality rates (per 100 000), Swedish women, by age and year.](image)

Mortality and incidence data may be affected by revisions in data collection protocols and changes in definitions of diseases, and it is important to gather information on these matters and test their effects.
2.2 A classical approach

The traditional basis for age-period-cohort modelling is two matrices containing number of cases and mean population. The most common situation is to use 5-year classes for both age and period. With \( A \) age classes (e.g., 30-34, ..., 85-89) and \( P \) period classes (e.g., 1952-56, ..., 1997-2001), the rates are organized in a matrix of size \( P \times A \). The diagonals in this matrix can be interpreted as cohorts, if the net demographic change in the population can be assumed to be negligible. Synthetic partially overlapping cohorts are constructed from the age-period data.

Let \( r(a, p) \) be the hazard function for breast cancer mortality in the population, where \( a \) is age, \( p \) is time period and \( c \) is cohort, with \( c = p - a \). Plotting the rows and diagonals of the rate matrix gives curves with exponential shapes, so the model is defined as

\[
r(a, p) = \exp(f(a) + g(p) + h(c)).
\]  

(2.1)

In the classical approach,

\[
f(a) = \sum \beta_i D_{ai} \]  

(2.2)

\[
g(p) = \sum \gamma_j D_{pj} \]  

(2.3)

\[
h(c) = \sum \zeta_k D_{ck} \]  

(2.4)

where the \( D \)'s are dummies for the respective cells. It is well known that estimation of the full age-period-cohort model causes difficulties and all parameters are not uniquely identified. The partial age-period and age-cohort models cause no difficulties when it comes to estimation. For the estimation of the age-period and the age-cohort model, the dummy variable for the reference period or cohort is dropped. The full age-period-cohort model calls for one reference period and two reference cohorts, or vice versa. The risk for age group \( i \) is \( \exp(\beta_i) \) in the reference period and cohort and \( \exp(\gamma_j) \) and \( \exp(\zeta_k) \) are the period and cohort effects.

It is usually assumed that the number of cases is Poisson distributed and maximum likelihood estimation of this model normally causes no difficulties. An alternative that usually produces very similar results is to use weighted least squares estimation, with weights equal to the number of cases. This could be preferred because of the overdispersion in the data, since it leads to more reliable estimates of the standard errors.

2.3 Using 1-year period data

In recent years, interest has been turned towards the use of annual data in trend modelling. Annual age data is particularly important when analysing childhood diseases. Analysis of all three effects in 1-year intervals calls for alternative methods, since the number of parameters must be kept at a manageable
level. The functions \( f(a), g(p) \) and \( h(c) \) in equation 2.1 could be modelled with other parametric functions than equations 2.2-2.4, e.g. with local or fractional polynomials, or splines. However, too much smoothing is not desirable in the descriptive age-period-cohort model, since sudden changes in effects are perfectly sensible. Other alternatives are e.g. using the variogram [10] to evaluate the anisotropy of figure 2.2, or a frequency domain [11] representation of the data, where the age effect would appear as a strong signal at a frequency corresponding to a cycle of length \( A \), and shorter cycle signals would indicate period or cohort effects. Still, the display of the different effects calls for easy interpretation, since nonspecialists like health policy makers must be able to understand the results.

In paper I [4], 1-year period data were used in combination with the classical 5-year definition of periods, i.e. with a piecewise constancy assumption. The use of 1-year data produced more accurate estimates for the data in paper I, but these data are not typical in one respect: there is little trendwise change compared with what we find in many other situations, where the trendwise increase (or sometimes decrease) can often be very large. In general, this would make the risk of obtaining serially correlated errors greater, since if the true trend were linear, the residuals would go from large negative to large positive within each 5 year period. Unless corrected for when standard errors are computed, this would result in biased variability estimates. A formal testing of the randomness of the residuals is less easy than in many other situations as there are several dimensions in the data. The ordering of the observations will influence the results of a standard test like the Durbin-Watson test or tests based on runs and autocorrelation coefficients. Thus even if we just consider age and period, we will get different results depending on whether we order the data rowwise or columnwise. With the data ordered as a time series within each group, the Durbin-Watson statistic computed with account taken of the fact that we have \( A \) different segments of \( P \) observations each (what is often called accounting for ‘gaps’ in the data), we obtained an implication of a moderate degree of serial correlation, with not much difference between 5- and 1-year data.

