The alter egos of the regularized maximum likelihood density estimators: deregularized maximum-entropy, Shannon, Rényi, Simpson, Gini, and stretched strings

Roger Koenker, Ivan Mizera

Abstract: Various properties of maximum likelihood density estimators penalizing the total variation of some derivative of the logarithm of the estimated density are discussed, in particular the properties of their dual formulations and connections to stretched (taut) string methodology.

MSC 2000: 62G07, 62B10, 90C25, 94A17
Key words: Density estimation, Regularization, Penalized likelihood, Taut string

Grau, teurer Freund, ist alle Theorie
und grün des Lebens goldner Baum. [7]

1 Density estimation via regularized MLE

The objective of this note is to discuss certain aspects of the density estimation proposal put forward by the present authors in [16]. Let $x_1, x_2, \ldots, x_n$ be an observed sample. Regarding it as arising from a continuous distribution on $\Omega \subseteq \mathbb{R}^d$, we would like to estimate the underlying probability density. The regularized maximum likelihood approach provides an estimate f minimizing the penalized negative loglikelihood

$$-\frac{1}{n}\sum_{i=1}^{n} \log f(x_i) + \lambda J(f) \tag{1}$$

over the class of probability density functions, nonnegative functions that integrate to 1 over the whole Ω.

It is often practical to formulate a problem in terms of g such that $f = g^{[\kappa]}$; an appealing choice is the family of power transformations $f = g^{[\kappa]} = g^\kappa$, together with the limiting case $f = g^{[\infty]} = \mathrm{e}^g$. As it happens, from the potential continuum of possibilities only those corresponding to $\kappa = 1, 2$ and ∞ seem to be of practical interest. The choice $\kappa = 2$ circumvents enforcing the nonnegativity of f and as such was one of the principal motivations of the early proposal of Good [8]; see also Good and Gaskins [9], Eggermont and LaRiccia [6].

The choice $\kappa = \infty$, however, not only automatically enforces nonnegativity, but also simplifies the imposition of the integral constraint $\int f dx = 1$ by incorporating it directly into the objective function. This and several other intriguing properties

Research partially supported by NSF Grant SES-05-44673, and by the Natural Sciences and Engineering Research Council of Canada.

won it our eventual favor in [16]. Since $J(-g) = J(g)$ for any penalty considered below, we chose $-g$ rather g, mainly for aesthetic reasons; the resulting formulation in terms of $g = -\log f$ amounts to the unconstrained minimization of

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i) + \int \mathrm{e}^{-g(x)}dx + \lambda J(g). \tag{2}$$

The regularizing effect of the penalty prevents the ordinary maximum likelihood from degenerating into a linear combination of Dirac functions. Usual penalties employed in (2) have form $J(g) = \|Dg\|_p^p$, where D is a linear differential operator, and $\|\cdot\|_p$ is an integral norm. In dimension $d = 1$, this general scheme specializes to

$$J(g) = \|D^\nu g\|_p^p = \int |g^{(\nu)}(x)|^p dx, \tag{3}$$

where $g^{(\nu)}$ stands now for the ν-th derivative of g. Again, all investigated instances involve either $p = 1$ or 2, and $\nu = 1, 2$, and 3. For $p = 2$, the case with $\nu = 2$ was studied by Gu [10]; that with $\nu = 3$ by Silverman [20], see also Ramsay and Silverman [19]. The choice $p = 1$ corresponds to penalizing the total variation of the $(\nu-1)$-th derivative of g, and has received considerable recent attention due to the capability of total-variation penalties to capture qualitative features. Total variation is inherent to the "stretched string" methodology of Hartigan and Hartigan [12] and Hartigan [11], recently revived by Davies [2] and substantially enriched by Davies and Kovac [3, 4]; an inspiration for this note was the close resemblance of the outcomes of this methodology and those of the scheme (2)–(3) for $p = 1$ and $\nu = 1$. This is revealed in Section 4. Sardy and Tseng [17] recently proposed to penalize the total variation of the density itself—this proposal corresponds in our nomenclature to $\kappa = 1$, $\nu = 1$, and $p = 1$.

Our ultimate objective is the estimation of multidimensional densities densities, that is, $d > 2$. Some discussion, in the context of our proposal with $p = 1$ and $\nu = 2$ (and $\kappa = \infty$) can be found in [16]; the somewhat related case with $p = 2$ and $\nu = 2$ was studied by Gu [10]. However, space and simplicity considerations necessitate to omit this theme and also the discussion of practically very important, but complicated issues concerning automatic selection of the regularization parameter λ. The reader wishing to learn more is thus referred to Koenker and Mizera [16], as well as to all our future publications.

