Computer Science Technical Report TR2010-934, October 2010
Courant Institute of Mathematical Sciences, New York University
http://cs.nyu.edu/web/Research/TechReports/reports.html

An Empirical Bayesian interpretation and generalization of NL-means

Martin Raphan
Courant Inst. of Mathematical Sciences
New York University
raphan@cims.nyu.edu

Eero P. Simoncelli
Howard Hughes Medical Institute,
Center for Neural Science, and
Courant Inst. of Mathematical Sciences
New York University
eero.simoncelli@nyu.edu

Abstract

A number of recent algorithms in signal and image processing are based on the empirical distribution of localized patches. Here, we develop a nonparametric empirical Bayesian estimator for recovering an image corrupted by additive Gaussian noise, based on fitting the density over image patches with a local exponential model. The resulting solution is in the form of an adaptively weighted average of the observed patch with the mean of a set of similar patches, and thus both justifies and generalizes the recently proposed *nonlocal-means* (NL-means) method for image denoising. Unlike NL-means, our estimator includes a dependency on the size of the patch similarity neighborhood, and we show that this neighborhood size can be chosen in such a way that the estimator converges to the optimal Bayes least squares estimator as the amount of data grows. We demonstrate the increase in performance of our method compared to NL-means on a set of simulated examples.

1 Introduction

Local patches within a photographic image are often highly structured. Moreover, it is quite common to find many patches within the same image that have similar or even identical content. This fact has been exploited in developing new forms of signal representation [10], and in applications such as denoising [5], clustering [3], inpainting [4], and texture synthesis [15, 7, 8]. Most of these methods are based on assumptions (sometimes implicit) about the probability distribution over the intensity vectors associated with the patches. A recent example is a patched-based inference methodology, known as nonlocal means (NL-means), in which each pixel in a noisy image is replaced by a weighted average over other pixels in the image that occur in similar image patches [1, 2]. The weighting is determined by the similarity of the patches, as measured with a Gaussian kernel. This method is intuitively appealing, and gives good (although less than state-of-the-art) denoising performance on natural images. More recent variants have been developed that achieve state-of-the-art [5]. But as we show here, the NL-means has the drawback that the estimate depends crucially on the Gaussian kernel width, and even in the limit of infinite data, there is no way to choose this width so as to guarantee convergence to the optimal answer.

In this paper, we develop an alternative patch-based denoising algorithm as an empirical Bayes (EB) estimator, and show that it generalizes and improves on the NL-means solution. We use an exponential density model for local patches of a noisy image, and derive a maximum likelihood method for estimating the parameters from data. We then combine this with a nonparametric EB form of the optimal least-squares estimator for additive Gaussian noise [13]. This simple, but often overlooked form expresses the usual Bayes least squares estimator directly in terms of the density of noisy observations, without reference to the (clean) prior density. The resulting denoising method resembles NL-means, in that the denoised value is a function of the weighted average over other pixels in the same image that come from similar patches. But the function is nonlinear, and is scaled according to the size of the neighborhood over which intensities are averaged. As a result, we prove that the estimator converges to the Bayes least squares optimum as the amount of data increases, a feature not present in the standard NL-means approach.

2 NL-means estimation

Suppose Y is a noise-corrupted measurement of X, where either or both variables can be finite-dimensional vectors. Given this measurement, the estimate of X that minimizes the expected squared error is the Bayes Least Squared (BLS) estimator, which is the expected value of X conditioned on Y, $E\{X|Y\}$. In particular, for an observation $\mathbf{y}_0$, the estimate is

$$\hat{x}_{\mathrm{BLS}}(\mathbf{y}_0) = E\{X|Y = \mathbf{y}_0\}.$$

In the NL-means approach, a vector observation $Y = \mathbf{y}_0$ is denoised by averaging it together with patches that are "similar", i.e., that are in some neighborhood, N_h, of size h about the vector $\mathbf{y}_0$. As the number of data points available increases to infinity, the NL-means estimator will converge to

