FINAL REPORT

on BIRS 5-day workshop *Regularization in Statistics*

Ivan Mizera and Roger Koenker

1. Overview

An annoying feature of the real world is that it often expects computational solutions to problems whose mathematical formulation clearly indicates that such efforts should be hopeless. Even worse, such situations seem to be the rule rather than the exception.

Applied mathematicians, especially those with interests in partial differential equations, recognized this fact early on. Following Hadamard's (1902) formulation, the idioms "ill-posed", followed by "inverse problem", have gradually become a staple of mathematical jargon. Not satisfied by simply naming the problem, a few intrepid souls began to look for approaches to deal with it. This initial phase culminated in the work of Tikhonov (1963) and his school, and independently with the work of Phillips (1962). Their numerous followers have continued to cultivate the regularization approach.

As often happens with problems urgently needing solutions, there were parallel efforts throughout the archipelago of mathematical thought. The recipe of Whittaker (1921) for smoothing mortality tables in actuarial science was noticed by Schoenberg (1964) and by Parzen (1961), and led subsequent authors toward nonparametric function estimation in reproducing kernel Hilbert spaces—allowing prior notions about the smoothness of target functions to be expressed as roughness penalties in the regularization framework. A broader perspective brings the recognition of the fundamental role regularization has played in modern statistics: Stein's (1956) work in decision theory on admissibility of least squares methods demonstrated the value of so-called shrinkage procedures. And Bayesian methods were quick to be extended and reinterpreted in this light. Recent computational advances have exerted a profound influence over the development of these ideas.

Thus it can be tentatively concluded that researchers interested in regularization recruit from three—at least—main groups: mathematicians with applied flavor, statisticians and similar individuals concerned with data analysis, and computing scientists. The crucial fact, however, is that, modulo technicalities and discipline-specific jargons, all of the above observe the fundamental principle. In the simplest instance the regularization paradigm seeks an acceptable solution to the linear system

$$Ax = b. \tag{1}$$

The inverse of A exists, but A is known to be ill-conditioned. Optimization of the penalized quadratic form,

$$\min_x \|Ax - b\|^2 + \lambda\phi(x). \tag{2}$$

may be considered. Again, in the simplest instance $\phi(x)$ may be taken to be $\|x\|^2$, leading to solutions of the form,

$$x = (A'A + \lambda I)^{-1}A'b. \tag{3}$$

It should to be stressed that many variations of this theme are possible. The entities may be infinite-dimensional, necessitating operators not matrices; the quadratic measure of lack-of-fit may be altered, and the penalty $\phi(x)$ may take many forms in an effort to accurately represent auxiliary information about solutions of the problem that was not explicitly incorporated into the original formulation. In the end, it is also crucial to choose some value of the regularization parameter λ and finally find efficient methods for solving (2). In statistics the leading problems of this type involve nonparametric regression and density estimation, solution of inverse problems leading to integral equations, and deconvolution, but a vast array of other problems from imaging, machine learning, and other regions of applied mathematics abound.

Emerging problems have led to novel modifications of the regularization approach. The scope has widened considerably, while some old problems still are not solved completely. In image processing, for example, edge detection and image segmentation objectives have stimulated work on total variation forms of the $\phi(x)$ regularization penalty. A common denominator with, say, classification methods used in machine learning, or nonparametric regression problems in statistics is nonlinearity; the latter brings challenges not only in computing, but above all in selection of regularization parameter and also in the subsequent theoretical interpretation. Advances in numerical analysis have played a crucial role in these developments.

These considerations taken together suggested that the time was ripe for a meeting of specialists from a broad range of fields to discuss recent developments in regularization. Our ambition was not to organize yet another meeting of a well established network of people whose work was already mutually familiar, but to bring together as diverse body of participants as possible, while maintaining a statistical core of the meeting—an understandable constraint given our professional orientation. An approach like this inevitably brings certain risks: rather than relying on people from our own circles, people we know well from regular meetings and have a fairly precise estimates what to expect of, we actively sought out new personalities—people from other circles of interests, with their own priorities and value systems. In particular, it was by no means clear whether the reward our meeting may have from them would be at least comparable to that it would have for us, statisticians—and therefore carried an inherent risk that we would be not be able to attract a wide enough circle of participants to realize our objectives.

