CHAPTER 1

OUR STEPS ON THE BICKEL WAY

Kjell Doksum, Ya'acov Ritov

Department of Statistics
1220 Medical Sciences Center
1300 University Ave Madison,
WI 53706
USA
E-mail: doksum@stat.wisc.edu

Department of Statistics
The Hebrew University of Jerusalem
Mt. Scopus, Jerusalem 91905
Israel
E-mail: yaacov@mscc.huji.ac.il

1. Introduction

Peter has been a leading figure in the field of statistics in the forty-three years since he received his Ph.D. in Statistics at the age of 22 from UC Berkeley in 1963 under the guidance of Erich Lehmann. His contributions are in education, research, and service to the profession and society. In education we find more than fifty students who have received their Ph.D. in statistics under Peter's guidance and a number of these are contributing to this volume. In addition we have found that doing joint work with Peter is a real educational experience.

Peter's many collaborators have benefited from his statistical insights and wisdom. Peter has also contributed to the statistical education of a large number of statistics Ph.D. students who took courses using his joint text book "Mathematical Statistics: Basic Ideas and Selected Topics", Bickel and Doksum (2000).

His professional service includes twice being chair of the Berkeley statistics department, twice being dean of the Physical Sciences, being director of the Statistical Computing Facility and being director of the Statistical Laboratory, all at UC Berkeley. It includes also being a member and chair of a number of national and international committees and commissions (including the National Research Council, the National Institute of Statistical Sciences, the National Academy of Science, the American Association for the Advancement of Science, and Council of Scientific Advisors, EURANDOM. He has been associate editor of the Annals of Statistics, Bernoulli, Statistical Sinica and the Proceedings of the National Academy of Sciences. He has received most of the honors in our profession including Guggenheim, NATO, Miller and the MacArthur Fellowships. He has received the COPSS prize, been Wald lecturere and has been president of the Institute of mathematical Statistics and the Bernoulli Society. In addition he is a member of the National Academy of Arts and Sciences and the Royal Netherlands Accademie of the Arts and Sciences.

In the rest of this piece we will focus on Peter's research contribution to statistics. We are not going to discuss all of Peter's contribution to statistics — there is too much of these, to too many fields, to be included in one survey. We are not going even to discuss all of his important work — we do not know them enough. The following is just some of Peter's work that *we* found interesting and influential (for us). The subject of this summary would be what *we* learnt from Peter, not on what can be learnt from him. Even so, it would be shorter than his work deserves.

2. Doing well at a point and beyond

The paper Bickel (1983) is not known very well. The toy problem is simple. We wish to estimate μ, the mean of a normal variable with variance one. We look for an estimator which behaves reasonably well whatever is the true value of the parameter, but we want it to behave really well at a given point, say 0. Peter translates this to the minimization of the loss at 0, subject to a bound on the loss else. A reasonable approximation to the estimator is given by $\hat{\theta}(X) = (X - c)\mathbf{1}(X > c) + (X + c)\mathbf{1}(X < -c)$. Peter generalizes the model to the multinormal distribution with an estimator that does well on a subspace, but this is still a toy model.

This however is an example of some of the Peter's research. It deals with a real big philosophical problem by leaving out non-essential elements and contributing important theoretical consideration.

The "standard" way to deal with the above mentioned problem is pre-testing: Test whether $\mu = 0$. If the null is accepted, estimate μ by 0, otherwise by x. This method lacks a rigorous justification, and the subject deserved a rigorous treatment. The claim of the paper is a soft-Bayesian. You should not ignore your assumptions. If you believe that the parameter belongs to a subset, you should use this fact. However, you should not ignore the possibility that you are wrong, and therefore you should use an estimator, that although does well on the specific subset, it should behave reasonably well everywhere else.

The resulted estimator resembles the pre-testing estimate. However, it is different. If the observation is greater than the threshold, then $X - c$ is used and not just X. Moreover, the level of the test is not arbitrary, but is a result of the imposed natural constraint.

