# Locally Weighted Full Covariance Gaussian Density Estimation

**Pascal Vincent and Yoshua Bengio**
Département d'Informatique et Recherche Opérationnelle
Centre de Recherche Mathématiques
Université de Montréal
Montréal, Québec, Canada, H3C 3J7
{vincentp,bengioy}@iro.umontreal.ca

**Technical Report 1240**,
Département d'Informatique et Recherche Opérationnelle

## Abstract

We describe an interesting application of the principle of local learning to density estimation. Locally weighted fitting of a Gaussian with a regularized full covariance matrix yields a density estimator which displays improved behavior in the case where much of the probability mass is concentrated along a low dimensional manifold. While the proposed estimator is not guaranteed to integrate to 1 with a finite sample size, we prove asymptotic convergence to the true density. Experimental results illustrating the advantages of this estimator over classic non-parametric estimators are presented.

## 1 Introduction

Most machine-learning problems, as they occur in nature, are posed in a very high dimensional space. However, overcoming the curse of dimensionality has been an open problem since it was first described in [2]. The popularity of the new generation of kernel methods, in particular the Support Vector Machines [3, 17], is due in part to relatively good performance on high dimensional problems, while the traditional kernel methods (s.a. Parzen windows [13]) often perform more poorly. Another, recently revived, and very promising research trend in dealing with the curse, is that of manifold learning. It is based on the idea that the data lives on (or close to) a non-linear manifold of much lower dimensionality, embedded in the high dimensional space. This trend is exemplified by Locally Linear Embedding [14] and Isomap [15] but also underlies the idea of mixtures of factor analyzers and similar algorithms [7, 8, 16, 6].

Our line of research attempts to integrate the notions of manifold modeling with the traditional non-parametric kernel and distance based methods such as k-nearest-neighbors and Parzen windows. We have already proposed improved algorithms for classification [18] and density estimation [19] but the latter one suffers from serious practical difficulties[1]. In this paper, we propose a different approach to density estimation which does not pose the same memory

[1] Its memory requirement scales in $\mathcal{O}(n.d^2)$ where $d$ is the input dimensionality, and $n$ is the number of training samples, making it impossible to use with large, high dimensional data sets.

requirement problems as [19], and is based on a general principle taking the point of view of local learning.

*Local learning* [4, 1, 12]. can be understood as a general principle that allows to extend learning techniques designed for simple models, to the case of complex data for which the model's assumptions would not necessarily hold globally, but can be thought as valid *locally*. A simple example is the assumption of linear separability, which in general is *not* satisfied globally in classification problems with rich data. Yet any classification algorithm able to find only a linear separation, can be used inside a local learning procedure, yielding an algorithm able to model complex non-linear class boundaries.

Similarly, for density estimation, while it is in general unreasonable to assume that the data follows a Gaussian distribution globally, the Gaussian approximation holds locally. Note that if the data lies close to a low dimensional manifold, then the shape of that local Gaussian will be a flattened pancake, and it's crucial to use a non-spherical Gaussian, to capture the local principal directions of the manifold.

Traditional parametric density estimation can be formulated as the question: "What is the likelihood of a test point $x$ under a model fitted to the whole training data". We formulate the principle of locally weighted density estimation in a similar manner as "What is the likelihood of a test point $x$ under a *simple model* fitted only to the *local* training data in the neighborhood of $x$."

Notice that locally weighted density estimation yields an unnormalized density estimate: in general it won't integrate to 1, as is also the case of several classical non-parametric density estimators, similar in spirit, like the nearest neighbor density estimator [5, 11]. (see [10] for a survey of non-parametric density estimation techniques).