A comparison of the estimates based on the 5-year and 1-year data revealed that the point estimates are very similar and the differences of no real consequence. The standard errors on the other hand are smaller when the estimates are based on 1-year data, which may increase the number of significant effects. The degree of overdispersion is much smaller with 1-year data. With 5-year data, it is considerable, which is an added argument for using weighted least squares rather than maximum likelihood based on a Poisson assumption.

With 1-year period effects, estimates close in time tend to be similar for these data, which means that the more restrictive standard model gives rather similar results. The reduction in the residual sum of squares for the standard age-period model when 1-year period effects are used instead of 1-year data with 5-year period effects is not much greater than would be expected due
to the difference in number of parameters included. Thus there is no obvious loss if the standard age-period set-up is used in combination with yearly data. After correction for the scaling the results are almost identical.
3. A new approach to the analysis of FROC curves

In the twentieth century, most breast cancers were detected by palpation. The outcome of these cancers was poor, because of the progression to advanced stage systemic diseases with metastases. Prevention of the disease from progressing requires early detection, which offers considerably better survival for affected women [12]. Mammography (x-ray imaging of the breast) makes it possible to find tumours long before they become palpable and the method had in the late 1970s become sufficiently developed to detect a high proportion of cancers. The screening needs to be repeated at regular intervals, and properly performed and interpreted. Detection with mammography also gave an opportunity to study the early natural history of breast cancer.

A mammography screening typically involves taking two views of the breast, from above (cranial-caudal view, CC) and from a 45-degree angled view (mediolateral-oblique, MLO). A radiologist, who is a physician experienced in mammography and other x-ray examinations, analyses the images, describes any abnormalities, and suggests a likely diagnosis.

As an aid in the diagnosing process, a computer-aided detection (CAD) system can be used. Digitalisation of the mammograms (a step that will disappear when full-field digital mammography replaces film-screen mammography) and application of a CAD algorithm results in markings of suspicious areas on screen, which are re-examined by the radiologist before the diagnosis is proposed.

The CAD algorithm contains four steps: preparation, sifting, feature extraction and classification. In the second step, the CAD system extracts candidate detections, or regions of interest (ROI), based on edge and spiculation measures. An example of such an extraction is shown in figure 3.1. The mammogram in the figure is obtained from the Digital Database for Screening Mammography (DDSM) at the university of South Florida [13] and the CAD system used is developed by Bornefalk [14, 15]. The classification of the ROIs as normal or abnormal can e.g. be performed by a support vector machine (SVM). The SVM creates a maximum-margin hyperplane that lies in a transformed input space. Given training examples labelled either ‘normal’ or ‘abnormal’ and the location of each lesion, a maximum-margin hyperplane splits the normal and abnormal training examples, such that the distance from the closest examples (the margin) to the hyperplane is maximized. The input to the SVM is one vector per ROI, with numerical descriptions of several fea-
tures describing the region. The output from the SVM is real numbers, each corresponding to one ROI. The discriminant rule is the system’s operating point $\theta^*$, which divides the non-malignant output from the malignant in an optimal manner.

Figure 3.1: Locations extracted by the CAD system that are classified in a later step. The arrow indicates the location of the malignant lesion.

3.1 Performance evaluation of a CAD system

The performance of a CAD system is usually evaluated using free-response receiver operating characteristic (FROC) methodology [16], which originates from receiver operating characteristic (ROC) methodology [17]. The latter is basically a way to measure the quality of a binary classification. Egan et al. [2] introduced the term ‘free-response’ in an article on how well a listener can hear weak signals against a background of noise, a research supported by the US Air Force. The listener did not know when the tone would occur or how many tones would be presented during the long observation interval and he was instructed to press a key once each time he heard a tone, i.e. the same scenario as in cancer detection. Bunch et al. [3] adapted the method of free-response to medical imaging. With this approach, multiple responses as well as no responses are allowed for each image, which is not the case with the ROC methodology.