2 The discretization and its dual

As a computational strategy for (2)–(3), we adopted in [16] quite a straightforward finite-element approach, for $d = 1$ based on the representation of g as a piecewise linear function. Even if the final solution is not piecewise linear—in which case the pieces of linearity could be determined by the data points x_i themselves—sufficiently regular g can always be approximated in this manner on fine meshes. Another implementation problem is replacing a potentially infinite domain $\Omega \subseteq \mathbb{R}$

by a finite interval I; nevertheless, our experience suggests that maximized likelihood quite well guarantees that the estimated density vanishes outside its plausible support, provided that the convex hull of x_i is placed deep enough in the interior of I.

The discretized version of (2)–(3) seeks the estimate f in terms of the vector f, with components $\mathsf{f}_0, \mathsf{f}_1, \ldots, \mathsf{f}_m$; these relate to the values $\mathsf{g}_j = -\log \mathsf{f}_j$ of the piecewise linear approximation of g on the **mesh points** $\mathsf{v}_0 < \mathsf{v}_1 < \cdots < \mathsf{v}_m$. The working domain is thus $I = [\mathsf{v}_0, \mathsf{v}_m]$, and the integral of $f = \exp(g)$ is approximated by the sum of $\mathsf{c}_j \exp(\mathsf{g}_j)$ following the trapezoid formula—the components of the **coefficient vector** c are $\mathsf{c}_j = (w_j + w_{j-1})/2$, with $w_{-1} = w_m = 0$, and otherwise $w_j = \mathsf{v}_j - \mathsf{v}_{j-1}$. Other approximations are possible, and may very slightly alter the final result.

The negative loglikelihood at the point x_i is expressed through the result of evaluation functional l_i applied to the vector g; in the simplest instance, when the mesh points v_j are selected to contain x_i, the evaluation functional returns the value f_j/n, where $\mathsf{v}_j = x_i$. That is, the coefficients of l_i are all zero except for the j-th one which is $1/n$. For larger n, it may be more numerically stable to choose the mesh equidistant; in such a case, l_i may analogously relate to the value of f at the nearest mesh point, or may express the interpolation between the two nearest ones. All evaluation functionals l_i form together an **evaluation operator** L. Whichever construction from those outlined above is used, L always satisfies the following important properties: (i) $\mathsf{L1} = (1/n)\mathsf{1}_n$, where $\mathsf{1}_n$ is the vector of ones with length n; (ii) $\mathsf{Lv} = \mathsf{x}$, where x is the vector consisting of the sample points x_i. The consequence of (i) is that $\mathsf{1}_n^\mathsf{T}\mathsf{L1} = 1$; similarly, (ii) implies that $\mathsf{1}_n^\mathsf{T}\mathsf{Lv} = (1/n)\mathsf{1}_n^\mathsf{T}\mathsf{x}$, the sample mean of x.

The penalization is carried for the piecewise linear approximation of g in an *exact* manner; that is, given g, the integral of the absolute value of the first or second derivative can be expressed as the ℓ^1 norm of Pg, where P is a linear **penalization operator** whose coefficients depend only on the mesh v. In particular, the total variation of a piecewise linear g is the sum, for $j = 1, 2, \ldots, m$, of $|\mathsf{g}_j - \mathsf{g}_{j-1}|$; apparently, in this case P is the simple difference operator, hereafter denoted by ∇. Note that ∇ annihilates $\mathsf{1}$: $\nabla\mathsf{1} = \mathsf{0}$. As a rule, this is also true for other penalization operators. For instance, the operator P arising from penalization of the total variation of the derivative ($\nu = 2, p = 1$), whose formula can be found in [16], annihilates $\mathsf{1}$ as well; just note that the constant function is piecewise-linear and therefore the total variation of its derivative must be zero.

The exact approach to penalization for higher derivatives and/or their squares would require higher-order piecewise-polynomial finite elements; in dimension $d = 1$ it is numerically simpler to make the meshes fine and equidistant, and rather *approximate* the penalties in the spirit of finite-difference methods (note that the case $p = 2$ and $\nu = 1$ can be still carried out exactly).

The components listed above form the discretized version of (2)–(3), which

amounts to the convex optimization problem seeking f minimizing

$$\mathsf{1}^\mathsf{T}\mathsf{L}\mathsf{g} + \mathsf{c}^\mathsf{T}\mathsf{f} + \lambda\|\mathsf{P}\mathsf{g}\|_p^p, \qquad \mathsf{f} = \exp(\mathsf{g}) \quad \text{componentwise.} \tag{4}$$

The inputs are x_i (reflected through L) and λ; the formulation depends on the way how the evaluation operator L, the coefficient vector c and the penalization operator P are set, and also on the selected value of p.