2. The Workshop

The outcome, however, exceeded our most optimistic expectations. Despite some unlucky coincidences with other professional meetings, which could not be anticipated, we were fortunate to assemble—quoting from a recent email—"a fabulous collection of participants." The workshop brought together 34 participants from statistics and allied fields, including imaging, machine learning, numerical analysis, and applied mathematics. Departmental affiliations of the participants included mathematicians of both "pure" and "applied" flavor, computer scientists and electrical engineers, as well as a broad collection of statisticians from mathematics, statistics, biostatistics, economics, and psychology departments.

Inspired partly by Oberwolfach traditions, we decided to deal with program issues quite informally. The submission of abstracts and titles of talks was voluntary. Not everybody was automatically assumed to present a talk; this choice was left to participants themselves—and by no means could it be said that those opting for a discussant rôle remained silent. The decision to compose the program "on the fly" brought us some anxious moments at the very beginning of the workshop, but in the broader perspective it was instrumental in realizing our vision of a meeting in which talks are not only delivered but also listened to, a workshop as a community, not a railway station where people and trains come, stop, and go (we pursued this objective even at the cost of losing several desirable speakers).

Although some of the participants knew each other before—one could notice a group of statisticians with interests close to smoothing and functional data analysis, and a group of mathematicians working differential-equations approach to image processing—it was fascinating was to see people from different areas interact; speakers were often not addressing their peers, but rather the participants from other groups.

3. Day first: Sunday

The workshop started by a short opening address of Robert Moody, FRSC, the scientific director of the BIRS. This was immediately followed by the scientific program.

The opening talk by Rudy Beran "Regularized Bayes fits to incomplete unbalanced multi-way layouts" provided an ideal starting point for the workshop. Beran outlined how low-risk adaptive Bayes fits to large discrete multi-way ANOVA layouts provide fits whose risk converges to the minimum risk attainable over the candidate class, as the number of factor levels tends to infinity. Both ordinal and nominal factor levels, possibly unbalanced or incomplete, were considered. He also illustrated in a broader context how modern statistical computing provides a technological environment for advances in statistical thought beyond those supported by probability theory. Since the method employed are equivalent to regularization by penalized

least squares, with smallest estimated risk used to select the penalty parameter, the talk brought everyone together to the core of the regularization problem.

The next talk "Learning with determined inputs and generalized Shannon-Nyquist sampling", by Stephen Smale, discussed extensions of classical Shannon-Nyquist sampling in the context of regularization and machine learning. It was, for most participants, a new perspective. Smale's traditional chalk-and-blackboard style created an immediate rapport with the predominantly statistical audience, and a vigorous discussion ensued. We were delighted to see how quickly participants adapted to this informal style, which set a good "workshop" atmosphere for the subsequent sessions.

The morning session concluded with the talk "Regularization in statistics: metamorphoses and leftovers" of one of organizers, Ivan Mizera. The talk essentially aimed at outlining some issues that were to follow: trying, in the simplest regularization setting of nonparametric regression to indicate similarities and differences between various possible approaches. Topics addressed included the difference between dense (functional, gridded) and scattered (point) data formulation; various motivations for nonlinearity and how they relate to linear problems; certain bivariate aspects - the role of the domain and subsequent implications for the numerical developments; the classical desiderata for smoothing versus emerging alternative objectives.

The afternoon session was devoted to several areas of application. Enno Mammen in his talk "Optimal estimation in additive regression models" discussed optimal estimation of a additive nonparametric components. He discussed the additive regression model and showed that up to the first order an additive component can be estimated as if the other components were known, proving this claim for kernel smoothers, local polynomials, smoothing splines and orthogonal series estimators.