The outcomes are true positive (the ROI is in fact a lesion) and false positive. An FROC curve illustrates the relationship between the proportion of true positives (the sensitivity), which is defined as $tp = \#\text{true positives}/\#\text{tumours}$, and the mean number of false positives per image, which is defined as $fp$ per image $= \#\text{false positives}/\#\text{images}$, for a spectrum of different discriminant
rules, i.e. for different values of a threshold \( \theta \). The lower the threshold is, the higher is the sensitivity, but at the cost of more false positives. Usually, different FROC curves are compared to see which system (or which algorithm within a system) works best. In technical reports and articles, a point on the curve commonly summarizes the result. However, the constitution of the data set used in the training and evaluation of the system, as well as the criteria used for deciding when e.g. a marking is close enough to be considered a true positive, differs substantially between studies, and this affects the location of the FROC curve [18]. In [15] and [19], we suggest summarising the result with an interval for the expected number of false positives given a reasonable aim for the sensitivity, which often is 90\%. The presented techniques can also be used for estimation of the expected sensitivity given a tolerable level of false positives. The model developed is the first to acknowledge that FROC curves are three-dimensional.

We have a joint parametric probability distribution of \( tp, fp \) and \( \theta \):

\[
(tp, fp, \theta) \sim \phi(tp, fp, \theta), \theta \in \Theta.
\]  

(3.1)

The space \( \Theta \) is of finite dimension. Appropriate measurability properties are assumed to hold for all conditional densities (or mass functions). Write

\[
\phi(tp, fp, \theta) = g(fp | tp, \theta)f(tp | \theta)h(\theta)
\]  

(3.2)

where \( h(\theta) \) describes the values of \( \theta \), for which the data for \( tp \) and \( fp \) is generated conditionally upon.

The density of \( fp \) conditionally on \( tp \) is

\[
g(fp | tp) = \frac{\int_\Theta \phi(tp, fp, \theta)d\theta}{\int_\Theta f(tp | \theta)h(\theta)d\theta}
\]

\[
= \int_\Theta g(fp | tp, \theta) \frac{f(tp | \theta)h(\theta)}{\int_\Theta f(tp | \theta)h(\theta)d\theta}d\theta
\]

\[
= \int_\Theta g(fp | tp, \theta)h(\theta | tp)d\theta.
\]  

(3.3)

In a parametric approach we assume that \( fp \) and \( tp \) are conditionally independent given \( \theta \) (i.e. that the occurrences of the two different signal types in any image are independent events):

\[
g(fp | tp, \theta) = g(fp | \theta).
\]  

(3.4)

The data set from the leave-one-out construction of the FROC curve consists of a vector of thresholds (in our case of size 200), with one observation on \( tp \) and \( 2n \) observations on false positives per image corresponding to each \( \theta \)-value, where \( n \) is the number of cases (in our case 45), with two images on each case. We have pairwise evaluation, i.e. the tumour is considered found if it is detected in at least one of the CC and MLO views.

19
For each image, the number of false positives can be 0, 1, 2, ... and may be modelled with a Poisson distribution with parameter $f_p(\theta)$. The total number of false positives in $2n$ images would then be Poisson distributed with parameter $2n \cdot f_p(\theta)$. A rescaling gives the distribution $g(f_p | \theta)$.

An ROI is either malignant or not, i.e. the true positive indicator equals 1 or 0, and is therefore a Bernoulli variable. The sum of $n$ iid Bernoulli variables drawn from an endless population (or drawn with replacement) is binomially distributed. Given the threshold $\theta$, the detection probability is $t_p(\theta)$ and the number of true positives detected among $n$ cases is binomially distributed with parameters $n$ and $t_p(\theta)$. We rescale this to get the distribution of the fraction of tumours found given a specific threshold, $f(t_p | \theta)$.

In paper IV [19], the assumption of a Poisson distributed number of false positives per image was tested and the hypothesis was rejected, due to under-dispersion. The Poisson and binomial nature of the experiment can only be expected to hold if we use repeated data sets on the same classifier, and in our case the SVM is retrained for each of the runs. Moreover, significant correlation between the true positive indicator and the number of false positives in an image was evident in the area of interest (for those $\theta$s that can generate a sensitivity of 90%), thereby violating the assumption in equation 3.4. We therefore alter the approach to both a nonparametric one, using the bootstrap [21, 22, 23] to obtain the values of $f_p$ that could be generated with 90% sensitivity when acknowledging the three-dimensionality, and a semi-parametric one, where we use the model in equation 3.3 with estimated densities.