The conjugate dual of (4), derived in [16], minimizes, for $p = 1$,

$$\sum_j \mathsf{c}_j\mathsf{f}_j \log \mathsf{f}_j, \qquad \mathsf{c}_j\mathsf{f}_j = \mathsf{h}_j \text{ for all } j, \qquad \mathsf{h} = \mathsf{L}^\mathsf{T}\mathsf{1}_n + \mathsf{P}^\mathsf{T}\mathsf{w}, \qquad \|\mathsf{w}\|_\infty \le \lambda. \tag{5}$$

Similar dual formulations can be formulated for other values of p. While already the formulation (4) is within the scope of the modern optimization software (for the values like $m = 1000$, say, which is fully satisfactory for the graphing, for instance), the dual formulation saves about 25% of CPU time. Beyond that, it has interesting theoretical interpretations and consequences.

3 Deregularized maximum entropy

The idealized functional form of the dual (5) (which will be in the full formal vigor pursued elsewhere) can be viewed as the maximization of the (differential) Shannon entropy $H(f) = -\int f \log f$. This maximization takes place over the sieve $\mathsf{h} = \mathsf{L}^\mathsf{T}\mathsf{1}_n + \mathsf{P}^\mathsf{T}\mathsf{w}$, $\|\mathsf{w}\|_q \le \lambda$. The coefficient vector c corresponds to the dominating measure; f can be thus interpreted then as a density of the probability measure corresponding to h. To see that the adjective "probability" is justified, we observe the following. If a is annihilated by P, then $\mathsf{a}^\mathsf{T}\mathsf{h} = \mathsf{a}^\mathsf{T}\mathsf{L}^\mathsf{T}\mathsf{1}_n$. We noted already that $\mathsf{1}$ is annihilated by typical penalization operators, and that for any evaluation operator, $\mathsf{1}^\mathsf{T}\mathsf{L}\mathsf{1}_n = 1$. Therefore, the elements of h sum to $\mathsf{1}^\mathsf{T}\mathsf{h} = \mathsf{1}^\mathsf{T}\mathsf{L}^\mathsf{T}\mathsf{1}_n = 1$.

For penalties involving derivatives higher than first, this reasoning can be extended to moments (interpreting the foregoing summation as the 0th moment). For instance, the total variation of the derivative of any linear function is 0; since any linear function is piecewise linear, the corresponding penalization operator P annihilates any a of the form $\mathsf{a}_j = \alpha + \beta\mathsf{v}_j$. In particular, $\mathsf{v}^\mathsf{T}\mathsf{h} = \mathsf{v}^\mathsf{T}\mathsf{L}^\mathsf{T}\mathsf{1}_n = (1/n)\mathsf{x}^\mathsf{T}\mathsf{1}_n$; thus, the mean of h is equal to the sample mean of x.

Similar considerations give us (at least approximate) equality of higher-order moments, if higher-order derivatives are penalized. As can be best seen from the idealized functional version, penalizing the derivative of order ν makes the moments up to order $\nu - 1$ of the estimated density equal to their sample counterparts (irrespective of p, that is regardless of whether the ℓ^1 or ℓ^2 norm is penalized).

If we set $\lambda = 0$, then the sieve contains only one element, which corresponds to the empirical probability measure supported by x. By increasing λ we gradually relax the restrictions on the sieve, by adding a sequence of "stabilizers" w_i, filtered through the adjoint P^T of the operator P.

Note, however, that the equality of moments remains preserved for every λ. Therefore, for $\lambda \to \infty$ the estimate approaches the result of entropy maximization

subject to moment constraints—which is the classical maximum entropy estimate, dating back at least to Jaynes [13]. The results of Shannon [18] about entropy maximization, available in the modern form from Kagan, Linnik, and Rao [15], tell us in the functional case that (a) for bounded domains, the maximum entropy distribution without any moment distribution is uniform over the domain; (b) for unbounded domains, the maximum entropy distribution with first two moments fixed is the normal distribution with the corresponding mean and standard deviation. The case (a) is prototypic to $\nu = 1$, penalizing the first derivative; the case (b) to $\nu = 3$, penalizing the third derivative. Since only the form of the operator P (or D) matters here, the same limiting distribution is obtained either for $p = 2$, the proposal of Silverman [20], or for $p = 1$, the proposal of this note.