A glimpse into econometrics was brought by Joel Horowitz in his talk on "Nonparametric estimation in the presence of instrumental variables". He suggested two nonparametric approaches, based on kernel methods and orthogonal series, respectively, to estimating regression functions involving instrumental variables. For the first time in this class of problems optimal convergence rates were derived, and showed that they are attained by particular estimators.

Paul Speckman's talk on "Adaptive function estimation using smooth stochastic variability priors" considered Bayesian nonparametric regression, with priors on the unknown function corresponding to smoothing with L-splines, where the penalty term also includes a weight function that is optimized to give a locally adaptive smoother. He showed that appropriate priors can be constructed that are similar to stochastic volatility models in finance indicating their usefulness in nonparametric regression and also density estimation.

Finally, C. Samuel Kou in his talk "Statistical analysis of single molecule experiments in chemistry" spoke about statistical challenges connected with

the recent technological advances allowing scientists to follow biochemical process on a single molecule basis.

4. Day second: Monday

The second day began with the talk of Emmanuel Candés on "Chirplets: Multiscale detection and recovery of chirps". He considered the problem of detecting and recovering chirps—signals that are neither smoothly varying nor stationary but rather exhibit rapid oscillations and rapid changes in their frequency content. Such behavior is quite different than the traditionally notions of smoothness and homogeneity for noisy data. Building on recent advances in computational harmonic analysis, he designed libraries of multiscale chirplets, and introduced detection strategies that are more sensitive than existing feature detectors. Structured algorithms that exploit information in the chirplet dictionary are used to chain chirplets together adaptively so as to form chirps with polygonal instantaneous frequency; these structured algorithms are so sensitive that they allow to detect signals whenever their strength makes them detectable by any method, no matter how intractable. An applications to the detection of gravitational waves was discussed.

The two other morning talks were given by mathematicians who have made significant contributions to the rigorous analysis of total-variation methods for imaging. Antonin Chambolle in "Total variation minimization–an algorithm for total variation minimization and applications" presented an algorithm for minimizing total variation under a quadratic constraint, based on the dual formulation and discussed its applications to image processing.

Otmar Scherzer in "Tube methods for bounded variation regularization" used statistical modeling for developing regularization models for de-noising, dejittering, and de-blurring applications in image processing. Inspired by connections to the taut-string algorithm for total variation optimization in one dimension, analogous methods were proposed for image processing in dimension two. This provided a nice link to more statistically motivated discussion of taut string methods by Arne Kovac and Laurie Davies later in the conference.

In the afternoon, the second of the organizers, Roger Koenker spoke about "Total variation regularization for noisy, scattered data." He described two variants of total variation regularization for bivariate function estimation: piecewise constant functions on Voronoi tesselations, or voronograms, and piecewise linear functions on Delaunay tesselations, or triograms. The accent was on efficient computing using primal-dual interior point methods and the sparsity of the underlying linear algebra enabling quite large problems to be solved. This theme was carried forward by the computational character of the other afternoon talks.

Arne Kovac spoke on "Taut strings and modality", and addressed the vital question of alternative objectives. The usual goal in nonparametric regression and density estimation is to specify a function f that adequately represent the data, but do not contain spurious local extremes. Multiresolution versions of the taut string method were proposed to generate a sequence of functions with increasing number of local extreme values; the close connection to total variation methods was also emphasized.

Robert Nowak spoke on "Near minimax optimal learning with dyadic classification trees." He reported on a family of computationally practical classifiers that converge to the optimal (Bayes error) classifier at near-minimax optimal rates for a variety of distributions. The classifiers are based on dyadic classification trees, which involve adaptively pruned partitions of the feature space, whose key aspect is their spatial adaptivity, enabling local (rather than global) fitting of the decision boundary. Their risk analysis involves a spatial decomposition of the usual concentration inequalities, leading to a spatially adaptive, data-dependent pruning criterion. The classifiers are practical and the same time provide rate of convergence within a logarithmic factor of the minimax optimal rate.