$$\hat{x}_{\mathrm{NL}}(\mathbf{y}_0) \to E\{Y|Y \in N_h\}. \tag{1}$$

Notice that this is a conditional expectation of Y, not X (as required for the BLS estimate). The original presentation of NL-means [1] justifies this by considering the case of estimating the center pixel, X_c of the patch based on the surrounding pixels. In this case, and assuming zero-mean noise, we can write

$$E\{Y_c|Y_{-c} = \mathbf{y}_{-c}\} = E\{E\{Y_c|X_c, Y_{-c} = \mathbf{y}_{-c}\}|Y_{-c} = \mathbf{y}_{-c}\} = E\{X_c|Y_{-c} = \mathbf{y}_{-c}\}, \tag{2}$$

where Y_{-c} indicates the patch without the center pixel. Thus, by setting N_h to be a neighborhood which does not depend on the central coefficient of the vector, Eq. (1) will tend to the optimal estimator of the central coefficient as a function of the surrounding coefficients, Eq. (2).

In practice, NL-means is implemented to denoise the central coefficient in the patch, but includes the central coefficient in determining similarity of patches. Under these conditions, the solution does not, in general, converge to the correct BLS estimator, regardless of the size of the neighborhood. To see this, consider the simplified situation where the neighborhood is defined as

$$N_h = \{\mathbf{y} : |\mathbf{y} - \mathbf{y}_0| \leq h\}. \tag{3}$$

Clearly, $\hat{x}_{\mathrm{NL}}(\mathbf{y}_0)$ will depend on the value of h. For a fixed value of h, we end up with a conditional expectation of Y instead of X, as mentioned above. And, in the limit as $h \to 0$, the estimate will converge to the identity, i.e. $\hat{x}_{\mathrm{NL}}(\mathbf{y}_0) \to \mathbf{y}_0$. Intuitively, this is because the mean of data within a neighborhood must also lie within that neighborhood. Thus, the best we can hope for is to find some binwidth, h, that works reasonably well asymptotically, but this choice will be highly dependent on the (unknown) signal prior.

This is illustrated for a simple scalar example in Fig. 1. The prior here is made up of delta functions which are well separated, relative to the standard deviation of the noise. With the right choice of binwidth (about twice the width of the noise distribution), for any amount of data, NL-means provides a good approximation to the optimal denoiser. But note that the NL-means behavior depends critically on the choice of binwidth: Fig. 1 also shows a result for a smaller binwidth, which is clearly suboptimal. In general, there is no simple way to choose an optimal binwidth, and to do so in a way that makes the NL-means estimator converge to the BLS estimator.

3 Patch-based empirical Bayes least squares estimator

We wish to develop a patch-based estimator that makes explicit the dependence on neighborhood size, h, and for which we can demonstrate proper convergence in the limit as $h \to 0$. For the particular case of additive Gaussian noise, Miyasawa [13] developed an nonparametric empirical form of the Bayes least squares estimate that is written directly in terms of the measurement (noisy) density $P_Y(Y)$, with no explicit reference to the prior:

$$E\{X|Y = \mathbf{y}_0\} \quad = \quad \mathbf{y}_0 + \Lambda \nabla_{\mathbf{y}} \ln(P_Y(\mathbf{y}_0)), \tag{4}$$

where Λ is the covariance matrix of the noise (assumed known). Note that this expression is *not* an approximation: the equality is exact, assuming the noise is additive and Gaussian (with zero mean and covariance Λ), and assuming the measurement density is known.

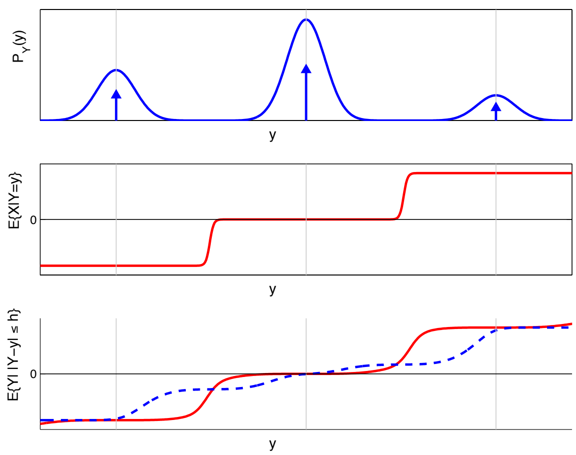


Fig. 1: Observed density and optimal estimator for prior made up of three isolated delta functions. Top panel: observed density (arrows represent prior). Middle panel: BLS estimator (assuming true prior is known). Bottom: NL-means estimates calculated for two different binsizes (thick solid and dashed lines). Note also that the NL-means estimator approaches the identity function as the binsize goes to zero.