Our numerical evidence supports this qualitative behavior. Even for the bounded domains of the discretized version, the constraint on the first two moments ($\nu = 3$) gives a quite stable truncated normal distribution close to the ideal one in the limit. The case with a constraint on the mean, but not on variance ($\nu = 2$), exhibits somewhat intermediate behavior. For moderate λ, the results do not differ much from those with $\nu = 3$; however, for $\lambda \to \infty$ those for $\nu = 3$ converge to the aforementioned truncated normal, while those for $\nu = 2$ eventually sink into the uniform, exhibiting sometimes an intermediate double-exponential behavior—to a greater extent for ℓ^1, to a lesser one for ℓ^2 penalties, corresponding the proposals of Koenker and Mizera [16] and Gu [10], respectively.

Note the opposite tendencies in λ: while the primal wants to increase λ to regularize the MLE off its overfitting behavior, the dual rather prefers to decrease λ to get the maximum entropy estimator closer to the reality in the data. For this reason, we tend to speak rather about *deregularized* than regularized maximum entropy.

Finally, further manipulation of quantities involved in the dual prescription shows the well-known equivalence of maximum entropy principle with that of minimum Kullback-Leibler divergence, in our discretized version deregularized to minimize, over the very same sieve, the discrepancy of the estimated density to the uniform on I. For more background, see Vajda [21], and the references there.

4 Tautology: some string theory and practice

The simplicity of the case $\nu = 1$ and $p = 1$, corresponding to penalizing the total variation of g itself, allows for some geometric understanding of the dual—in connection to the "stretched" or "taut" string methods.

Consider first the ideal functional version again. Integration by parts shows that the adjoint of the first derivative operator D^1 is $-D^1$; in the light of the symmetry of the penalization prescription and that of the sieve, we can safely drop the minus sign. In dimension $d = 1$, it is possible express the density f as the derivative of the cumulative distribution function, $f = F'$. The dominating measure is the Lebesgue one, hence c_j is equal to a constant; identifying f with h and applying the antiderivative operator to the both sides of the sieve identity, we obtain that $F =$

$F_n + w$ with $\|w\| \le \lambda$; here F_n is the empirical distribution function supported by x. In other words, the estimated F lies within the Kolmogorov distance λ of F_n, and minimizes the objective function $\int F'(x) \log F'(x) dx$, which makes F linear between the points where it touches the boundary of the Kolmogorov "tube"—this follows from an exercise in classical calculus of variations: minimizing $\int_\alpha^\beta f(x) \log f(x) dx$ under boundary conditions fixing $f(\alpha)$, $f(\beta)$ leads to the solution linear on $[\alpha, \beta]$. In other words, our F is the stretched string in the neighborhood of F_n given by λ!

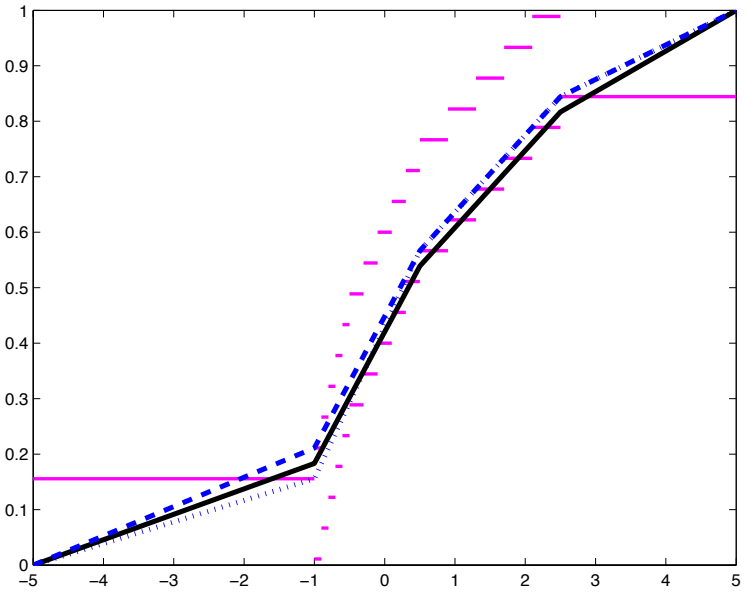

Figure 1: Stretched strings in three differently understood tubes.