Local learning typically comes in two flavors, depending on the notion of "neighborhood". The neighborhood is always based on some a-priori measure of locality (such as the Euclidean distance in input space), but it can be either defined as a "hard" neighborhood (the set of $k$ nearest neighbors of $x$ for instance), or as a "soft", weighted neighborhood (the set of all training points, but with an associated weight given by a prior, continuous weighting kernel $\mathcal{K}$, centered on $x$). The former can be seen as a special case of the latter, with a particular discontinuous weighting kernel giving only weights of $0$ or $1$). We would like to stress the importance of using a "soft" neighborhood to avoid discontinuities in the estimate, a problem that plagues the nearest neighbor density estimator. Indeed, while *some* statistics of the set of $k$ neighbors (such as the distance to the $k^{th}$ neighbor) vary smoothly with $x$, the set of $k$ neighbors doesn't: a small variation in $x$ may yield a totally different $k^{th}$ neighbor, and thus lead to a discontinuous estimate, if a "hard" neighborhood is used. Consequently the local model fitting procedure should accommodate sample weights.

For previous work on local learning applied to non-parametric density estimation, see also [9].

## 2 The Locally Weighted Density Estimatior

Let $D = \{x_1, \ldots, x_n\}$ a data set with $x_i$ sampled i.i.d. from an unknown distribution with continuous density $f(x)$.

Let $\mathcal{K}(x_i; x)$ a weighting Kernel centered on $x$, used to give a weight to every $x_i$.

In addition, we suppose that it is easy to fit a simple parametric model $\mathcal{M}$ to a data set endowed with sample weights (e.g. the maximum of the log-likelihood can be found analytically or cheaply).

The locally weighted density estimation principle computes an estimate $\hat{f}(x)$ of the density at a given test point $x$ as follows:

- Associate a weight $\mathcal{K}(x_i; x)$ to every $x_i \in D$
- Fit model $\mathcal{M}$ to the weighted data.
- The estimate is $\hat{f}(x) = \frac{1}{Z}\mathcal{M}(\S)$ with $Z$ a normalization factor to try to make $\hat{f}(x)$ integrate to 1 over the domain of $x$ (at least asymptotically).

In our particular case we use $\mathcal{K}(x_i; x) = \mathcal{N}(x_i; x, \sigma_n^2(x)I)$, with $\sigma_n(x) = \alpha d(x, x_{v_k})$, where $x_{v_k}$ denotes the $k^{th}$ neighbor of $x$ according to the Euclidean distance $d$. i.e. the *width* of our weighting kernel is proportional to the distance from its center to its $k^{th}$ neighbor.

The model $\mathcal{M}$ we fit to the weighted training samples is a Gaussian with a regularized full covariance matrix, i.e.: $\mathcal{M}(\S) = \mathcal{N}(\S; \mu_\backslash(\S), \mathcal{C}_\backslash(\S))$, where $\mu_n(x) = \frac{1}{\sum_i \mathcal{K}(x_i;x)} \sum_i \mathcal{K}(x_i; x)x_i$ is the weighted sample mean, and $C_n(x) = \frac{1}{\sum_i \mathcal{K}(x_i;x)} \sum_i \mathcal{K}(x_i; x)(x_i - \mu_n(x))(x_i - \mu_n(x))' + \gamma_n^2 I$ is the weighted sample covariance with an additional regularization parameter $\gamma_n$.

$\mathcal{N}(x; \mu, \Sigma)$ denotes the density of $x$ under a multivariate Normal density with center $\mu$ and covariance matrix $\Sigma$:

$$\mathcal{N}(x; \mu, \Sigma) = \frac{e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}}{(2\pi)^{d/2}|\Sigma|^{1/2}} \tag{1}$$

We denote the resulting estimator $f^{WN}$,

$$f^{WN}(x) = \frac{1}{Z}\mathcal{N}(x; \mu_n(x), C_n(x)), \tag{2}$$

and will prove its asymptotic convergence to the true density $f(x)$, when the normalization factor is estimated with $Z = \frac{nV_d(1)}{k_n(2\pi)^{d/2}\alpha^d}$, with $V_d(r)$ being the volume of a $d$-ball of radius $r$ in $\mathbf{R}^d$, i.e. $V_d(r) = r^d V_d(1) = r^d \frac{\pi^{d/2}}{\Gamma(1+\frac{d}{2})}$.