3.1 Learning the estimator function from data

The formulation of Eq. (4) offers a means of computing the BLS estimator from noisy samples, if one can construct an approximation of the noisy distribution, $P_Y(Y)$. Simple histograms will not suffice for this approximation, because Eq.(4) requires us to compute the logarithmic derivative of the estimated density. Instead, we use an exponential model for the local density centered on an observation $Y_k = \mathbf{y}_0$ within some neighborhood, N_h of size h, centered on $\mathbf{y}_0$:

$$P_{Y|Y\in N_h}(\mathbf{y}) \quad = \quad \frac{e^{\mathbf{a}(\mathbf{y}-\mathbf{y}_0)}}{Z(\mathbf{a},h)}\mathbf{1}_{N_h} \tag{5}$$

where $\mathbf{1}_{N_h}$ denotes the indicator function of N_h, and $Z(\mathbf{a},h)$ normalizes the density over the interval:

$$Z(\mathbf{a},h) = \int_{N_h} e^{\mathbf{a}\cdot(\mathbf{y}-\mathbf{y}_0)}dy. \tag{6}$$

This type of local exponential approximation is similar to that used in [11] for density estimation. The shape of the neighborhood, N_h, is unspecified, but the linear size should scale with h. For example, it could be a spherical ball of radius h around $\mathbf{y}_0$ or the hypercube centered on $\mathbf{y}_0$ of length $2h$ on each side. In any case, the logarithmic gradient of this density evaluated at $\mathbf{y}_0$ is simply $\mathbf{a}$.

We can then fit the parameter $\mathbf{a}$ by maximizing the likelihood of the conditional density over the neighborhood:

$$\hat{\mathbf{a}} \quad = \quad \arg\max_{\mathbf{a}} \sum_{\{n:Y_n\in N_h\}} \ln(P_{Y|Y\in N_h}(Y_n)) \tag{7}$$

Substituting Eq. (5), taking the gradient of the sum, and setting it equal to zero gives

$$\nabla_{\mathbf{a}} \ln(Z(\hat{\mathbf{a}},h)) \quad = \quad \bar{Y} - \mathbf{y}_0 \tag{8}$$

where $\bar{Y}$ is the average of the data that lie in N_h. Note that substituting Eq. (6) in for $Z(\hat{a},h)$ and adding $\mathbf{y}_0$ to both sides gives

$$E\{Y|Y\in N_h\} = \bar{Y}.$$

Thus, the ML solution for $\hat{\mathbf{a}}$ matches the local mean of the fitted density to the local empirical mean.

Equation (8) provides an implicit definition of $\hat{\mathbf{a}}$ which involves the function $Z(\mathbf{a},h)$. Noting that Z scales according to $Z(\mathbf{a},h) = h^d Z(h\mathbf{a},1)$, we can solve explicitly for $\hat{\mathbf{a}}$:

$$\hat{\mathbf{a}} \quad = \quad \frac{1}{h}F^{-1}(\frac{\bar{Y}-\mathbf{y}_0}{h}) \tag{9}$$

where

$$F(\mathbf{a}) = \nabla_{\mathbf{a}} \ln(Z(\mathbf{a}, 1)), \tag{10}$$

does not depend on h.

Finally, replacing $\nabla_{\mathbf{y}} \ln(P_Y(\mathbf{y}))$ in Eq. (4) by $\hat{\mathbf{a}}$ gives the empirical Bayes estimator:

$$\hat{x}_{\text{EB}}(\mathbf{y}_0) = \mathbf{y}_0 + \Lambda \frac{1}{h} F^{-1}(\frac{\bar{Y} - \mathbf{y}_0}{h}). \tag{11}$$

Note that it may be difficult to calculate F^{-1} for a general neighborhood. We'll return to this issue in section 3.2.

