Mode Estimation in High-dimensional Spaces with Flat-top Kernels: Application to Image Denoising

Arnaud de Decker[1], John Aldo Lee[2], Damien François[1], and Michel Verleysen[1] *

1- Université catholique de Louvain, Machine Learning Group,
Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium

2- Université catholique de Louvain, Imagerie Moléculaire et Radiothérapie Expérimentale
Avenue Hippocrate 55, B-1200 Bruxelles, Belgium

Abstract. Data denoising can be achieved by approximating the data distribution and replacing each data item with an estimate of its closest mode. This idea has already been successfully applied to image denoising. The data then consists of pixel intensities or image patches, that is, vectorized groups of pixel intensities. The latter case raises the issue of mode estimation in a high-dimensional space, since patches can contain about 10 to more than 100 pixels. This paper shows that the widely used Gaussian kernel is outperformed by flat-top kernels that are specifically tailored in order to fight the curse of dimensionality.

1 Mode Estimation

Starting from a data sample, mode estimation can be achieved by first approximating the underlying data distribution. Next, a hill-climbing procedure can be run from any point in order to obtain an estimate of the closest mode. If $\mathbf{X} = [\mathbf{x}_i]_{1\leq i\leq N}$ denotes the data sample, then the widely known Parzen's window kernel probability density estimator (KPDE) [1] can be written as $\hat{p}(\mathbf{x}_i) = C\sum_{j=1}^{N} \Psi_\sigma(\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/2)$, where C is a normalization factor that ensures that $\int \hat{p}(\mathbf{x})d\mathbf{x} = 1$ and kernel Ψ_σ is a positive and monotically decreasing function. Parameter σ denotes the (band)width of the kernel. Equating the derivative of this KPDE with 0 provides a fixed-point update that is written as

$$\hat{\mathbf{x}}^{(t+1)} = \frac{\sum_{i=1}^{N} \Psi'_\sigma(\|\hat{\mathbf{x}}^{(t)} - \mathbf{x}_i\|_2^2/2)\mathbf{x}_i}{\sum_{i=1}^{N} \Psi'_\sigma(\|\hat{\mathbf{x}}^{(t)} - \mathbf{x}_i\|_2^2/2)} \quad , \tag{1}$$

where t is the iteration index. If $\hat{\mathbf{x}}^{(0)} = \mathbf{x}_i$, then iterating the update amounts to climbing towards the closest hill top on the pdf estimate. Such an iterative mode estimator can also emerge from the field of robust statistics [2]. The idea is to refine the classical least-squares estimator of the mean in order to make it more robust against outliers. For this purpose, the quadratic terms are replaced with upper-bounded functions such as Leclerc's $\sigma^2(1-\exp(-x^2/\sigma^2))$. The fixed-point minimization of this generalized non-convex estimator leads to a similar update as the hill-climbing procedure on a KPDE.

The literature contains many publications that give recommendations as to the choice of the optimal kernel and bandwidth [3; 4; 5]. It is noteworthy that the recommendations are different for KPDE than for robust statistics. In the former case, they aim at producing the best estimator, achieving for instance a minimal KL divergence with

*A. de Decker is funded by a Belgian F.R.I.A. grant. J.A. Lee is a Research Associate with the Belgian National Fund of Scientific Research (FNRS).

respect to the true pdf. In the latter case, however, smoothness is important as well, in order to prevent the fixed-point procedure to get stuck in poor local minima. Therefore, robust statistics (or equivalently hill-climbing on a KPDE) requires a tradeoff between accuracy and smoothness to be found. Most of the recommendations, however, deal with mono- or two-dimensional case. In contrast, this paper investigates the choice of the kernel for high-dimensional data. We suggest using flat-top kernels as a way to fight the curse of dimensionality [6; 7] and in particular the phenomenon of norm concentration [6; 7]. An application to image denoising, with so-called patch-based filters [8; 9; 10], shows that flat-top kernels lead to better performances.