A slight variant of the above $f^{WN}$ will also be considered (in particular for the convergence proof) in which $\mu_n(x)$ is fixed on $\mu_n(x) = x$ rather than on the weighted sample average.

## 3 Other Classical Estimators Considered

In addition we'll consider the following classical non-parametric estimators, using the above definitions.

1. The fixed width Parzen density estimator [13]:

$$f^{FP}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{N}(x; x_i, r_n^2 I). \tag{3}$$

2. The fixed width nearest neighbor density estimator [13]:

$$f^{FN}(x) = \frac{|N(x, r_n)|}{nV_d(r_n)} \tag{4}$$

where $N(x, r_n) = \{x_i \in D : ||x - x_i|| < r_n\}$ are the neighbors of $x$ closer than $r_n$ and $\frac{|N(x,r_n)|}{n}$ estimates unbiasedly the probability of falling in the $d$-ball of radius $r_n$ around $x$.

3. The k-nearest neighbor density estimator [5, 11]:

$$f^{KN}(x) = \frac{k_n}{nV_d(\sigma_n(x))} \tag{5}$$

where $\frac{k_n}{n}$ estimates the probability of falling in the $d$-ball of radius $\sigma_n(x)$ around $x$, and we select $\alpha = 1$ in $\sigma_n(x)$.

# 4 Asymptotic Convergence Analysis

To prove convergence, a number of the quantities defined above will be subscripted with $n$. We have to impose conditions on the rate of modification of $k_n$ which must increase with $n$, and the way it affects $\sigma_n$ i.e. $\lim_{n\to\infty} k_n = \infty$ but slower than $n$, i.e. $\lim_{n\to\infty} \frac{k_n}{n} = 0$. We let $k_n$ increase at a slow enough rate such that $n\sigma_n(x)^d \to \infty$. Besides, $\gamma_n$ must go to zero faster than $\sigma_n$.

We already know from the classical litterature [13, 5, 11] that $f^{FP}$, $f^{FN}$ and $f^{KN}$ are consistent i.e. that their limit as $n \to \infty$ is $f$ for every $x$ (with the hyperparameters converging at a proper rate, as defined above).

**Lemma 1** . *Let $r_n$ be a probabilistic lower bound on $\sigma_n$, i.e. $\sigma_n(x) > r_n$ with high probability $1-\delta$. Let $g(x,y)$ be a function that is $O(y^m)$ and is $O(\sigma_n(x)^{-p})$. The limit of a locally weighted average of $g(x,x_i)$ with $\sigma_n(x)$ converges to the same limit as $f(x)\int g(x,y)\mathcal{N}(y;x,\sigma_n^2(x)I)dy$:*

$$\lim_{n\to\infty}\frac{1}{n}\sum_i g(x,x_i)\mathcal{N}(x_i;x,\sigma_n^2(x)I) = f(x)\lim_{n\to\infty}\int g(x,y)\mathcal{N}(y;x,\sigma_n^2(x)I)dy \qquad (6)$$

*To obtain convergence, then $r_n$ must decrease at a slow enough rate such that $nr_n^{d+2(p-m)} \to \infty$.*

**Sketch of the Proof**

*Consider the left-hand-side of 6 with $\sigma_n(x)$ replaced by $r_n$ (even in $g$). Its expected value is $\int f(y)g(x,y)\mathcal{N}(y;x,r_n(x)^2I)dy$. Let us show that the average converges to its expected value. For this we will show that the average error (i.e. the variance) goes to 0. Below we will show that for a data-independent spread $r_n$ in the local weighting kernel $\mathcal{N}$, the variance of the average is inversely proportional to a power of $r_n$. Therefore with probability $1-\delta$, the variance of the average using $\sigma_n(x)$ is less than the variance of the average using $r_n$. Since the latter will be shown to converge to zero we will obtain convergence to zero of the variance of the desired average. Consider the variance of the average with data-independent spread $r_n$:*

$$\begin{aligned} Var[\frac{1}{n}\sum_i g(x,x_i)\mathcal{N}(x_i;x,r_n^2I)] &= \frac{Var[g(x,x_i)\mathcal{N}(x_i;x,r_n^2I)]}{n} \\ &< \frac{\int f(y)g(x,y)^2\mathcal{N}(y;x,r_n^2)^2dy}{n} \end{aligned}$$