Finally, Sylvain Sardy spoke about "L_1 penalized likelihood nonparametric function estimation" and discussed the relationship between Laplace Markov random fields and total variation regularization for regression, density estimation and linear inverse problems. He proposed a relaxation algorithm to solve a dual optimization problem and considered ways of choosing the regularization parameter λ—contributing thus to the final characterization of Tuesday as a "total-variation day".

On Monday evening, we decided to hold Round-table discussion on λ selection. How salient this topic is is illustrated by the fact that the question "How do you select λ?" appeared in the discussion to nearly every talk at the workshop. The discussion was led in a very exquisite manner by Grace Wahba. Since we consider this event one of the high moments of the meeting, we decided to capture its elusive character on the video recording. Once available on web, we believe that there will be quite demand for this *coup de theatre*.

5. Day third: Tuesday

Chong Gu in his talk "Model Diagnostics for Penalized Likelihood Estimates" spoke about functional ANOVA decompositions that can be incorporated into multivariate function estimation through penalized likelihood methods. He presented some simple diagnostics for the "testing" of selected model terms in the decomposition; the elimination of practically insignificant terms generally enhances the interpretability of the estimates, and sometimes may also have inferential implications. The diagnostics were illustrated in the settings of regression, density estimation, and hazard model estimation.

Dennis Cox's talk "Functional-data ANOVA via randomization based multiple comparisons procedure" considered an approach to a functional-data ANOVA using univariate ANOVA methods pointwise on a grid and applying the randomization based multiple comparisons method of Westfall and Young to obtain a desired significance level. This method is relatively straightforward and gives very interpretable results (regions in the space of the independent variable where there are significant differences), but there have been concerns that there are potential problems. Cox conjectured that the Westfall and Young procedure in fact overcomes this problem and as the grid becomes finer, the corrected p-values converge to an approximately continuous function and gave some motivation and numerical examples in support of this conjecture.

"Bayesian logic regression" by Charles Kooperberg addressed fitting regression models of single nucleiotide polymorphism data: using many, essentially binary, covariates. MCMC is used on the class of regression models of interest, and rather than summarizing all models by, say, a mean all the MCMC models are summarized in a qualitative way.

Finally, in the talk "From Data to Differential Equations Differential equations" Jim Ramsay argued that regularization penalties could be adapted to differential equations representing the underlying processes generating observed functional data, and as such offer a number of potential advantages over conventional parametric or basis expansion models. They explicitly model the behavior of derivatives, and derivative estimates based on them are often superior to those derived from conventional data smoothers; they have the capacity to model curve-to-curve variation as well as within subject variability and the known structural features can be built into them more easily than is usually the case for conventional functional models. Finally they offer a wider range of ways to introduce stochastic behavior into models. Some illustrations of the performance of these methods were given from process control in chemical engineering and for medical data on treatment regimes for lupus.

Tuesday morning underscored the remarkable achievements of the classical, quadratic, Hilbert-space based approach to regularization in statistics; Tuesday afternoon was originally planned to be free. Adverse weather, however, altered this plan, and the program was adaptively restructured to shift the free afternoon on Wednesday, in the hope of improved weather.

The afternoon session began with Selim Esedoglu's talk on "Decomposition of images by the anisotropic Rudin-Osher-Fatemi model." He reported on generalizations of total variation based image de-noising model of Rudin, Osher, and Fatemi, designed to privilege certain edge directions. He consider the resulting anisotropic energies and studied properties of their minimizers.