3.2 Convergence with proper choice of binwidth

The empirical Bayes estimator of Eq. (11) depends an the choice of binwidth, h, which not only appears explicitly in the expression, but also implicitly affects the choice of which data vectors are included in calculating the local mean, $\bar{Y}$. This binwidth can vary for each data point, but in this paper, we will consider only the simpler case of a constant binwidth for all data. In order for our EB estimator to converge to the BLS estimator, we need $\hat{\mathbf{a}}$, as defined in Eq. (9), to converge to the logarithmic derivative of the true density, and this means the binwidth must shrink to zero as the total amount of data grows. The rate of shrinkage should be set in such a way that the amount of data within each neighborhood goes to infinity. In this subsection, we calculate the rate at which the binwidths should shrink, in order to ensure convergence.

First, we note that the factor of $\frac{1}{h}$ in Eq. (9) implies that, for $\hat{\mathbf{a}}$ to converge at all, it is necessary that

$$F^{-1}\left(\frac{\bar{Y} - \mathbf{y}_0}{h}\right) \rightarrow \mathbf{0}. \tag{12}$$

Since it can be shown that $\mathbf{0}$ is the only zero of F, we need only show that

$$\frac{\bar{Y} - \mathbf{y}_0}{h} \rightarrow \mathbf{0}. \tag{13}$$

Looking again at Eq. (9), and considering a Taylor approximation of F about zero, we see that for for $\hat{\mathbf{a}}$ to converge, it is further necessary that $\frac{\bar{Y} - \mathbf{y}_0}{h^2}$ have some finite limit

$$\frac{\bar{Y} - \mathbf{y}_0}{h^2} \rightarrow \mathbf{v}, \tag{14}$$

for some value $\mathbf{v}$. In this case, a bit of calculation shows that the Taylor approximation of Eq. (9) may be written

$$\hat{\mathbf{a}} \rightarrow V_1 C_1^{-1} \mathbf{v},$$

where C_1/V_1 is the Jacobian matrix of F evaluated at zero, with V_1 denoting the volume of the neighborhood N_1 (N_h, with size $h = 1$), and

$$C_1 = \int_{N_1} (\tilde{\mathbf{y}} - \mathbf{y}_0)(\tilde{\mathbf{y}} - \mathbf{y}_0)^T d\tilde{\mathbf{y}}.$$

Finally, combining the Taylor approximation with the definition of $\hat{\mathbf{a}}$ tells us that our EB estimator converges to the optimal BLS estimator if and only if we can find a choice of binwidth so that

$$\frac{1}{h^2} V_1 C_1^{-1} (\bar{Y} - \mathbf{y}_0) \rightarrow \nabla \ln(P_Y(\mathbf{y}_0)). \tag{15}$$

We now discuss how the binwidth should shrink as the amount of data increases in order to ensure this condition is met. First, we decompose the asymptotic mean squared error into variance and squared bias terms:

$$\begin{aligned} & E\left\{\left(\frac{\bar{Y} - \mathbf{y}_0}{h^2} - \frac{C_1}{V_1} \nabla \ln(P_Y(\mathbf{y}_0))\right)^2\right\} \\ = \; & Var\left\{\frac{\bar{Y} - \mathbf{y}_0}{h^2}\right\} + \left(E\left\{\frac{\bar{Y} - \mathbf{y}_0}{h^2}\right\} - \frac{C_1}{V_1} \nabla \ln(P_Y(\mathbf{y}_0))\right)^2, \end{aligned} \tag{16}$$

We expect (and will show) that the variance term should decrease as h grows (since more data will fall into each bin), whereas the squared bias should increase as the neighborhood grows. This is illustrated in Fig. 2. The trick is to find a way to shrink the binwidths as the amount of data increases, so that the bias goes to zero, but slowly enough to ensure that the variance still goes to zero as well.