### 3.2 Bootstrapping the leave-one-out

We have one observation on $t_p$ and $f_p$ at each threshold level from the leave-one-out runs on the CAD system. To estimate the distributions $f(t_p | \theta)$ and $g(f_p | t_p, \theta)$, we use the bootstrap. The original data consists of two matrices of size $200 \times 45$, one with the mean number of false positive markings for each of the 45 cases at 200 threshold levels, and one with the true positive indicator for each case at each level of the threshold. In each of the 10 000 bootstrap runs, we randomly draw 45 columns, with replacement, from each of the two matrices. For the row in the matrix of true positives that gives 90% ones, we record the $\theta$ that corresponds to this row and the mean number of false positives for the same row in the false positive matrix.

An initial verification of the use of the bootstrap was given by the fact that approximately the same results were obtained when bootstrapping from a subset of the observations as from the whole set.
3.3 Smoothing-after-bootstrap with kernel density estimation

With a data set that is based on 45 cases, we can not obtain exactly 90% sensitivity, since this is between 40/45 and 41/45. We could add random noise to the bootstrap estimates to overcome this problem with discrete values [25], but we chose to smooth the bootstrap distributions with kernel density estimation [26]. The reason for this is that we found that kernel density estimation adjusts the percentiles of the final estimate of the density of $fp$ conditionally on $tp$ in a desirable manner, thereby reducing bias. As a summary of this method, consider a random sample $X_1, ..., X_n$ from a density $f(x)$, which we want to estimate. A kernel function $K$ satisfies the condition

$$\int_{-\infty}^{\infty} K(t) dt = 1. \quad (3.5)$$

The kernel density estimator with kernel $K$ is defined by

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x - X_i}{h} \right). \quad (3.6)$$

where $h$ is the bandwidth (also called window width or smoothing parameter). The mean integrated squared error is the most common choice of criterion for judging the goodness of an estimate:

$$MISE = E \left[ \int [f(x) - \hat{f}(x)]^2 \right] dx. \quad (3.7)$$

The asymptotic form of $h$ which minimizes $MISE$ is given by

$$h_{opt} = \left[ \frac{\int x^2 K(x) dx}{\int x^2 K(x) dx} \right]^{1/5}. \quad (3.8)$$

From equation 3.8 we note that $h_{opt}$ depends on the unknown density being estimated. A natural approach is to choose $h$ with reference to some standard family of densities. Choosing the normal may lead to some oversmoothing if the distribution is in fact multimodal or heavily skewed, but the resulting $h$ will serve as a good starting point.

As is apparent from equation 3.6, the kernel estimator is a sum of bumps placed at the observations. The kernel function $K$ determines the shape of the bumps while the bandwidth $h$ determines their width. The choice of kernel shape can be based on the degree of differentiability required, or the computational effort involved. The choice of bandwidth is a crucial point. When $h$ tends to zero, the estimate tends to a sum of spikes at the observations. Hence, if $h$ is chosen too small, spurious fine structures become visible. As $h$ becomes large, all detail, spurious or otherwise, is obscured. The choice of bandwidth...
can be based on one of a large number of methods for automatic bandwidth selection or on examination of several plots of the same data, all smoothed by different amounts. When density estimation is used for the presentation of conclusions it is best to undersmooth somewhat, and let the reader do further smoothing by eye.

For most practical purposes, non-negative kernels are used, so that the kernel itself is a probability density function. Then it immediately follows from the definition that the kernel density estimate will be a probability density function.

With a long-tailed distribution, spurious noise can appear in the tails of the kernel density estimate, since the bandwidth is fixed across the entire sample. If we smooth enough to deal with this, the main part of the distribution is over-smoothed. Adaptive kernel estimates, which is used in paper II [27], smooth to a greater degree in the tails of the distribution by allowing the bandwidth of the kernels to vary from one point to another. Broader kernels are used in regions of low density.

When applying the adaptive kernel method, we first have to decide whether the observation is in a region of low density. Using an initial estimate to get a rough idea of the density does this.