Peter extends the ideas to practical frameworks in the paper Bickel (1984). Here he considers procedures that are "optimal" with respect to a given risk function over a submodel M_0 of models subject to doing "well" over a larger class M_1 of models. One of the problems considered is the familiar linear model problem where we are to choose between using a model M_0 with r parameters, or a model M_1 with s parameters, where $r < s$. He first solves the problem for known covariance matrices, then shows that the results hold asymptotically when these matrices are replaced by estimates for models where the parameter is of the Pitman form $\theta_n = \theta_0 + an^{-1/2}$ with $\theta_0 \in R^r$ and $a \in R^s$. In the case where the risk is mean square error, he finds that for a certain class of estimates, the asymptotic solution to the "optimal" at M_0 and good over M_1, formulation is a linear combination of the MLE's under the models M_0 and M_1, with the weights in the combination determined by a function of the Wald statistic for testing M_0 vs. M_1. Peter's idea of using parameters in the Pitman form is what makes it possible to derive asymptotic procedures and results for this difficult problem where inference and model selection problems are tackled together from a robustness point of view. More recently, Claeskens and Hjorth (2003) have used parameters in Pitman form to deal with estimation and model selection problems from the point of view of minimizing asymptotic mean squared error under the general model M_1. A comparison of these procedures with those of Peter is of interest.

3. Robustness, transformations, oracle-free inference, and stable parameters

Much of Peter's work is concerned with robustness, that is, the performance of statistical procedures over a wide class of models that typically are semiparametric models that describe neighborhoods of ideal models. The frameworks considered include one and two sample experiments, regression experiments, and time series as well as general frameworks. In the case of Box and Cox (1964) transformation regression models where a transformation $h(Y;\lambda)$ of the response Y follow a linear regression model with coefficient vector β, error variance σ^2, transformation parameter λ and error distribution F, Peter considered in the joint work Bickel and Doksum (1981) robust estimation of the Euclidean parameters for semiparametric models where F is general. In particular, they established the extra variability of the estimate of β due to the unknown λ. This work was controversial, cf. Box and Cox (1982) and Hinkley and Runger (1984), for a brief period, because it seemed that the coefficient vector β depends on λ and results that depend on a model that assumes that lambda is unknown were claimed to be "scientifically irrelevant". Fortunately the coefficient vector has an intuitive interpretation independent of λ and there is no need to depend on an oracle to provide the true λ: Brillinger (1983) shows how to do inference when a transformation is unknown, by showing that the parameter $\alpha = \beta/|\beta|$, where $|\ |$ is the Euclidean norm, is identifiable and independent of the transformation. It has the interpretation of giving the relative importance of the covariates provided they have been standardized to have SD's equal to one. He also provided $\sqrt{n}$ inference for α and gave the inflation in variability due to the unknown transformation. These ideas and results were developed further by Stoker (1986) and lead to the field of index models in statistics and econometrics as well as the concept of stable parameters,e.g. Cox and Reid (1987), Doksum and Johnson (2002). Peter, Bickel and Ritov (1997), developed asymptotically efficient rank based estimates of alpha for general transformation regression models with increasing transformations, a project that had been approximated by Doksum (1987).

4. Distribution free tests, higher order expansions, and challenging projects

Peter is not one to shy away from challenging problems. One of the most challenging projects ever carried out in statistics was Peter's joint work Bickel and van Zwet (1978) on higher order expansions of the power of distribution free tests in the two sample case and the joint work Albers, Bickel and van Zwet (1976) in the one sample case. They considered situations where the $n^{-1/2}$ term in the asymptotic expansion of the power is zero and n^{-1} term is required. In a classic understatement they wrote"the proofs are a highly technical matter". They used their results to derive the asymptotic Hodges-Lehmann deficiency d of Pitman efficient distribution free tests with respect to optimal parametric tests for given parametric models. Here d is defined as the limit of the difference between the sample size required by a Pitman efficient distribution free test to reach a given power to the sample size required by the optimal parametric test to reach the same power. They found that in normal models, the optimal permutation tests have deficiency zero, while this is not true for locally optimal rank(normal scores) tests.

5. From adaptive estimation to semiparametric models

In his 1980 Wald lecture, Bickel (1982), Peter discussed the ideas of Stein (1957) on adaptive estimation. It was already well established (Beran (1974),Stone (1975),Sacks (1975)) that adaptive estimation is possible for the estimation of the center θ of symmetric distribution on the line. That is, asymptotically θ can be estimated as well when the density f of $x - \theta$ is unknown as when f is known. Peter discussed in his paper the conditions needed to ensure that adaptive estimation is generally possible. The meeting with Jon Wellner resulted in extending the scope to the general semiparametric model. A project of almost 10 years started in which the estimation in the presence of non-Euclidean nuisance parameters was discussed and analyzed to the fine details. It included a general analysis, mainly in Bickel, Claassen, Ritov and Wellner (1993), and discussion of specific models.

The statistical models considered in the book and the relevant papers were interesting themselves, but almost all of them presented a more general issue. They were chosen not only because they had interesting application, but because they were fitted to present a new theoretical aspect.