For large amounts of data, we expect h to be small, and so we may use small h approximations for the bias and variance. Since the local average, $\bar{Y}$ is calculated using only data which fall in N_h, we may write the expected value as

$$E\{\bar{Y}\} = E\{Y|Y \in N_h\} = \frac{\int_{N_h} \tilde{\mathbf{y}} P_Y(\tilde{\mathbf{y}}) d\tilde{\mathbf{y}}}{\int_{N_h} P_Y(\tilde{\mathbf{y}}) d\tilde{\mathbf{y}}} \tag{17}$$

To get an idea of how this behaves asymptotically, we can take the local Taylor approximation of the density and substitute into Eq. (17). Since the neighborhoods are centered on $\mathbf{y}_0$, the integrals of terms with odd degree will vanish, leaving

$$E\{\bar{Y}\} - \mathbf{y}_0 \approx \frac{C_h \nabla P_Y(\mathbf{y}_0) + O(h^{d+4})}{P_Y(\mathbf{y}_0) V_h + O(h^{d+2})} \tag{18}$$

where V_h denotes the volume of the neighborhood and

$$C_h = \int_{N_h} (\tilde{\mathbf{y}} - \mathbf{y}_0)(\tilde{\mathbf{y}} - \mathbf{y}_0)^T d\tilde{\mathbf{y}}$$

For symmetric neighborhoods, C_h will be a diagonal matrix. If the N_h has the same shape for all h just rescaled by the factor h, a simple change of variables shows that

$$V_h = h^d V_1$$

and

$$C_h = h^{d+2} C_1$$

so that

$$E\{\frac{\bar{Y} - \mathbf{y}_0}{h^2}\} \approx \frac{C_1}{V_1} \nabla \ln(P_Y(\mathbf{y}_0)) + O(h^2) \tag{19}$$

so that the bias term in Eq. (16) will shrink to zero as long as the binwidth shrinks to zero.

Now consider the variance term in Eq. (16). We need to choose binwidths that shrink to zero slowly enough to allow

$$Var\left\{\frac{\bar{Y} - \mathbf{y}_0}{h^2}\right\} \to 0.$$

First note that

$$E\{|Y - \mathbf{y}_0|^2 | Y \in N_h\} = O(h^2) \tag{20}$$

Combining this with Eq. (19) gives

$$\begin{aligned} Var(\frac{\bar{Y} - \mathbf{y}_0}{h^2}) &= \frac{1}{nh^4} Var\,(Y - \mathbf{y}_0 | Y \in N_h) \\ &= O(\frac{1}{nh^2}) + O(\frac{1}{n}) \end{aligned} \tag{21}$$

where n is the number of data points which fall in the bin. Thus, we see that as long as

$$nh^2 \to \infty \tag{22}$$

the variance of our approximation will also tend to zero. If there are N total data points, we can approximate the number falling in N_h as

$$n \approx P_Y(\mathbf{y}_0) N V_1 h^d \tag{23}$$

and, inserting this into Eq. (22) gives

$$Nh^{d+2} \to \infty,\ h \to 0 \tag{24}$$

For example, we might use

$$h \propto N^{-\frac{1}{m+d+2}},\ m > 0 \tag{25}$$

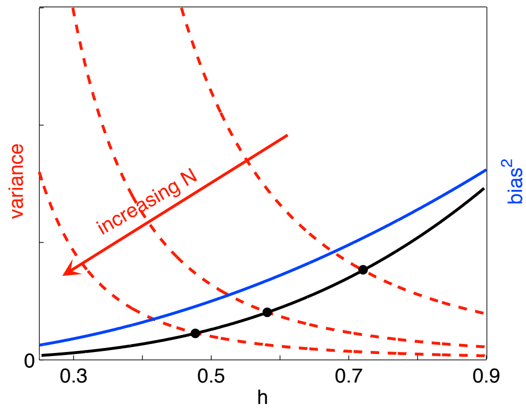


Fig. 2: Bias-variance tradeoff, as a function of binwidth, h. Solid blue line indicates squared bias. Dashed lines indicate variance, for different values of N (number of data points). Solid black line (with black points) shows the variance resulting when h is chosen as a function of N, as indicated in Eq. (25) with $m = 2$. Under these conditions, the variance falls to zero as h shrinks.