Since all this may appear pretty unconvincing (and on the other hand probably not worth a formal proof), let us try to implement (5) numerically and compare it with the taut string estimate, as returned by the function `pmden()` of the R package `ftnonpar` of Davies and Kovac [5]. However, here we stumble on aspects indicated by the motto in the preamble. While the theoretical descriptions in [11] and [2] seem clear, [4] in passing indicates that the outcome of `pmden()` is a result of some closer unspecified (and undocumented in `ftnonpar`) aftersmoothing. Fortunately, the open source character of the R code of `pmden()` enables reverse engineering—after which we are able to conjecture that the key to the "real" taut lies in its variable `fts`.

The next problem lies in the interpretation of "tube". We omit the details and rather refer the reader to Figure 1, which shows three possible interpretations that

come immediately to mind. However, it turns out that `pmden()` uses a fourth one—not that unnatural in the given context, but also not that immediately coming to mind.

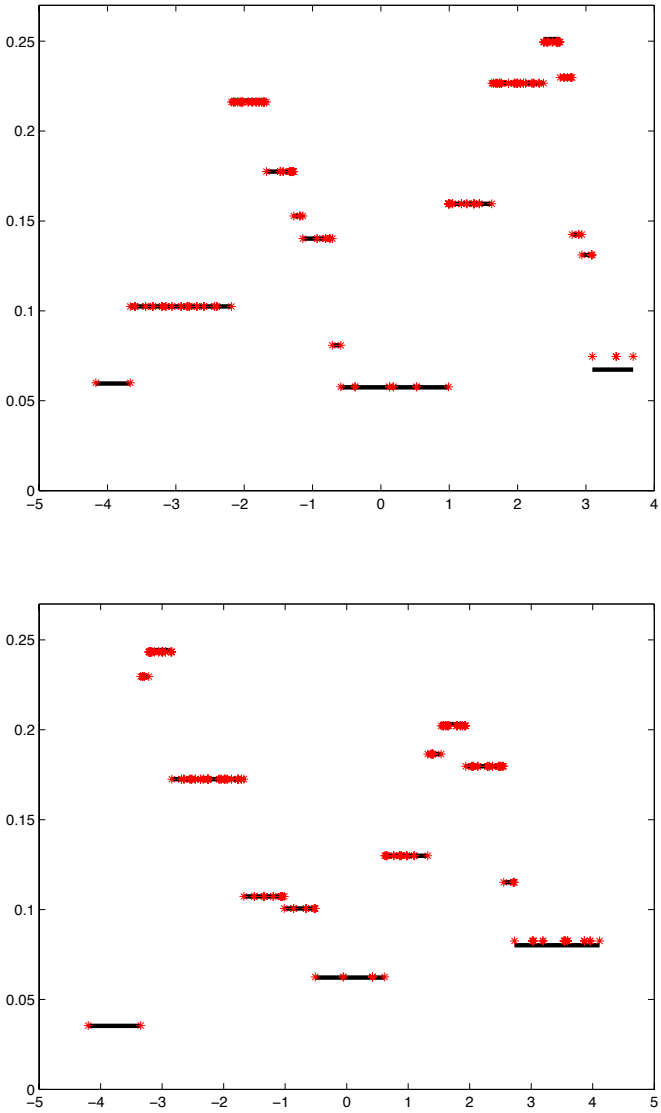

Figure 2: Short simulation experiment (2 runs): `string` plotted over `taut`.

Having gathered all this intelligence, it is not hard to reproduce it as an output of our MATLAB function `taut` and try it on several examples against the original variable `fts` from `pmden()`. While we find the agreement in this case fully satisfactory, the same cannot be said yet with respect to the output of our implementation of (5), with $\mathsf{P} = \nabla$. The problem is not only in different understanding of the concept of the tube, but also in the fact that (5) computes density *at* the mesh points, while the taut string algorithm *between* them. Nevertheless, a slight modification of our coefficient vector (that is, the integral approximation formula) to $\mathsf{c}_i = ((n-1)/n)(\mathsf{v}_{i+1} - \mathsf{v}_i)$ for $i = 0, 1, \ldots, n-1$, and $\mathsf{c}_n = (1/n)(\mathsf{v}_n - \mathsf{v}_0)$ produces, on the mesh generated by the sorted x_i, fits that in repeated runs of our new MATLAB function `string` (see Figure 2 for a sample) give 99% satisfactory agreement; we attribute the missing 1% to numerical discrepancies, different algorithms and similar phenomena between Earth and Heaven. At this point, we—if not the reader—are sufficiently convinced that this very special case of (5) (recall: $\nu = 1$, $p = 1$, $\mathsf{v} = \mathsf{x}$ and tweaked c) can be viewed as essentially identical with the stretched string method.