The rest of this paper is organized as follows. Section 2 deals with the counter-intuitive properties of norms and distances in high-dimensional spaces. It also defines similarity kernels that take these properties into account. Section 3 introduces image filtering with patch-based approaches. Section 4 experimentally compares the denoising performance of patch-based filtering with either a Gaussian kernel or the proposed similarity functions. Finally, Section 5 draws the conclusions.

2 Norms, distances, and similarities in high-dimensional spaces

The curse of dimensionality [6; 7] refers to the counter-intuitive properties of high-dimensional spaces. Among many other weird manifestations, the phenomenon of norm concentration is particularly annoying. It states that usual norms and distances tend to be poorly discriminative in high-dimensional spaces. More precisely, when the dimensionality grows, the distribution of distances has an increasing mean. As an example, in the case of a D-dimensional zero-mean unit-variance Gaussian distribution, Euclidean norms are χ_D-distributed. This is illustrated in Fig. 1. It is easy to see that pairwise distances in an isotropic Gaussian D-dimensional distribution are also χ_D-distributed, up to a scaling factor.

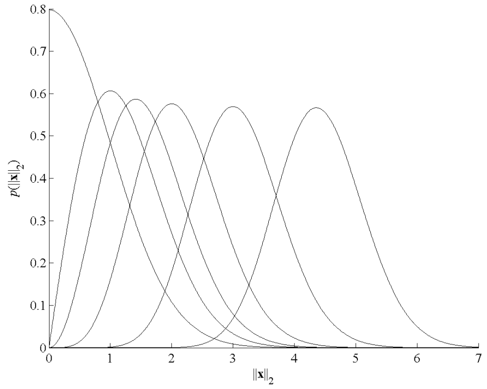

Fig. 1: Probability densities of $\|\mathbf{x}\|_2$ if $\mathbf{x}$ is multivariate Gaussian with zero-mean, unit-variance and dimensionalities equal to $\{1,2,3,5,10,20\}$ from left to right.

Similarities (also known as affinities or proximities) are typically inversely proportional to distances. When they are not ad-hoc measurements, similarities often result from the application of a decaying kernel to pairwise distances. The most widely used kernel is without any doubt the Gaussian function $\exp(-x^2/2)$; it is smooth, easy to compute, and owns a lot of other useful properties. The application of the Gaussian function casts Euclidean distances within the interval $[0,1]$. In order to be discriminant and thus useful, a similarity is expected to occupy this interval as uniformly as possible. For the sake of simplicity, let us consider a multimodal distribution, where each mode is Gaussian. If two vectors stem from a different mode, then we consider them to be dissimilar. Two vectors are similar if they are drawn from the same mode, that is, they are both noisy measurements of the same quantity. This allows us to focus on a single mode. In order to be discriminant and thus useful, we expect the similarity measure to be close to 1 for distances between 0 and quantile 0.05 of the distance distribution. Similarly, the similarity should be close to 0 beyond quantile 0.95. Hence, most of the similarity decay must occur within the interval given by quantiles 5% and 95%. In a high-dimensional space, a Gaussian kernel cannot satisfy these requirements, whatever the value of its width is.

Keeping the same assumptions, i.e. Gaussian modes of width σ, we see that D-dimensional vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ are similar if $\mathbf{x}_i - \mathbf{x}_j \sim N(0, \sqrt{2}\sigma\mathbf{I})$ and therefore $\|\mathbf{x}_i - \mathbf{x}_j\|_2 \sim \chi_D$. Within this framework, we define the similarity as the probability of observing a larger distance than the one that is measured, that is, $\mathrm{sim}(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{P}[\sqrt{2}\sigma c \leq \|\mathbf{x}_i - \mathbf{x}_j\|_2]$, where $c \sim \chi_D$. The similarity is thus given by the complementary cumulative distribution function (CCDF) of a scaled χ_D variable.

$$\mathrm{sim}(\mathbf{x}_i, \mathbf{x}_j) = \int_{\|\mathbf{x}_i - \mathbf{x}_j\|_2}^{\infty} \frac{\sqrt{2}}{\Gamma(D/2)} \left(\frac{c}{2\sigma}\right)^{D-1} \exp\left(-\frac{c^2}{4\sigma^2}\right) dc = Q\left(\frac{c^2}{4\sigma^2}, \frac{D}{2}\right) \quad , \qquad (2)$$

where Q is the regularized upper incomplete Gamma function [11]. The Gaussian kernel corresponds to the case $D = 2$.