The variance resulting from this choice (for $m = 2$) is illustrated in Fig. 2, and can be seen to approach zero as h decreases. With this choice of binwidth, it now follows that the estimator in Eq. (11) converges asymptotically to the optimal BLS solution. Note that a consequence of our asymptotic analysis is that, when the estimator in Eq. (11) converges, we can use the Taylor series approximation of the estimator, so that for the same choice of binwidths

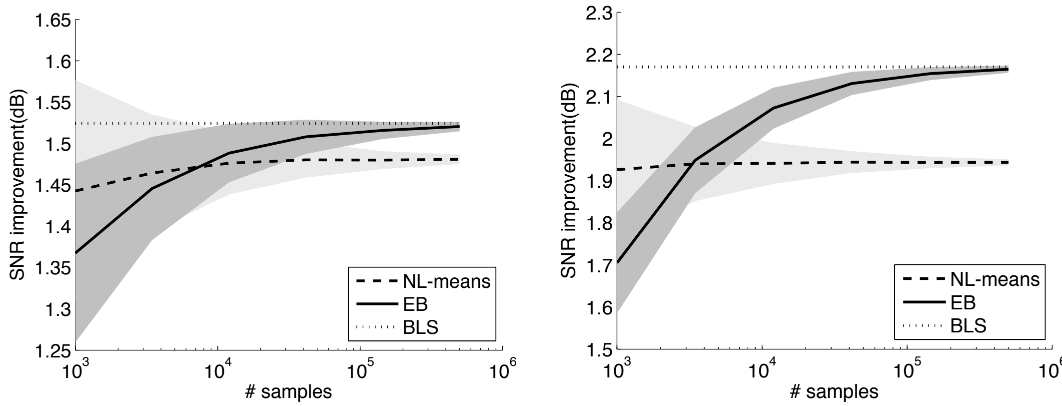
$$\hat{x}_{\text{EB}}(\mathbf{y}_0) \approx \mathbf{y}_0 + \Lambda V_1 C_1^{-1} (\frac{\bar{Y} - \mathbf{y}_0}{h^2}) \tag{26}$$

will also converge asymptotically to the BLS estimator. In the case of spherical neighborhoods we get an estimate of the logarithmic derivative very similar to that derived for the mean-shift method of clustering[3]. Note, though, that the derivation used for that approach uses the derivative of the Epanechnikov kernel, which would only seem to work for spherical neighborhoods. Our method, on the other hand, works for general neighborhoods. Note also that mean-shift is meant to be used as an iterative gradient ascent, moving data towards peaks in the density. For our purposes, however, we have shown that when the logarithmic gradient is scaled by the noise variance, we achieve BLS denoising in one step.

4 Empirical Bayes versus NL means

We have now introduced both the Empirical Bayes method for approximating the BLS estimator, Eq. (11), and the NL-means estimator Eq. (1). For pedagogical reasons, we will compare and contrast the two using low dimensional examples which allow for both easier visualization and easier computation. As discussed in the last section, we expect our EB estimator to converge to the ideal solution as the amount of data goes to infinity. NL-means, on the other hand does not have this asymptotic property in general, since if the binwidths were to shrink the NL-means method would do no denoising. Therefore, we expect the optimal optimal binwidth to asymptote to some finite, nonzero optimal value that depends heavily on both the prior and the amount of noise. For this reason, this method may be able to give us improved performance for low to moderate amounts of data, but will in general be asymptotically biased.