1. Produce a pilot estimate \( \tilde{f}(x) \) such that \( \tilde{f}(X_i) > 0 \) ∀\( i \).

The adaptive kernel method is insensitive to the fine detail of the pilot density, so any convenient estimate can be used.

Assuming \( f \) in equation 3.8 to be normal, and using the Gaussian kernel, where

\[
K(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}
\]

we have

\[
h_{opt} = 1.06 \sigma n^{-1/5}
\]

where \( \sigma \) can be estimated from the sample.

Replacing \( \sigma \) with \( A_{\text{std}} \) (standard deviation, interquartile range/1.34) and reducing the factor 1.06 in equation 3.10 improves the ability of coping with both skewness and multimodality [26]. The result is inserted into equation 3.6 to obtain the pilot density.

2. Compute local bandwidth factors \( \lambda_i \),

\[
\lambda_i = \left[ \frac{\tilde{f}(X_i)}{g} \right]^{-\alpha}
\]

where \( g \) is the geometric mean of the \( \tilde{f}(X_i) \):

\[
\log g = \frac{1}{n} \sum_{i=1}^{n} \log \tilde{f}(X_i)
\]

and \( \alpha \) is the sensitivity parameter, a number satisfying \( 0 < \alpha \leq 1 \).
That is, the local bandwidth factors depend on a power of the pilot density. There will be greater difference between bandwidths used in different parts of the sample when \( \alpha \) is large, giving the pilot density more importance. In paper II, \( \alpha \) is set to 0.5, since this gives an estimate whose bias is of smaller order than that of the fixed-width kernel estimate.

The adaptive kernel estimate is defined by
\[
\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h\lambda_i} K\left(\frac{x-X_i}{h\lambda_i}\right). \tag{3.13}
\]
This implies that the width of the kernel placed at \( X_i \) is \( h\lambda_i \). Again, the Gaussian kernel in equation 3.9 is used. Since the estimate inherits all the continuity and differentiability properties of the kernel, this means that in this case the final estimate has derivatives of all orders. These are available by differentiating equation 3.13 with respect to \( x \).

### 3.4 Verification with cross-validations

Since we apply the bootstrap to the data generated in the leave-one-out construction of the FROC curve, and kernel density estimation to smooth the resulting densities, we need to verify that this approach is legitimate. We do this with numerous cross-validations (CV) on the original data, using a Monte Carlo (MC) resampling scheme [15, 20]. The CVMC is in itself a method of estimating the effect of sampling error, yet considerably more time consuming. It may represent an answer closer to the truth. There are 2 000 runs in the resampling scheme. For each run the following is done:

1. Random partitioning: training set 35 cases, test set 10 cases.
2. Leave-one-out training scheme on the training set, 34+1, i.e., trained 35 times on 34 observations. The one observation left out gives a one or a zero when run through the system, which results in 35 ones and zeroes for each \( \theta \). Then \( \theta^* \) is the value of \( \theta \) that gives 90% ones.
3. Retrain the system with all 35 observations, so that the system is maximally trained.
4. Run the test set through the system with \( \theta^* \) set. This gives an estimate of \( fp \) per image for \( tp = 90\% \).

2 000 runs give 2 000 estimated values of \( fp \) per image given \( tp = 90\% \). The reduction in training sample size from 44 in the original leave-one-out training of the system to 35 in the CVMC should lead to greater variance of the error rate, but the 2 000 training sets are not independent since the same cases appear in them. This dependency leads to less variability in the distribution of the system operating point, and the two shortcomings seem to cancel out.
3.5 Simulations with the XFROC model

To further verify our suggested evaluation approach and to enable fair comparisons between parametric, semi-parametric and nonparametric versions, we need the true answer. From Chakraborty’s XFROC model [24], we simulate data that resembles the FROC data [20]. Let $x_{ij}$ denote the classification machine output from the $j$th noise site in the $i$th image. This is a false positive marking if $x_{ij}$ is larger than some threshold. Let $y_{ij}$ denote the output value from the $j$th signal site (cancer) in the $i$th image. $x_{ij}$ and $y_{ij}$ are drawn from two normal distributions

$$x_{ij} \sim N(\xi_i, \sigma_{LN})$$

$$y_{ij} \sim N(\psi_i, \sigma_{LS})$$

where the means $\xi_i$ and $\psi_i$ are themselves random variables:

$$(\xi_i, \psi_i) \sim N_2(0, \mu, \sigma_{CN}, \sigma_{CS}, \rho_{NS}).$$

That is, the positions of the location ($L$) distributions depend on the samples from the case ($C$) distribution. $N$ and $S$ stands for noise and signal, respectively.