*where the variance is over $D$, and we simply dropped $E[g(x,x_i)\mathcal{N}(x_i;x,r_n^2I)]^2$. Let us now make the change of variable $z = \sqrt{2}(y-x)/r_n$ (this is a vector in $\mathbf{R}^d$, i.e. $dy_1\ldots dy_d = 2^{-d/2}r_n^d dz_1\ldots dz_d$), yielding the bound on the variance*

$$\frac{1}{(2\pi)^{d/2}n2^{d/2}r_n^d}\int f(\frac{r_nz}{\sqrt{2}}+x)g(x,\frac{r_nz}{\sqrt{2}}+x)^2\mathcal{N}(z;0,1)dz$$

*using $\mathcal{N}(y;x,r_n^2)^2 = \frac{e^{-||y-x||^2/r_n^2}}{(2\pi)^d r_n^{2d}} = \frac{1}{(2\pi)^{d/2}r_n^{2d}}\frac{e^{-\frac{||z||^2}{2}}}{(2\pi)^{d/2}} = \frac{1}{(2\pi)^{d/2}r_n^{2d}}\mathcal{N}(z;0,1)$. If $g$ varies with $n$ in $\frac{1}{r_n^p}$ and in the power $m$ of its second argument, then the integrand varies in $\frac{1}{r_n^{2(p-m)}}$: the condition on $r_n$ is indeed $nr_n^{d+2(p-m)} \to \infty$.*

*Q.E.D.*

Note that as a special case we obtain the convergence of $f^{FP}$, with $g(x,x_i) = 1$.

**Lemma 2** *If $n\sigma_n^{d+1}(x) \to \infty$ and $\sigma_n(x) \to 0$ the locally weighted version of $\mu_n(x)$ converges to $x$ and $\frac{\mu_n(x)-x}{\sigma_n(x)}$ converges to 0, in probability:*

$$\lim_{n\to\infty} \frac{\sum_i x_i \mathcal{N}(x_i; x, \sigma_n^2(x)I)}{\sum_i \mathcal{N}(x_i; x, \sigma_n^2(x)I)} = x$$

*and*

$$\lim_{n\to\infty} \frac{\mu_n(x) - x}{\sigma_n(x)} = 0,$$

*in probability.*

**Proof**

*The denominator of $\mu_n(x)$ times $\frac{1}{n}$ is $f^{FP}$, which converges to $f(x)$. For the numerator, we apply Lemma 1 with $g(x, x_i) = \frac{x_i - x}{\sigma_n(x)}$ (i.e. $p = m = 1$) and obtain $f(x) \lim_{n\to\infty} \int \frac{y-x}{\sigma_n(x)} \mathcal{N}(y; x, \sigma_n(x)^2 I) dy$. We apply the change of variable $z = \frac{y-x}{\sigma_n(x)}$ and obtain $f(x) \lim_{n\to\infty} \int \sigma_n(x)^d z \mathcal{N}(z; 0, I) dz = 0$.*

*Q.E.D.*

**Lemma 3** *The locally weighted covariance $C_n(x)$ has the same limit as $\sigma_n^2(x)I$.*

**Proof**

*We rewrite the numerator and denominator in the first term of $C_n$ as averages. Using Lemma 1 and Lemma 2 the numerator has the same limit as*

$$f(x) \int (x-y)(x-y)' \mathcal{N}(y; x, \sigma_n^2(x)I) dy.$$

*Apply the change of variable $z = y - x$, yielding*

$$f(x) \int zz' \mathcal{N}(z; 0, \sigma_n(x)^2 I) dz \to f(x)\sigma_n(x)^2 I.$$

*As in previous proofs, the denominator converges to $f(x)$.*

$$C_n(x) \to (\sigma_n^2(x) f(x) + \gamma_n^2) I$$

*Since we have assumed $\frac{\gamma_n}{\sigma_n} \to 0$, the second term can be ignored.*