To illustrate the convergence behavior as a function of the amount of data, we have simulated a scalar estimation problem. Signal values are drawn from a generalized Gaussian density, $p(x) \propto \exp(-|x|^{\beta})$, with exponent $\beta = 0.5$. For each N, we draw N samples from this distribution, and add Gaussian white noise to each. The noise was chosen so that the noisy Signal to Noise Ratio (SNR) (10 log of squared error divided by signal variance) was 7.8 dB. We then apply our estimator, with binwidth chosen according to Eq. (25). As a reference, we also show the performance of the ideal BLS estimator (which knows the true prior). We also show the result obtained using NL-means (Eq. (1)), with binwidth chosen to optimize performance (i.e., by comparing the denoising result to the clean signal). For each N, this simulation is performed 1000 times, so as to obtain estimates of the average performance, as well as the variability. The SNR performance is plotted, as a function of N, in Fig. 3. Note that the average performance of NL-means is reasonable for small N



(a) (b)

Fig. 3: Performance of Empirical Bayes method compared with NL-means, and the optimal BLS estimator, for (a) a generalized Gaussian prior; (b) the sum of two generalized Gaussians. Lines (solid, dashed, and dotted) indicate mean performance of each estimator over 1000 simulations. Gray regions (for EB and NL-means) indicate plus/minus one standard deviation. See text.

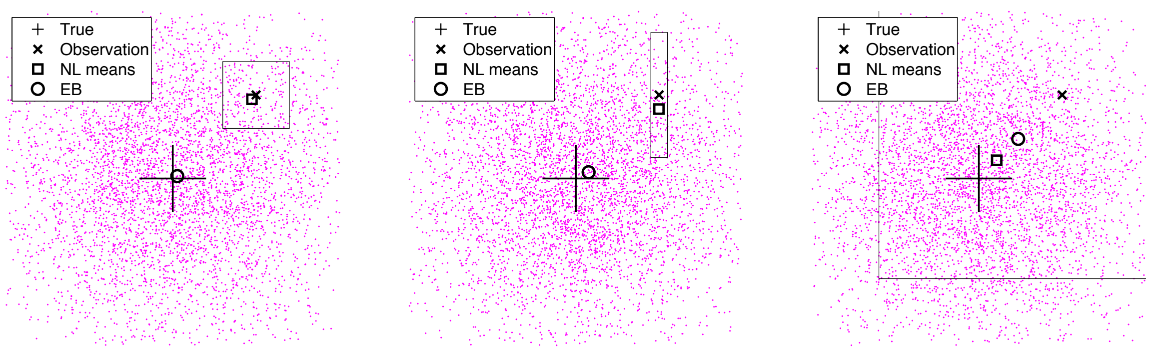


Fig. 4: Behavior of EB method compared with NL-means, for two-dimensional data, with three different neighborhood choices. Prior is a delta at the origin (indicated by cross). Only a subsample of the data points are shown, for visual clarity. Boxes indicate neighborhoods. Note that in the third example, the neighborhood extends beyond the display. See text.

(although the variability is quite high), but saturates as N grows, and thus does not converge to the optimal performance of the BLS estimator. By comparison, our estimator starts out with a lower SNR at small N, but asymptotically converges (in both mean and variance) to the optimal solution. Figure 3 shows another example, for a signal drawn from a mixture of two generalized Gaussian variables, with mean values ± 5. Here, we see a more substantial asymptotic bias for the NL-means estimator.

Figure 4 shows a comparison of our method to NL-means in two dimensions. The signal samples in this example are drawn from a delta function at the origin. We see that the NL-means solution is highly dependent on the size and shape of the neighborhood. It is possible to achieve good results in this particular case of a delta function prior by using a very large neighborhood (3rd panel). But for the smaller neighborhoods, the NL-means estimate is restricted to lie within the neighborhood, and is thus heavily biased, while the EB method is not.

5 Discussion

We've derived a patch-based method for signal denoising by assuming a local exponential density model for the noisy data. We've proven that, despite the fact that we do not include any assumption about the overall form of the signal prior, our estimator converges to the optimal (Bayes) least squares estimator as the amount of data increases, assuming the binsize is reduced appropriately. The estimator is a nonlinear function of the average of similar patches, and thus may be viewed as a generalization of the NL-means and mean-shift algorithms. The main drawback of the method (as with NL-means) is the high computational cost associated with gathering the data patches that are within an h-neighborhood of the patch being denoised. We believe that some of the methods that have been developed for accelerating NL-means [18, 14, 6] (or the other patch-based applications mentioned in the introduction) may be utilized in our estimator as well, and we are currently working on such an implementation so as to examine empirical performance of our method on image denoising.