Frits Ruymgaart's talk "Improving regression function estimators in indirect models" explained that the traditional estimators of linear functionals of the regression function in general, and of their Fourier coefficients in particular, are not asymptotically efficient in the sense of LeCam-Hájek-van der

Vaart, except, possibly, when the error distribution is normal. In particular, when repeated measurements are available, an improvement procedure can be carried out for error distributions with heavy tails, starting with a preliminary estimator based on linear combinations of the order statistics in the subsamples.

Finally, Curt Vogel spoke on "A Wavefront reconstruction problem in adaptive optics." He reviewed the basic concepts behind the adaptive optics, which can dramatically improve the resolution of ground-based optical telescopes. He described in some detail the computation of the wavefront reconstructor, the mapping from sensor measurements to mirror deformations, and highlighted its connection to statistical regularization.

The evening was devoted to a social program event—the song recital of a mezzosoprano Kathryn Whitney, a native of Calgary recently elected Creative Arts Fellow in Music at Wolfson College of the University of Oxford. The location of BIRS within Banff Centre greatly facilitates the organization of events like this. BIRS manager Andrea Lundquist was very instrumental in making the event possible; we owe the Department of Music our deepest thanks for providing the venue and technical support. We were glad to invite to this event the members of the focused research group on the "Arithmetic of fundamental groups ", with whom we shared BIRS, and were delighted to see that they enjoyed the concert with us.

6. Day fourth: Wednesday

The morning session started with an illuminating overview "The method of moments and statistical computation", by Gene Golub. The talk brought out many stimulating ideas related to the fundamental issues of implementation of every single method mentioned in the workshop—computational linear algebra.

Werner Stuetzle then spoke about "Spline smoothing on surfaces", presenting a method for estimating functions on topologically and/or geometrically complex surfaces from possibly noisy observations. An extension of spline smoothing, using a finite element method. The results of an experiment comparing finite element approximations to exact smoothing splines on the sphere were shown, and examples suggesting that generalized cross-validation is an effective way of determining the optimal degree of smoothing for function estimation on surfaces were given.

The rest of the morning belonged to another two representatives of the vital UCLA group. Jianhong Jackie Shen spoke about "A young mathematician's reflection on vision: Illposedness and regularizations" giving an overview of the topological, geometric, and statistical/Bayesian regularization techniques of recent work on visual perception.

Finally, Tony Chan in "Geometric and total variation regularization in inverse problems" reviewed recent advances in the area of inverse problems in which one wishes to recover functions that are piecewise smooth separated by lower

dimensional interfaces. For these problems, it is important to choose a regularization technique that will respect and preserve the discontinuities of the function values, as well as control the geometric regularity of the interfaces. Notable examples of this class of regularization include minimizing the total variation of the function and minimizing the surface area of the interfaces; applications include image restoration and segmentation, elliptic inverse problems, and medical image tomography problems.

The afternoon was free—and fortunately the weather improved, so participants could also enjoy the marvelous Rocky Mountain surrounding of the Banff Centre.

7. Day fifth and last: Thursday

In the first morning talk on "Oracle inequalities for regularized estimators", Sara van de Geer emphasized the special role of the L_1 penalty. She investigated least squares estimators with general penalties, using smoothing splines as special case and examined which penalties achieve an oracle inequality for the estimator. The latter results showed that the L_1 penalty adapts to both the smoothness as well as to possibly non-quadratic margin behavior.

Laurie Davies talk "Approximating data" considered the basic idea that a model can be taken as an adequate approximation for a data set if typical samples generated under the model share the essential features of the data set. This somewhat loose formulation was made precise and related problems of topologies in data analysis and inference were discussed. Examples of the idea of data approximation involve nonparametric regression, densities and financial data.

The workshop concluded by the talk of Grace Wahba on "The multicategory support vector machine and the polychotomous penalized likelihood estimate". She described two modern methods for statistical model building and classification, penalized likelihood methods and support vector machines, both of which are obtained as solutions to optimization problems in reproducing kernel Hilbert spaces. Two category support vector machines are very well known, while the multi-category ones were introduced in the talk: they include modifications for unequal misclassification costs and unrepresentative training sets, is new. The applications of each method were described, in a demographic study and meteorology.