*Q.E.D.*

**Theorem 1** *The locally weighted full covariance matrix estimator $f^{WN}$ is consistent (converges to $f$) for both versions of $\mu_n(x)$.*

**Proof**

*Consider the numerator of $\mathcal{N}(x; \mu_n(x), C_n(x))$ in eq. 1. When $\mu_n(x) = x$ it is simply equal to $e^0 = 1$. For the other versions, Lemma 2 shows that it converges to 1. Using Lemma 3, and $|aB| = a^d|B|$ for a $d \times d$ matrix $B$ and a scalar $a$, the denominator $(2\pi)^{d/2}\sqrt{|C_n(x)|}$ has the same limit as*

$$(2\pi)^{d/2} \sqrt{|\sigma_n^2(x) I|} = (2\pi)^{d/2} \sigma_n^d(x).$$

*Finally, we use the main result in [11], i.e.*

$$\lim_{n\to\infty} \frac{k_n}{n V_d(1) \sigma_n^d(x)} = f(x).$$

*when $\alpha = 1$, i.e. equal to $f(x)/\alpha^d$ when using a different value of $\alpha$. Putting these together with the formula for $f^{WN}$, we obtain*

$$\begin{aligned}\lim_{n\to\infty} f^{WN}(x) &= \lim_{n\to\infty} \frac{k_n \alpha^d (2\pi)^{d/2}}{nV_d(1)} \mathcal{N}(x; \mu_n(x), C_n(x)) \\ &= \lim_{n\to\infty} \frac{k_n \alpha^d (2\pi)^{d/2}}{nV_d(1)(2\pi)^{d/2}\sigma_n^d(x)} \\ &= f(x).\end{aligned}$$

*Q.E.D.*

## 5 Experiments

To assess the performance of the proposed algorithm, we performed the following experiment on a 2D spiral problem:

A training set of 300 points, a validation set of 300 points (reserved for tuning hyper-parameters), and a test set of 10000 points were generated from the following distribution of two dimensional $(x, y)$ points:

$$x = 0.04\ t\ \sin(t) + \epsilon_x,\ \ y = 0.04\ t\ \cos(t) + \epsilon_y$$

where $t \sim U(3, 15)$, $\epsilon_x \sim \mathcal{N}(0, 0.01)$, $\epsilon_y \sim \mathcal{N}(0, 0.01)$, $U(a, b)$ is uniform in the interval $(a, b)$ and $\mathcal{N}(\mu, \sigma)$ is a normal density.

As the density estimators under consideration are not guaranteed to integrate to 1 (except for $f^F P$), we compute an approximation of their integral by sampling on a $300 \times 300$ regularly spaced grid (covering the $x$ and $y$ range of the training set plus $10\%$), and divide the raw estimators by the obtained integral approximation, yielding normalized estimators for our comparison study. Evaluation of the estimators over that same grid is also used to produce the graphs of Figure 1. The performance of each normalized estimator $\hat{f}$ is then evaluated using the average log likelihood (ALL) over the test set $T$ (of size $m = 10000$):

$$ALL(\hat{f}, T) = \frac{1}{m} \sum_{x \in T} \log \hat{f}(x)$$

For each estimator type, we tried several values of the hyper-parameters, keeping the choice that yielded the largest possible ALL over the validation set.