Thus the title of Bickel and Ritov (1988) is "Estimating integrated squared density derivatives". Although the estimation of the integral of the square of the density can be motivated, and was done in the past, this was not the reason the paper was written. Information bounds in regular parametric model are achievable. That is, there are estimators which achieve these bounds. There was a conjecture that the same is true also for semiparametric models. Before this paper, different estimators for different parameters and models were presented which achieved the information bound, but always more conditions were needed in the construction of the estimator section, than were needed to establish the bounds. An example was needed to clarify this point, and in Bickel and Ritov (1988) it was shown that if not enough smoothness is assumed, there may be rate bounds which are not even of the $\sqrt{n}$ rate. The title of Bickel and Ritov (1990) was less modest and present a more general claim (although the paper just gave some more examples): "Achieving information bounds in semi and non parametric models."

Similarly Bickel and Ritov (1993) dealt with a relatively minor situation (although, it generalizes nicely to deal with many censoring situations). The title is "Efficient estimation using both direct and indirect observations.". The interest in this problem was in effect different. The situation is thus that the information bound for the parameter of interest using only the indirect observation is 0. The question asked was whether a sub-sample which seeming carries no information is important when it can be combined with an informative sample. The surprising answer was yes.

The book by Bickel, Klaassen, Ritov and Wellner on efficient and adaptive estimation was a real project, which ended in 1993. It needed careful and tedious work. The effort was to cover all relevant aspects of semiparametric models, from information bounds to efficient estimation of both the Euclidean and non-Euclidean parameters.

The book dealt only with i.i.d. observations. Bickel and Kwon (2001) extended the ideas beyond the i.i.d. model. In this paper the authors formulate a 'calculus' similar to that of the i.i.d. case that enables them to analyze the efficiency of procedures in general semiparametric models when a nonparametric model has been defined. The extension includes regression models (in their functional form), counting process models in survival analysis and submodels of Markov chains, which traditionally require elaborate special arguments.

6. Hidden Markov Models

The hidden Markov model (HMM) is a simple extension of the parametric Markov model. The only difference is that we have noisy observations of the states of the chain. Formally, let $X_1, X_2, \ldots, X_n$ be an unobserved Markov chain with a finite state space. Let $Y_1, Y_2, \ldots, Y_n$ be independent given the Markov chain, where the conditional distribution of Y_i given $X_1, X_2, \ldots, X_n$ depends only on X_i and maybe on an unknown parameter $\theta \in \Theta \subseteq R^d$.

After the long time spent on semi-parametric models, this seemed to be just a simple parametric model. The theory of regular parametric models for Markov chains is a natural extension of that of the i.i.d. case. HMMs looked to be the next natural step, and developing their theory seemed to be an interlude in the drama of semi-parametric research. It was not. The model needed serious work. The first paper, Bickel and Ritov (1996) was a long tedious analysis with more than 20 lemmeta, which resulted at the end with a too weak result (the LAN condition is satisfied for the HMM models). What was needed was the help of Tobias Rydén, who could write a one pages formula without a mistake, Bickel, Ritov and Rydén (1998). The third paper, Bickel, Ritov and Rydén (2002), was elegant, if its complicated notation is deciphered.

7. Non- and semi-parametric Testing

The paper Bickel and Rosenblatt (1973) is seemingly about density estimation. But it set the stage to our further work on testing. Limit theorems are obtained for the maximum of the normalized deviation of the density estimate from its expected value, and for quadratic norms of the same quantity. This sets the stage for chi-square type of tests and for a consideration of local large deviation in some unknown location from the null hypotheses. Ait-Sahalia, Bickel and Stoker (2001), applied these ideas to regression models and econometrical data.

There is a real theoretical problem with testing. It is not clear what can be really achieved in non-parametric testing: should the researcher be greedy and look for (almost) every thing? That is, all deviations are considered equally likely. The price for such an omnipotent test, is having an impotent test—no real power in any direction. In particular deviations on the $\sqrt{n}$ level cannot be detected. The alternative is looking for tests which concentrate their power in a restricted set of directions. The latter test would typically be able to detect deviation from the null in all directions on the $\sqrt{n}$ scale, but their power would be mostly minor. However, in a few directions they would be powerful. Thus we meet again the theme of doing well at a point. Bickel, Ritov and Stoker (2005a),Bickel, Ritov and Stoker (2005b) consider this problem, argue that there is no notion of optimal or efficient test,and the test to use should be Taylor made to the problem at hand.