8. Final thoughts

The talks spanned a broad spectrum of both theoretical and applied work on regularization. The enthusiastic response of the participants, including, but not limited to their emails after the conference, fully justified our initial confidence in the timeliness of the workshop and its success. In particular, we are confident that the workshop accomplished its main objective of cross-fertilization. Some representatives of the L_2 approach expressed their

interest by L_1 techniques, and conversely, some achievements in the classical setting provided a great inspiration for new ramifications. Problems arising in applications provided a better focus for theoreticians, and, on the other hand, applied people assured the former that theoretical insights are much needed and appreciated. Everyone learned valuable lessons about numerical methods, about relevant mathematical techniques, and about areas of application.

To better document the workshop in a publicly accessible place, and also for the future reference, we constructed the follow-up webpages at

`http://www.stat.ualberta.ca/ mizera/confer.html#banff`

containing documentation about program, abstracts, social events, and a fairly complete collection of transparencies for talks.

A full evaluation of the success of the workshop will certainly be only possible after some time. But even given our current perspective, it may be of some interest to note that under the auspices of the special semester devoted to inverse problems, at the Institute for Pure and Applied Mathematics of the University of California, Los Angeles, a special, previously unplanned, two-day workshop on statistical aspects of regularization was organized in November 2003. We would like to believe that the inspiration came partly from our workshop.

We are indebted to all participants; from those who helped us beyond the usual standard and thus deserve special thanks, we would like to especially thank Rudy Beran, Tony Chan, Andrew Conn, Steve Portnoy, Jim Ramsay, and Grace Wahba.

We can hardly express our thanks to the BIRS for the opportunity to organize the workshop. The scientific director, Robert Moody, was very supportive from the very beginning; but also substantially helped us with technicalities on the site, which included his introduction to the BIRS facilities on the eve of the workshop. Amanda Kanuka flawlessly handled the correspondence, and Brent Kearney assured that the excellent technical facilities worked smoothly. Last but not least, it was a great pleasure to work with the BIRS manager Andrea Lundquist.

Appendix: List of participants

Beran, Rudolf (University of California, Davis)
Candes, Emmanuel (California Institute of Technology)
Chambolle, Antonin (Université de Paris-Dauphine)
Chan, Tony (University of California Los Angeles)
Conn, Andrew R. (IBM Thomas J. Watson Research Center)
Cox, Dennis (Rice University)
Davies, P. Laurie (University of Essen)
Esedoglu, Selim (University of California, Los Angeles)
Golub, Gene (Stanford University)
Gu, Chong (Purdue University)
He, Xuming (University of Illinois at Urbana-Champaign)
Horowitz, Joel L. (Northwestern University)
Koenker, Roger (University of Illinois at Urbana-Champaign)
Kooperberg, Charles (Fred Hutchinson Cancer Research Center)
Kou, Samuel (Harvard University)
Kovac, Arne (University of Essen)
Mammen, Enno (Universität Heidelberg)
Meise, Monika (University of Essen)
Mizera, Ivan (University of Alberta)
Nowak, Robert (University of Wisconsin-Madison)
Oh, Hee-Seok (University of Alberta)
Portnoy, Steve (University of Illinois at Urbana-Champaign)
Ramsay, Jim (McGill University)
Ramsay, Tim (University of Ottawa)
Ruymgaart, Frits (Texas Tech University)
Sardy, Sylvain (Swiss Federal Institute of Technology)
Scherzer, Otmar (University of Innsbruck)
Shen, Jianhong (University of Minnesota)
Smale, Stephen (Toyota Technological Institute at Chicago)
Speckman, Paul (University of Missouri)
Stuetzle, Werner (University of Washington)
Vogel, Curtis (Montana State University)
Wahba, Grace (University of Wisconsin-Madison)
van de Geer, Sara (University of Leiden)