Table 1: Performance of various estimators on the spiral data, measured as average log-likelihood over the test set (standard errors are in parenthesis).

| Estimator | Hyper-parameters used | ALL on test set |
|---|---|---|
| $f^{FP}$ | $r = 0.0151$ | 1.292 (0.012) |
| $f^{KN}$ | $k = 2$ | 1.058 (0.011) |
| $f^{FN}$ | $r = 0.065$ | 0.739 (0.018) |
| $f^{WN}$ fixed $\sigma$ | $\sigma = 0.04, \gamma = 6e - 5$ | 1.461 (0.006) |
| $f^{WN}$ | $\alpha = 0.23, k = 25, \gamma^2 = 3e - 5$ | 1.561 (0.008) |
| $f^{WN}$ fixed $\sigma$, $\mu = x$ | $\sigma = .045, \gamma^2 = 1e - 9$ | 0.759 (0.003) |
| $f^{WN}, \mu = x$ | $\alpha = 1, k = 2, \gamma^2 = 1e - 5$ | 1.201 (0.010) |

Results are reported in Table 1. $f^{WN}$ (with local $\sigma(x)$ and using the local average for $\mu(x)$) appears to perform significantly better than the other classical estimators, indicating that it was able to more accurately capture and model the underlying true distibution. The difference between $f^{FP}$ and $f^{WN}$ can also be better appreciated qualitatively in Figure 1. For $f^{WN}$, the experiments suggest it is better to use $\alpha < 1$, which yields a less variable $\sigma(x)$, and it is better to use an adaptive $\sigma(x)$ than a fixed $\sigma$. The validation hyper-parameter selection also chooses a non-zero $\gamma$, which suggests that it is also useful.

## 6 Conclusions

To summarize, we have introduced and analyzed a family of non-parametric locally weighted density estimators that are appropriate to model the local manifold structure of data, and experiments suggest that it performs well against classical estimators, in addition to being much more memory-efficient than the related estimator proposed in [19].

## References

[1] C. G. Atkeson, A. W. Moore, and S. Schaal. Locally weighted learning for control. *Artificial Intelligence Review*, 11:75–113, 1997. 2

[2] R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, New Jersey, 1961. 1

[3] B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, 1992. 1

[4] L. Bottou and V. Vapnik. Local learning algorithms. *Neural Computation*, 4(6):888–900, 1992. 2

[5] E. Fix and J.L. Hodges. Discriminatory analysis, non-parametric discrimination, consistency properties. Technical Report Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1951. 2, 3, 4

[6] Z. Ghahramani and M. J. Beal. Variational inference for Bayesian mixtures of factor analysers. In *Advances in Neural Information Processing Systems 12*, Cambridge, MA, 2000. MIT Press. 1

[7] Z. Ghahramani and G.E. Hinton. The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, Dpt. of Comp. Sci., Univ. of Toronto, 21 1996. 1

[8] G.E. Hinton, M. Revow, and P. Dayan. Recognizing handwritten digits using mixtures of linear models. In G. Tesauro, D.S. Touretzky, and T.K. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 1015–1022. MIT Press, Cambridge, MA, 1995. 1

[9] N. L. Hjort and M. C. Jones. Locally parametric nonparametric density estimation. *Annals of Statistics*, 24(4):1619–1647, 1996. 2

[10] A.J. Inzenman. Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, 86(413):205–224, 1991. 2

[11] D.O. Loftsgaarden and C.P. Quesenberry. A nonparametric estimate of a multivariate density function. *Annals of Mathematical Statistics*, 36:1049–1051, 1965. 2, 3, 4, 5

[12] D. Ormoneit and T. Hastie. Optimal kernel shapes for local linear regression. In S.A. Solla, T.K. Leen, and K-R. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000. 2

[13] E. Parzen. On the estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1064–1076, 1962. 1, 3, 4