In a different direction, Bickel and Bühlmann (1997) argue that given even an infinitely long data sequence, it is impossible (with any test statistic) to distinguish perfectly between linear and nonlinear processes. Their approach was to consider the set of moving-average (linear) processes and study its closure. The closure is, surprisingly very large and contains nonergodic processes.

8. The road to real life

Much of Peter's work is theoretical. There are real world examples, but they are just that, examples. In recent years, however, Peter devoted much of his time to real world projects, where his main interest was in the subject matter per se, and not as a experimental lab for statistical ideas.

The first field to consider was traffic analysis. The first problem was travel time estimation using loop detector micro-data, Petty, Bickel, Kwon, Ostland, Rice, Ritov and Schoenberg (1998). In Kwon, Coifman and BIckel (2000), the estimation of future travel time was considered. The estimation use as input flow and occupancy data on one hand and historical travel-time information on the other hand.

The second field which really fascinates Peter is molecular biology. The practical work of Peter as a statistician in the field is in topics like motif discoveries, Kechris, van Zwet E., Bickel and Eisen (2004), important sites in protein sequences, Bickel, Kechris, Spector, Wedemayer and Glazer (2003), or finding critical structural features of HIV proteins as targets for therapeutic intervention, Bickel, Cosman, Olshen, Spector, Rodrigo and Mullins (1996).

References

Ait-Sahalia, Bickel and Stoker (2001). Ait-Sahalia, Y., Bickel, P. J., and Stoker, T. M. (2001). Goodness-of-fit tests for kernel regression with an application to option implied volatilities. *J. Econometrics*, **105**, 363–412.

Albers, Bickel and van Zwet (1976). Albers, W., Bickel, P. J., and van Zwet, W. R. (1976). Asymptotic expansions for the power of distribution free tests in the one-sample problem. *Ann. Statist.*, **4**, 108–156.

Beran (1974). Beran, R. (1974). Asymptotically efficient adaptive rank estimates in location models. *Ann. Statist.*, **2**, 63–74.

Bickel (1982). Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.*, **10**, 641–671.

Bickel (1983). Bickel, P. J. (1983). Minimax estimation of the mean of a normal distribution subject to doing well at a point. In *Recent advances in statistics.* Academic Press, New York.

Bickel (1984). Bickel, P. J. (1984). Parametric robustness: small biases can be worthwhile. *Ann. Statist.*, **12**, 864–879.

Bickel and Bühlmann (1997). Bickel, P. J. and Bühlmann, P. (1997). Closure of linear processes. *J. Theoret. Probab.*, **10**, 445–479.

Bickel, Claassen, Ritov and Wellner (1993). Bickel, P. J., Claassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, volume Reprinted 1998. Johns Hopkins University Press, Baltimore.

Bickel, Cosman, Olshen, Spector, Rodrigo and Mullins (1996). Bickel, P. J., Cosman, P. C., Olshen, R. A., Spector, P. C., Rodrigo, A. G., and Mullins, J. I. (1996). Covariability of v3 loop amino acids. *AIDS Res Hum Retroviruses*, **12**, 1401–1411.

Bickel and Doksum (1981). Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited. *J. Amer. Statist. Assoc.*, **76**, 296–311.

Bickel and Doksum (2000). Bickel, P. J. and Doksum, K. A. (2000). *Mathematical Statistics: Basic Ideas and Selected Topics*, volume I. Prentice Hall, Upper Saddle River.

Bickel, Kechris, Spector, Wedemayer and Glazer (2003). Bickel, P. J., Kechris, K., Spector, P. C., Wedemayer, G. J., and Glazer, A, N. (2003). Finding important sites in protein sequences. *Proceedings of The National Academy of Sciences*, **99**, 14764–14771.

Bickel and Kwon (2001). Bickel, P. J. and Kwon, J. (2001). Inference for semiparametric models: some questions and an answer. *Statist. Sinica*, **11**, 863–960.

Bickel and Ritov (1988). Bickel, P. J. and Ritov, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhya*, **A50**, 391–393.

Bickel and Ritov (1990). Bickel, P. J. and Ritov, Y. (1990). Achieving information bounds in non and semiparametric models. *Ann. Statist.*, **18**, 925–938.

Bickel and Ritov (1993). Bickel, P. J. and Ritov, Y. (1993). Efficient estimation using both direct and indirect observations. (russian); english version in theory probab. appl. 38 (1993), no. 2, 194–213. *Teor. Veroyatnost. i Primenen.*, **38**, 233–238.

Bickel and Ritov (1996). Bickel, P. J. and Ritov, Y. (1996). Inference in hidden markov models. i. local asymptotic normality in the stationary case. *Bernoulli*, **2**, 199–228.