5 Beyond Shannon: brave new worlds

We have to admit that our intention to make all our research reproducible is hitting a serious obstacle at this point: our computational procedures require an access to proprietary software, in particular to a reasonable convex optimization solver: not only MATLAB and its optimization toolbox, but unfortunately beyond. While we gratefully acknowledge the use of Danish MOSEK [1] we may also contemplate potential modifications of our estimating prescriptions. A possible direction is to consider a generalized form of the entropy functional based on Rényi's entropy; this is defined for $\alpha \neq 1$ as

$$H^\alpha(f) = \frac{1}{1-\alpha} \log \int f^\alpha dx,$$

the limit for $\alpha \to 1$ being equal to the Shannon entropy $H(f)$, which can be thus viewed as $H^1(f)$. Using the discrete version of H^α in the objective function of (5), we arrive to a continuum of alternatives, from which again only few are of interest (if any): apart from H^1, it is most prominently H^2, whose versions are known in ecology as Simpson's diversity index and in economics as Gini's index. From the computational perspective, note that replacing the Shannon entropy by that of Rényi-Simpson-Gini moves (5) from the "nonlinear convex" to "quadratic" programming denomination —hence possibly easing the implementation. *

Somewhat unexpectedly, we find that replacing H^1 by H^2 does not change the outcome for the taut string case ($\nu = 1$, $p = 1$). Indeed: the aforementioned exercise to find a minimum of $\int_\alpha^\beta f(x) \log f(x) dx$ under fixed $f(\alpha)$, $f(\beta)$ yields same

*Not in the MATLAB optimization toolbox, though. While MOSEK's quadratic programming function successfully converges to the solutions shown at Figures 4 and 5, MATLAB's crashes at the same input data reporting mysterious error messages.

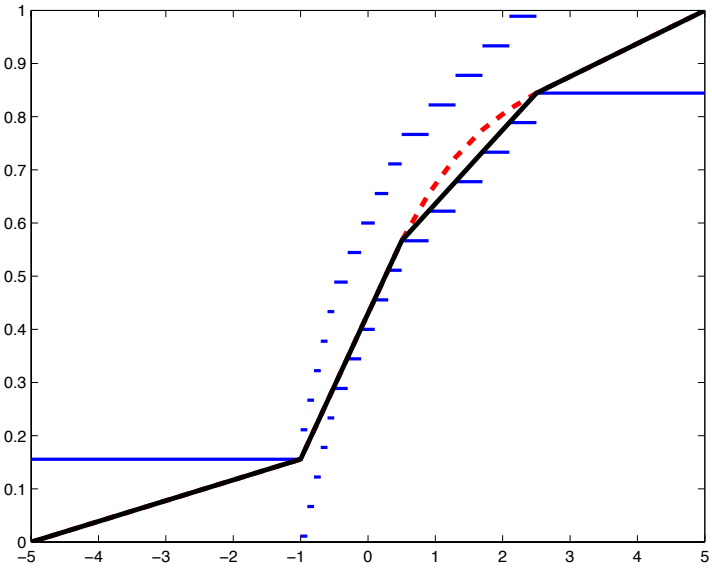

Figure 3: Total variation does not stretch the string unambiguously.

linear solution when the minimized functional is replaced by $\int_\alpha^\beta (f(x))^2 dx$. In this respect, an interesting question is whether the minimized functional could not be the L^1 functional $\int_\alpha^\beta |f(x)| dx$, as suggested on page 19 of Davies and Kovac [3]. Note that this would mean further simplification, downshifting the denomination from "quadratic" to "linear" programming. Unfortunately, the answer is negative, for the reason that can be seen in Figure 3. For the L^1 functional, the taut string is only *a* solution, not *the* solution; albeit this might not be an issue in many cases, there is always a danger that an L^1-based density estimation algorithm would return different results, depending on the particular linear programming implementation.

Despite these findings, it is important to remind ourselves and the reader that the world does not end at the first derivative ($\nu = 1$). After all, one can ask whether the Shannon and Rényi-Simpson-Gini density estimators, shown in Figure 4 (for $\nu = 2$) and Figure 5 (for $\nu = 3$), and computed from the data used in the left panel of Figure 2 do not each resemble the 50:50 mixture of N(−2,1) and N(2,1) densities (from which the data were simulated) more than anything from Figure 2. Note that for $\nu > 1$ different entropies really give different results—albeit perhaps not that much.