The results from these simulations were in agreement with the results for the real-life data, and they suggest that the semi-parametric bootstrap smoothed conditional (BSC) interval, that is based on the proposed model in equation 3.3 with smoothing-after-bootstrap, allows a fair assessment of the performance of a CAD system on unseen data sets.
4. Summary of the papers

4.1 Paper I: Trends in breast cancer mortality among Swedish women 1953-92: analyses by age, period and birth cohort

Paper I is concerned with the modelling of breast cancer mortality trends. The trends are divided into three factors - age, period and cohort. We found that the mortality was remarkably stable over the study period. Age alone explains almost all of the mortality. A slight decrease was noted in the latest periods. As the magnitude of the effects associated with period and cohort were small, we did not specify restrictions in the full model that would have solved the identification problem, but reported the effects from the respective sub models. The improvement of adding effects to the sub models was tested.

The analysis of the effect of the change in coding practice in 1981 indicates a significant reduction of the likelihood of breast cancer being registered as a cause of death for the oldest age group.

Breast cancer mortality rates are fairly insensitive to changes in autopsy rates, since the proportion of all breast cancers found incidentally at autopsy has been below 1% since the mid-1970s.

The absence of mortality increase, even though there is a constant increase in incidence, may be explained by several factors. Besides earlier detection through mammography, possible factors are better survival due to surgical, radiological, adjuvant cytotoxic or hormonal treatment. However, since the divergence between incidence and mortality has been ongoing for several decades, other influences might be important, such as the occurrence of successively less aggressive tumours or an enhanced host defence against tumour cells. Considering the markedly varying pattern in incidence and mortality trends among countries, nutritional factors, e.g. calorie or fat intake, seem to be likely aetiological factors.

4.2 Paper II: The distributions of perceptions and expectations of inflation in Sweden

Paper II is included as an illustration of how kernel density estimation [26] is applied. The method is also used in papers III-V. The application is in econometrics, with the estimation of the distributions of perceptions and expecta-
tions of inflation in Sweden, and the distribution estimates are used to evaluate
deviations from the normal distribution.

We found that standardised skewness and kurtosis rejects the hypothesis of
the data coming from a normal distribution. The density estimates are skewed
to the right and clearly indicate that the data originates from a distribution that
is more peaked than the normal. The similarity of density estimates between
the perceived and expected inflation formed at the same time point indicates
that respondents seem to give approximately the same answer to both ques-
tions, though being a little more cautious in predicting the future. The truncation
of observations used by Statistics Sweden was found to be unreasonable
and we proposed the median of the density estimate as a better estimator of
what people act upon.

4.3 Paper III: On the comparison of FROC curves in
mammography CAD systems

In paper III, a model for the estimation of the distribution of the expected
number of false positive markings per image by a CAD system at a predeter-
mined sensitivity level is proposed and evaluated. The threshold distribution
is incorporated in the model. Two alternative procedures for estimating the
densities needed for the construction of the confidence interval are presented.
The first is based on the assumption of a Poisson distributed number of false
positive markings per image and in addition, on the assumption of indepen-
dence between false positives and the true positive indicator at each threshold
level. The second procedure uses the bootstrap applied to the data generated
in the leave-one-out construction of the FROC curve. The approach is verified
through cross-validation Monte Carlo.