Bickel and Ritov (1997). Bickel, P. J. and Ritov, Y. (1997). Local asymptotic normality of ranks and covariates in transformation models. In D. Pollard, E. Torgerson, and G. Yang (Eds.), *Festschrift for Lucien Le Cam*. Springer, New York.

Bickel, Ritov and Rydén (1998). Bickel, P. J., Ritov, Y., and Rydén, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden markov models. *Ann. Statist.*, **26**, 1614–1635.

Bickel, Ritov and Rydén (2002). Bickel, P. J., Ritov, Y., and Rydén, T. (2002). Hidden markov model likelihoods and their derivatives behave like i.i.d. ones. *Ann. Inst. H. Poincaré Probab. Statist.*, **38**, 825–846.

Bickel, Ritov and Stoker (2005a). Bickel, P. J., Ritov, Y., and Stoker, T. (2005a). Nonparametric testing of an index model. In D. W. K. Andrews and

J. H. Stock (Eds.), *Identification and Inference for Econometric Models:A Festschrift in Honor of Thomas J.Rothenberg*. Cambridge University Press, Cambridge.

Bickel, Ritov and Stoker (2005b). Bickel, P. J., Ritov, Y., and Stoker, T. (2005b). Tailor-made tests for goodness-of-fit to semiparametric hypotheses. *Ann. Statist.*

Bickel and Rosenblatt (1973). Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.*, **1**, 1071–1095.

Bickel and van Zwet (1978). Bickel, P. J. and van Zwet, W. R. (1978). Asymptotic expansions for the power of distribution free tests in the two-sample problem. *Ann. Statist.*, **6**, 937–1004.

Box and Cox (1964). Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. (with discussion). *J. Roy. Statist. Soc. Ser. B*, **26**, 211–252.

Box and Cox (1982). Box, G. E. P. and Cox, D. R. (1982). Comment on: "an analysis of transformations revisited". *J. Amer. Statist. Assoc.*, **77**, 209–211.

Brillinger (1983). Brillinger, D. R. (1983). A generalized linear model with "gaussian" regressor variables. In P. J. Bickel, J. L. Doksum, and J. Hodges (Eds.), *A Festschrift for Erich L. Lehmann*. Wadsworth, Belmont.

Claeskens and Hjorth (2003). Claeskens, G. and Hjorth, N. L. (2003). The focused information criterion. *J. Amer. Statist. Assoc.*, **98**, 900–916.

Cox and Reid (1987). Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. with a discussion. *J. Roy. Statist. Soc. Ser. B*, **49**, 1–39.

Doksum (1987). Doksum, K. A. (1987). An extension of partial likelihood methods for proportional hazard models to general transformation models. *Ann. Statist.*, **15**, 325–345.

Doksum and Johnson (2002). Doksum, K. A. and Johnson, R. (2002). Comments on "box-cox transformations in linear models: large sample theory and tests of normality" by chen, g. and lockhart, r. a. and stephens, m. a. *Canad. J. Statist.*, **30**, 177–234.

Hinkley and Runger (1984). Hinkley, D. V. and Runger, G. (1984). The analysis of transformed data. *J. Amer. Statist. Assoc.*, **79**, 302–320.

Kechris, van Zwet E., Bickel and Eisen (2004). Kechris, K. J., van Zwet E., Bickel, P. J., and Eisen, M. B. (2004). Detecting dna regulatory motifs by incorporating positional trends in information content. *Genome Biology*, **5**, R50.

Kwon, Coifman and BIckel (2000). Kwon, J., Coifman, B., and BIckel, P. J. (2000). Day-to-day travel-time trends and travel-time prediction from loop-detector data. *Transportation Research Record*, **1717**, 120–129.

Petty, Bickel, Kwon, Ostland, Rice, Ritov and Schoenberg (1998). Petty, K. F., Bickel, P. J., Kwon, J., Ostland, M., Rice, J., Ritov, Y., and Schoenberg, F. (1998). Accurate estimation of travel times from single-loop detectors. *Transportation Research: Part A—Policy and practice*, **32**, 1–17.

Sacks (1975). Sacks, J. (1975). An asymptotically efficient sequence of estimators of a location parameter. *Ann. Statist.*, **3**, 285–298.

Stein (1957). Stein, C. (1957). Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I, pp. 187–195*, Berkeley and Los Angeles. University of California Press.

Stoker (1986). Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica*, **54**, 1461–1481.

Stone (1975). Stone, C. J. (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist.*, **3**, 267–284.