Before proceeding to the final conclusion, let us mention that our findings about the nature of stretched strings may have some relevance also for the nonparametric

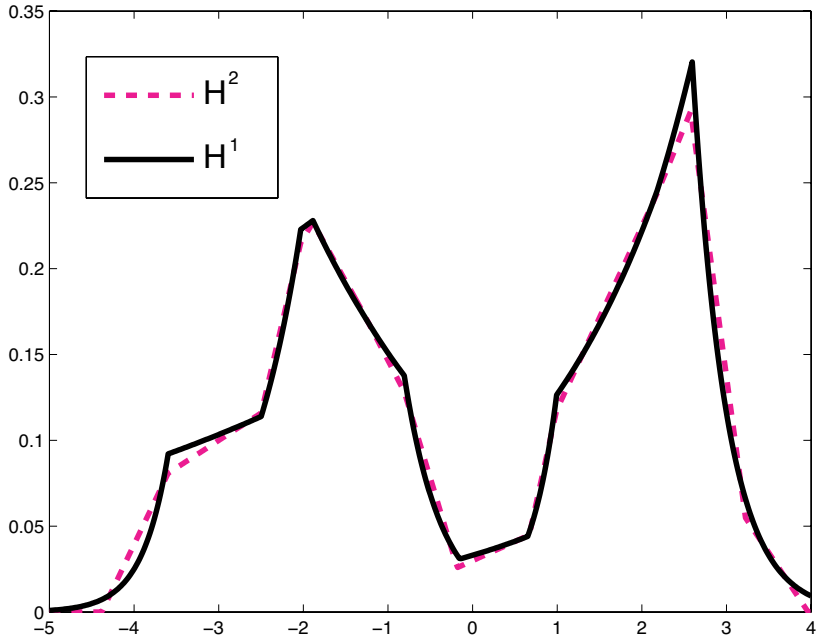

Figure 4: The Shannon (H^1, solid) and the Rényi-Simpson-Gini (H^2, broken) deregularized maximum-entropy estimates penalizing the total variation of the *first* derivative of the logarithm of the estimated density.

regression methods based on this principle; knowing the equivalence of strengths and weakness of various possible dual prescriptions entices to trace them back into their primal formulations.

6 Conclusion: density estimation in the new millenium

Density estimation methods based on penalized maximum likelihood offer a flexible and data-analytically very apealing methodological tool—as can be realized from various viewpoints and is also documented by the recent wave of interest in these methods. While the fortuitous aspects of the L^2 approach (a possibility to express exact solutions via Hilbert space theory; linear estimating algorithms) are in the density estimation context for the most part lost due to the presence of nonnegativity and integrability constraints, the L^1 formulation offers not only a computationally feasible alternative, but through its conjugate dual also a possibility to apply the sophisticated λ selection adaptive strategies developed by Davies and

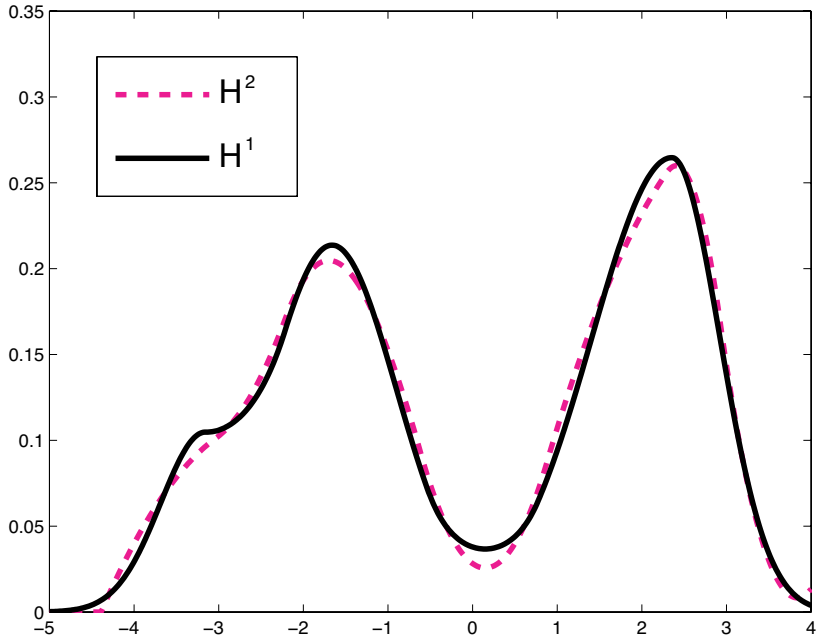

Figure 5: The Shannon (H^1, solid) and the Rényi-Simpson-Gini (H^2, broken) deregularized maximum-entropy estimates penalizing the total variation of the *second* derivative of the logarithm of the estimated density.