The above CCDF is rather expensive to compute. It can be approximated reasonably well with the CCDF of an 'all-purpose' distribution over $\mathbb{R}^+$, such as the Burr type XII distribution [12]. Its scaled CCDF is given by $F(b, \lambda, \tau, \theta) = (1 - (b/\lambda)^\tau)^{-\theta}$. Choosing $\tau = D$, $\theta = -1$, and $\lambda = \sqrt{2}\sigma$ yields a good approximation of the χ_D distribution. The Burr CCDF reproduces the shape of the χ_D CCDF, with first a flat top, followed by a steep descent, and eventually a thin tail.

It is noteworthy that with the kernel (2), the similarity equals $1/2$ when the distance equals the median value of the distribution. Depending on mode overlap, this choice can be suboptimal. Bandwidth σ can then be adjusted in order to shift the kernel decay to the left or to the right, in the same way as the bandwidth is adjusted and/or optimized in Gaussian kernels.

3 Patch-based filtering

The intuitive idea behind patch-based filtering [8] is that a filtered image can be obtained by averaging pixels sharing similar neighborhoods. For this purpose, patch-based filters usually compute similarities with a kernel whose argument is a Euclidean distance between two vectorized image patches. This methodology turns out to be strongly related to robust statistics and hill-climbing on a KPDE, with the specificity of working in very high-dimensional spaces. To see this, we define an image I as a set of pixel locations. Each pixel is referred to with its coordinate vector $\mathbf{i}$. The distance between the $\mathbf{i}$th and $\mathbf{j}$th pixel is then defined as $\|\mathbf{i}-\mathbf{j}\|_0$. A (square) neighborhood around the $\mathbf{i}$th pixel is defined as $P_\mathbf{i} = \{\mathbf{j} \text{ s.t. } \|\mathbf{i}-\mathbf{j}\|_\infty \leq r\}$, where r is some radius. The intensity of the $\mathbf{i}$th pixel is denoted by $x_\mathbf{i}$. A patch can then be denoted by $\mathbf{x_i} = [x_\mathbf{j}]_{\mathbf{j}\in P_\mathbf{i}}$.

A local KPDE in the patch space can be written as $\hat{p}(\mathbf{x_i}) = \sum_{\mathbf{j}\in I} \Psi_\sigma(\|\mathbf{x_i}-\mathbf{x_j}\|_2^2/2)$. Equating the partial derivative with respect to $x_\mathbf{k}$ leads to the fixed-point update

$$x_\mathbf{k}^{(t+1)} = \frac{\sum_{\mathbf{i}\in P_\mathbf{k}} \sum_{\mathbf{j}\in I} w_\mathbf{ij} \Psi'_\sigma(\|\mathbf{x}_\mathbf{i}^{(t)} - \mathbf{x_j}\|_2^2/2) x_{\mathbf{j+i-k}}^{(t)}}{\sum_{\mathbf{i}\in P_\mathbf{k}} \sum_{\mathbf{j}\in I} w_\mathbf{ij} \Psi'_\sigma(\|\mathbf{x}_\mathbf{i}^{(t)} - \mathbf{x_j}\|_2^2/2)} \ , \tag{3}$$

where t is the iteration index and $x_\mathbf{k}^{(0)}$ is initialized to $x_\mathbf{k}$. In order to keep an acceptable computational load, weight $w_\mathbf{ij}$ can be chosen as a decaying function of $\|\mathbf{i}-\mathbf{j}\|_2$ and negligible terms can be omitted in the double sum.