[14] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, Dec. 2000. 1

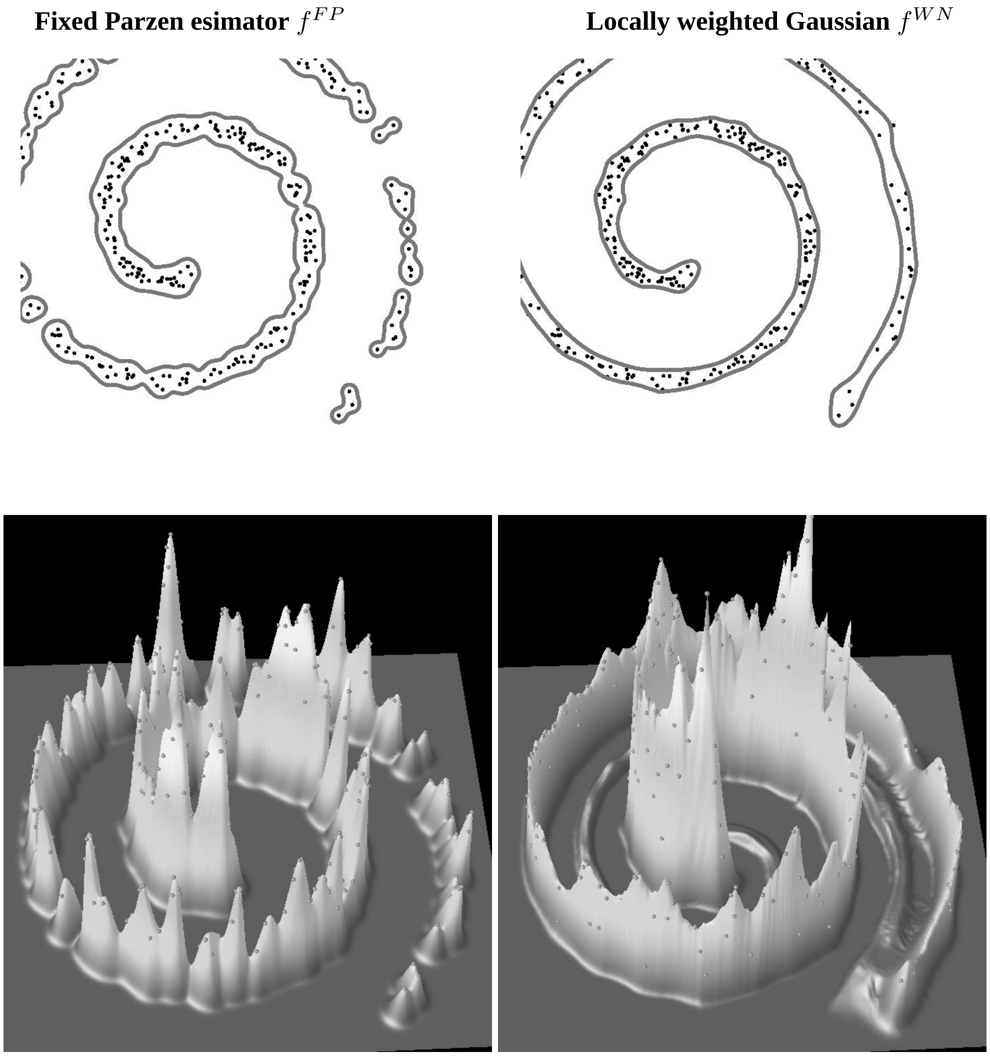
[15] J. Tenenbaum, V. de Silva, and J.C.L. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, Dec. 2000. 1

[16] M.E. Tipping and C.M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482, 1999. 1

[17] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995. 1

[18] P. Vincent and Y. Bengio. K-local hyperplane and convex distance nearest neighbor algorithms. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, Cambridge, MA, 2002. The MIT Press. 1

[19] P. Vincent and Y. Bengio. Manfold parzen windows. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15, Cambridge, MA, 2003. The MIT Press. 1, 2, 7



**Fixed Parzen esimator** $f^{FP}$

**Locally weighted Gaussian** $f^{WN}$

Figure 1: Illustration of the density estimated by ordinary Parzen Windows (left) and locally weighted Gaussian (right). The top images show the 300 training points together with an isoline corresponding to a normalized density estimate of 1. The bottom images show the estimated densities as the elevation. The $f^{WN}$ estimate appears much sharper along the manifold (thinner walls) and significantly less bumpy. It appears better able to capture the structure of the underlying distribution, and to successfully "extrapolate" in regions with few data points but high true density. $f^{FP}$ on the contrary, appears to waste more probability mass away from the manifold (due to its clearly visible *spherical bump* nature).