4.4 Paper IV: Statistical aspects of threshold
independent performance assessment of mammography
CAD systems

In paper IV, we discuss mammography, CAD, the differences between ROC
and FROC methodology, the data structure, the scoring protocol, as well as
more on the model introduced in paper III. The assumptions underlying a para-
metric approach are tested. The hypothesis of a Poisson distributed number of
false positive markings per image is rejected with the bootstrap goodness-of-
fit test [28]. Furthermore, around those threshold values that are most likely to
produce 90% sensitivity, significant positive correlation (i.e. that finding the
tumour increases the risk of false positive markings in an image at a given
threshold) is discovered with point-biserial correlation estimation [29]. We
conclude that since the nature of the experiment changes with every image that is run through the CAD system, an assumption-free approach to the evaluation is preferable. The evaluation should not depend on a specific threshold, but should be done for all possible thresholds, thereby incorporating the effect of the sampling error. We call the intervals ‘threshold independent’.

4.5 Paper V: Evaluation of the bootstrap smoothed conditional (BSC) interval and others in small sample FROC analysis of mammography CAD system performance

In paper V, four confidence interval calculation methods are compared using original data and via simulations from the XFROC model [24]. The sample size needed to obtain an interval that is reasonably narrow is in focus. The bootstrapping is also further commented.

We found that the nonparametric bootstrap intervals are sensitive to the piecewise flatness of the FROC curve. When bootstrapping cases from the original sample, the flat curve tends to be preserved in the pseudo-samples, which results in bimodality and outliers in the upper region of the bootstrap estimate of the distribution of the mean false positives at the selected sensitivity. The performance of the nonparametric intervals is improved in the XFROC simulations, since the sample curves are somewhat smoother than the FROC curve based on the original sample. Our conclusion is that the model-based intervals are preferable for small sample FROC analysis, and in particular the bootstrap smoothed conditional interval (BSC), as suggested by both real-life data and simulations.

For larger samples, the simulations suggest undercoverage due to unbalance for the BSC approach, but this is not expected for real-life data. Furthermore, the BSC approach produces the shortest intervals. Thus, for larger samples, the result suggests that the assumption-free alternatives are the BSC or the BCa intervals. The smoothing seems more important than the bias reduction and acceleration adjustments present in the BCa interval.

A threshold independent interval gives an indication of whether more images should be used for the FROC curve construction. In the end, this would assure optimal performance of the CAD system in a clinical setting.
5. Bibliography


Acknowledgements

I am grateful to the Department of Statistics and the head of the department Professor Anders Christoffersson for my first two years as a PhD student, and to the new Department of Information Science with head Bo Wallentin for the finishing period. I am also thankful to the Department of Statistics and Operations Research, Stern School of Business, New York University, for my third year as a PhD student. Supervisor during the finalization of this thesis has been Professor Rolf Larsson and as my assisting supervisor, Ulf Olsson made administrative arrangements for my dissertation. Thank you both!

I was fortunate to work with the late Professor Reinhold Bergström, co-author on Paper I, during my first two years. Some of the things he said to me will always be cherished. His greatly missed. My gratitude also extends to my second co-author on Paper I, Professor Ingemar Persson and, in particular, to my brother Hans Bornefalk for the fruitful collaboration on Paper III.

I wish to thank my colleagues Fan Yang Wallentin, whose positive attitude and generosity mean very much to me, Lisbeth Hansson, who has been a great source of support throughout the years, Anna Gunsjö, for all the fun we had during the first period of my doctorate studies, Johan Lyhagen, Professor Adam Taube, Roland Pettersson, Bertil Andersson and Professor Anders Ågren. I also wish to express my appreciation to Gunilla Klaar and Ingrid Lukell for their help with administrative matters and for their kind support.

Encouragement from my best friends Camilla Douhan and Sonja Eaker, my mother Erica Johansson, my father Bengt Bornefalk and my sisters and brothers, especially Hans Bornefalk, Eva Bornefalk, Daniela Eklöf and Anders Bornefalk, my brother-in-law Gunnar Hermansson and my mother-in-law Gunborg Hermansson, who has helped me with numerous things for many years, is also gratefully acknowledged.

My dear husband Jarl Hermansson and my most beloved son Mattis Hermansson, thank you for all love and inspiration! When Mattis was five years old he was asked “What does your mother work with?” and he answered, “She reads and taps on her computer”. That pretty much sums it up!

Uppsala, November 6, 2006
Anna Bornefalk Hermansson
Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences 23

Editor: The Dean of the Faculty of Social Sciences

A doctoral dissertation from the Faculty of Social Sciences, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-7452