The update rule (3) turns out to be an iterative generalization of an image denoising technique known as the nonlocal means [8]. Several versions of this filter exist, such as UINTA [9], SAFIR [13], and many others [10; 14; 15; 16; 17; 18]. Most of the variants, however, keep using the Gaussian kernel in spite of the high dimensionality of the patches. Nevertheless, a few attempts to fight the curse of dimensionality can be found in the literature. For example, in [19], a principal component analysis is achieved in the patch space, in order to reduce the dimensionality and hence the distance concentration. In [10], the authors suggest shifting the distribution of patch distances towards zero, just by subtracting a dimensionality-dependent constant from each distance. Instead, we propose to replace the classical Gaussian kernel $\Psi_\sigma(u^2/2) = \exp(-u^2/2/\sigma^2)$ with

$$\Psi_\sigma(u^2/2) = \int_0^u Q\left(\frac{v^2}{4\sigma^2}, \frac{D}{2}\right) v\,dv \tag{4}$$

or its Burr approximation given by $\Psi_\sigma(u^2/2) = \int_0^u (1-(v/\sigma)^D)^{-1} v\,dv$.

4 Experiments

The experiments feature three images with 512 × 512 pixels and 256 gray levels. The first one is composed of vertical one-pixel-wide stripes. The second image is a stair, with a constant low intensity on the left half and a higher one on the right half. For these two artificial images, the low and high pixel intensities are equal to 50 and 100, respectively. The third image is Barbara's widely used picture. White Gaussian noise is added to all images with three standard deviation levels: $\nu = \{35, 50, 65\}$ for the artificial images and $\nu = \{10, 15, 20\}$ for Barbara's less contrasted portrait.

Image are filtered by iterating (3) three times. The metaparameters of the filter are the kernel function (Gaussian, χ_D CCDF, or Burr CCDF) and its bandwidth σ. For each image, kernel and noise level, denoising is carried out for $\sigma = \{1, 5, 10, 15 \ldots, 2\nu)$.

The denoising quality is assessed with the root mean square error that is defined as RMSE$(\mathbf{z},\mathbf{x}^{(t)}) = (\sum_{\mathbf{i}\in I}(z_{\mathbf{i}} - x_{\mathbf{i}}^{(t)})^2/|I|)^{1/2}$, where $\mathbf{z}$ and $\mathbf{x}^{(t)}$ are the noisefree and filtered images, respectively; $|I|$ denotes their number of pixels.

Results for the stripes images are shown in Fig. 2. As expected from the developments in Section 2, the minimum values of the RMSE are found at about $\sigma/2$ for the Gaussian kernel while they are around σ for the χ_D CCDF and Burr CCDF. The Burr CCDF approximation leads to results similar to those obtained with the computationally more intensive χ_D CCDF. As this behavior is very similar for all tested images, and because of space limitations, other figures are not shown here. RMSE results for all images are presented in Table 4. For all noise levels, the lowest RMSE values are achieved using the Chi or the Burr kernel.

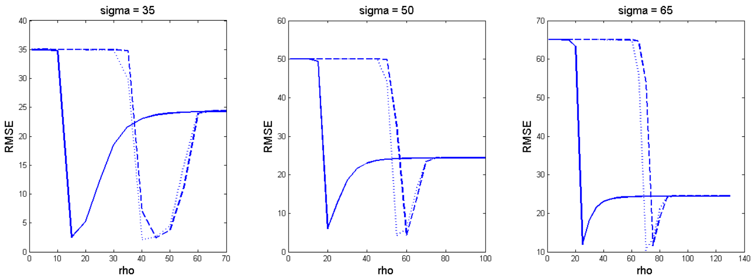

Fig. 2: Denoising results for the stripes image (RMSE vs range kernel width). The plain line shows the Gaussian kernel results, the dotted line the χ_D CCDF results, and the dashed line the results obtained with the Burr CCDF. Left. Gaussian noise, std = 35. Middle. Gaussian noise, std = 50. Right, Gaussian noise, std = 65.