Kovac [3], [4], as exemplified by the taut string connection. The dual formulation brings not only computational savings and intriguing theoretical interpretations, but the use of alternative entropies broadens the scope of the methods and may lead to further conceptual and computational improvements.

From various possibilities in this vein, the L^1 analog of the proposal of Silverman [20], minimizing

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i) + \int \mathrm{e}^{-g(x)}dx + \bigvee g',$$

the last term standing for the total variation of the *second derivative* of g, may have additional aesthetic qualities, in particular arising from the fact that its regularization limit is normal. Moreover, this extends to the multivariate case, and recent results of Johnson and Vignat [14], and others establish similar properties also for the Rényi-Simpson-Gini alter ego, with the role of normal played by t distributions.

From the pragmatic standpoint, however, any definitive recommendations will be possible only after a more rigorous scrutiny of possible alternatives, complemented by appropriate automatic λ selection rules. We hope that this note gives some impetus for such an undertaking.

References

[1] E. D. Andersen and K. D. Andersen, The MOSEK interior point optimizer for linear programming: *Applied Optimization*, 33, 1999.

[2] P. L. Davies. Data features. *Statist. Neerlandica*, 49:185–245, 1995.

[3] P. L. Davies, A. Kovac. Local Extremes, Runs, Strings and Multiresolution (with discussion). *Ann. Statist.*, 29:1–65, 2001.

[4] P. L. Davies, A. Kovac. Densities, spectral densities and modality. *Ann. Statist.*, 32:1093–1136, 2004.

[5] P. L. Davies, A. Kovac. `ftnonpar`, R package, version 0.1-4. `http://www.r-project.org`, 2005.

[6] P. P. B. Eggermont, V. N. LaRiccia. *Maximum Penalized Likelihood Estimation Volume I: Density Estimation*. Springer-Verlag, New York, 2001.

[7] J. W. von Goethe. *Faust*. Akademie-Verlag, Berlin, 1954. (in German)

[8] I. J. Good. A nonparametric roughness penalty for probability densities. *Nature*, 229:29–30, 1971.

[9] I. J. Good, and R. A. Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika*, 58:255–277, 1971.

[10] C. Gu. *Smoothing Spline ANOVA Models*. Springer-Verlag, New York, 2002.

[11] J. A. Hartigan. Testing for antimodes. In *Data Analysis: Scientific Modeling and Practical Applications* (W. Gaul, O.Opitz, and M. Schader, eds.) 169–181. Springer-Verlag, New York, 2000.

[12] J. A. Hartigan, P. M. Hartigan. The dip test of unimodality. *Ann. Statist.* 13:70–84, 1985.

[13] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, 1957.

[14] O. Johnson, C. Vignat. Some results concerning maximum Rényi entropy distributions. arXiv:math.PR/0507400 v1, 2005.

[15] A. M. Kagan, Yu. V. Linnik, and C. R. Rao. *Characterization Problems in Mathematical Statistics*. Nauka, Moskva, 1972. (in Russian)

[16] R. Koenker, I. Mizera. Density estimation by total variation regularization. University of Alberta Statistics Centre Technical Reports 06.02. `http://www.stat.ualbertaqq.ca/stats_centre/tech.htm`, 2006.

[17] S. Sardy, P. Tseng. Estimation of nonsmooth densities by total variation penalized likelihood driven by the sparsity L_1 information criterion. `http://statwww.epfl.ch/people/sardy/PDF/preprintSLIC06.pdf`, 2005.

[18] C. E. Shannon. The Mathematical Theory of Communication. In C. E. Shannon, W. Weaver, *The Mathematical Theory of Communication*. University of Illinois Press, Urbana-Chicago-London, 1949.

[19] J. Ramsay, B. Silverman. *Functional Data Analysis, Second Edition*. Springer-Verlag, New York, 2005.

[20] B. Silverman. On the estimation of a probability density function by the maximum likelihood method. *Ann. Statist.*, 10:795-810, 1982.

[21] I. Vajda. *Theory of Statistical Inference and Information*. Bratislava, Alfa, 1982. (in Slovak)

Roger Koenker: University of Illinois at Urbana-Champaign, Departments of Economics and Statistics, Champaign, Illinois, 61620, USA, rkoenker@uiuc.edu

Ivan Mizera: University of Alberta, Department of Mathematical and Statistical Sciences, CAB 632, Edmonton, T6J0Z2, Canada, mizera@stat.ualberta.ca