We envision a number of ways in which our method may be extended. First, the binsizes need not be uniform for all patches, but can be adapted according to the local sample density. Specifically, for each patch that is to be denoised, a rule such as that of Eq. (25) could be used to adjust the binsize to as to achieve the best bias/variance tradeoff (i.e., so as to best approximate the BLS estimator). This extension necessarily increases the computational cost (since the binsize must now be optimized for each patch) but may lead to substantial increases in performance. Second, the NL-means method computes a *weighted* average of patches, where the weighting is a function of both the patch similarity, and the patch proximity within the signal/image (for example, bilateral filtering may be viewed as an NL-means method, with one-pixel patches). We are exploring means by which our method can be generalized to include both of these forms of weighting. Third, EB estimators have been developed for observation processes other than additive Gaussian noise [17, 12, 9, 16], and each of these could be developed into a patch-based estimator (although most of these would not be written in terms of local means). And finally, we believe our local exponential density estimate will prove useful as a substrate for generalizing other patch-based empirical methods that have become popular in computer vision, graphics, and image processing.

References

[1] A. Buades, B. Coll, and J. M. Morel. A review of image denoising algorithms, with a new one. *Multiscale Modeling and Simulation*, 4(2):490–530, July 2005. 1, 2

[2] A. Buades, B. Coll, and J. M. Morel. Nonlocal image and movie denoising. *Intl Journal of Computer Vision*, 76(2):123–139, 2008. 1

[3] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Pat. Anal. Mach. Intell.*, 24:603–619, 2002. 1, 6

[4] A. Criminisi, P. Perez, and K. Toyama. Region filling and object removal by exemplar-based inpainting. *IEEE Trand Image Proc*, 13:1200–1212, 2004. 1

[5] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3D transform-domain collaborative filtering. *IEEE Trans. Image Proc.*, 16(8), August 2007. 1

[6] A. Dauwe, B. Goossens, H. Luong, and W. Philips. A fast non-local image denoising algorithm. In *Proc SPIE Electronic Imaging*, volume 6812, 2008. 8

[7] J. S. De Bonet. Multiresolution sampling procedure for analysis and synthesis of texture images. In *Computer Graphics*. ACM SIGGRAPH, 1997. 1

[8] A. A. Efros and T. K. Leung. Texture synthesis by non-parameteric sampling. In *Proc. Int'l Conference on Computer Vision*, Corfu, 1999. 1

[9] Y. C. Eldar. Generalized SURE for exponential families: Applications to regularization. *IEEE Trans. on Signal Processing*, 57(2):471–481, Feb 2009. 8

[10] N. Jojic, B.Frey, and A. Kannan. Epitomic analysis of appearance and shape. In *ICCV*, 2003. 1

[11] C. R. Loader. Local likelihood density estimation. *Annals of Statistics*, 24(4):1602–1618, 1996. 3

[12] J. S. Maritz and T. Lwin. *Empirical Bayes Methods*. Chapman & Hall, 2nd edition, 1989. 8

[13] K. Miyasawa. An empirical Bayes estimator of the mean of a normal population. *Bull. Inst. Internat. Statist.*, 38:181–188, 1961. 1, 2

[14] J. Orcharda, M. Ebrahimi, and A. Wong. Efcient nonlocal-means denoising using the SVD. In *Proc IEEE Intl Conf on Image Processing*, 2008. 8

[15] K. Popat and R. W. Picard. Cluster-based probability model and its application to image and texture processing. *IEEE Trans Im Proc*, 6(2):268–284, 1997. 1

[16] M. Raphan and E. P. Simoncelli. Least squares estimation without priors or supervision. *Neural Computation*, 2010. In Press. 8

[17] H. Robbins. An empirical Bayes approach to statistics. *Proc. Third Berkley Symposium on Mathematcal Statistics*, 1:157–163, 1956. 8

[18] J. Wang, Y. Guo, Y. Ying, Y. Liu, and Q. Peng. Fast non-local algorithm for image denoising. In *Proc IEEE Intl Conf on Image Processing*, 2006. 8