	Stripes			Stair			Barbara		
ν	35	50	65	35	50	65	10	15	20
Gauss	2.49	6.00	11.92	2.24	2.99	3.62	7.41	9.88	12.72
Chi	2.08	4.12	10.35	2.08	2.88	3.52	7.31	9.72	11.30
Burr	2.39	4.16	11.55	2.17	2.92	3.59	7.32	9.82	11.50

Table 1: Minimum RMSEs for each image, kernel, and noise level.

5 Conclusions

Mode estimation is an effective denoising method. It can be achieved by running a hill-climbing procedure on a kernel density estimator. This paper investigates the behavior of this technique in high-dimensional spaces. In particular, we propose kernels that are specifically crafted to deal with this case. Experiments with several image denoising tasks show that the proposed kernels outperform the widely used Gaussian function.

References

[1] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, Sep 1962.

[2] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics*. Wiley Series in Probability and Mathematical Statistics. Wiley & Sons, New York., 2005.

[3] B. U. Park and B. A. Turlach. Practical performance of several data driven bandwidth selectors. *Computational Statistics*, 7:251–270, 1992.

[4] R. Cao, A. Cuevas, and W. G. Manteiga. A comparative study of several smoothing methods in density estimation. *Computational Statistics and Data Analysis*, 17:153–176, 1994.

[5] M. Farmen and J. S. Marron. An assessment of finite sample performance of adaptive methods in density estimation. *Comp. Stat. and Data Analysis*, 30:143–168, 1999.

[6] D. Francois, V. Wertez, and M. Verleysen. About the locality of kernels in high-dimensional spaces. In *International Symposium on Applied Stochastic Models and Data Analysis*, pages 238–245, Brest, France, May 2005.

[7] D. Francois. *High-dimensional data analysis : from optimal metrics to feature selection*. Verlag-VDM, 2008.

[8] A. Buades, B. Coll, and J.M. Morel. A review of image denoising algorithms, with a new one. *Multiscale Modeling & Simulation*, 4(2):490–530, 2005.

[9] S. A. Awate and R. T. Whitaker. Image denoising with unsupervised, information-theoric, adaptative filtering. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(3):364 – 376, 2006.

[10] C. Kervrann, J. Boulanger, and P. Coup. Bayesian non-local means filter, image redundancy and adaptative dictionaries for noise removal. In *Conf. Scale-Space and Variational Methods*, volume 4485, pages 520–532, Ischia, Italy, May 2007.

[11] G. Arfken. *The Incomplete Gamma Function and Related Functions*, volume 10.5 of *3rd ed.* 1985.

[12] I.W. Burr. Cumulative frequency functions. *Annals of Mathematical Statistics*, 13:215–232, 1942.

[13] C. Kervrann and J. Boulanger. Unsupervised patch-based image regularization and representation. In *In Proc. European Conf. Comp. Vision (ECCV'06)*, pages 555–567, Graz , Austria, May 2006.

[14] C. Kervrann and J. Boulanger. Optimal spatial adaptation for patch-based image denoising. *IEEE Trans. on Image Processing*, 15(10):2866–2878, 2006.

[15] C.-A. Deledalle, L. Denis, and F. Tupin. Iterative weighted maximum likelihood denoising with probabilistic patch-based weights. *IEEE Trans. on Image Processing*, To appear, 2009.

[16] B. Goossens, H. Luong, A. Pizurica, and W. Philips. An improved non-local denoising algorithm. In *Int. Workshop on Local and Non-Local Approximation in Image Processing*, pages 25–29, Lausanne, Switzerland, 2008.

[17] P. Chatterjee and P. Milanfar. A generalization of non-local means via kernel regression. In *SPIE Conference Series*, volume 6814, Mar 2008.

[18] T. Brox, O. Kleinschmidt, and D. Cremers. Efficient non-local means for denoising of textural patterns. *IEEE Trans. on Image Processing*, 17(7):1083–1092, 2008.

[19] T. Tasdizen. Principal components for non-local means image denoising. In *IEEE Int. Conf. on Image Processing*, pages 1728–1731, San Diego, California, Oct 2008.