Classification: Statistics

Nonparametric Regression to the Mean

Hans-Georg Müller[1†], Ian Abramson[2], and Rahman Azari[1]

[1]Department of Statistics, University of California, 1 Shields Ave., Davis, CA 95616, U.S.A.

[2]Department of Mathematics, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093, U.S.A.

ABSTRACT

Available data may reflect a true but unknown random variable of interest plus an additive error which is a nuisance. The problem to predict the unknown random variable arises in many applied situations where measurements are contaminated with errors; it is known as the regression-to-the-mean problem. There exists a well-known solution when both the distributions of the true underlying random variable and of the contaminating errors are normal. This solution is given by the classical regression-to-the-mean formula, which has a data shrinkage interpretation. We discuss the extension of this solution to cases where one or both of these distributions are unknown, and demonstrate that the fully nonparametric case can be solved for the case of small contaminating errors. The resulting nonparametric regression-to-the-mean paradigm can be implemented by a straightforward data sharpening algorithm that is based on local sample means. Asymptotic justifications and practical illustrations are provided.

[†]To whom reprint requests should be addressed, e-mail: mueller@wald.ucdavis.edu

Introduction

The regression-to-the-mean phenomenon was named by Galton (1), who noticed that the height of sons tends to be closer to the population mean than the height of the father. The phenomenon is observed in uncontrolled clinical trials, where subjects with a pathological measurement tend to yield close-to-normal subsequent measurements (2,3) and motivates controlled clinical trials for the evaluation of therapeutic interventions (4,5). Classical regression-to-the-mean has been mainly studied in the context of multivariate normal distributions (6).

In the typical regression-to-the-mean situation one has observations which are contaminated by random errors. The well-known basic result for the situation of a multivariate normal distribution corresponds to shrinkage to the mean and provides the best prediction for a new observation based on past observations and also a method for denoising contaminated observations.

Extensions of the normality-based regression-to-the-mean strategies have been studied by various authors. While the contaminating errors are still assumed to be normal, Das and Mulder (7) derived a regression-to-the-mean formula allowing for an arbitrary distribution of the underlying observations. This result was combined with an Edgeworth approximation of this unkown distribution in (8), and it forms the starting point of our investigation as well, see Eq. 2 below. Regression-to-the-mean for more complex treatment effects has been studied in (9,10).

We propose a new procedure for the case where both the distribution of the true underlying uncontaminated observations (which are to be predicted), as well as the distribution of the contaminating errors are unknown. As we demonstrate, if repeated observations are available, it is possible to obtain consistent predictors under minimal assumptions on the distributions if either the error variance declines or the number of repeated measurements increases asymptotically. We establish asymptotic normality and propose an intuitively appealing and simple implementation based on local sample moments, that is illustrated with a data set consisting of a bivariate sample of repeated blood sugar measurements for pregnant women.

The Regression-to-the-Mean Problem

The general problem can be stated as follows: Given unknown independently and identically distributed random variables X_i, we observe a sample $\{\tilde{X}_1, ..., \tilde{X}_n\}$ of data contaminated with errors δ_i,

$$\tilde{X}_i = X_i + \delta_i, \quad i = 1, ..., n.$$

Here, X_i and δ_i are independent and the contaminating errors δ_i are independently and identically distributed with zero means. The goal is to predict the uncontaminated values X_i from the observed contaminated data $\tilde{X}_i$. The best linear unbiased predictor for X_i is given by the Bayes estimator $E(X_i|\tilde{X}_i)$. Assuming the existence of probability density functions (pdf's) $f_{\tilde{X}}$ for $\tilde{X}$, f_X for X, and f_δ for δ, we find by elementary calculations

$$f_{\tilde{X}}(x) = \int f_\delta(x-y) f_X(y)dy,$$

and

$$f_{\tilde{X},X}(x_1, x_2) = f_\delta(x_1 - x_2) f_X(x_2),$$

where we denote the joint pdf of $(\tilde{X}, X)$ by $f_{\tilde{X},X}$. This leads to the following general form for the regression-to-the-mean function:

$$E(X|\tilde{X} = x_0) = \frac{\int y f_\delta(x_0 - y) f_X(y)dy}{\int f_\delta(x_0 - y) f_X(y)dy}. \tag{1}$$

We show that the difficulty that is caused by the fact that both f_δ and f_X are unknown can be addressed with a nonparametric method. The proposed method produces consistent predictors of the uncontaminated X, whenever the errors δ can be assumed to be shrinking asymptotically, as in situations where an increasing number of repeated measurements become available. In classical regression-to-the-mean a critial assumption is that the contaminating pdf f_δ is Gaussian; even then its variance is typically unknown and must be estimated, requiring the availability of repeated measurements for at least some subjects.

The key argument for the Gaussian case can be found in (7), see also (11) and (12). We reproduce the argument here for the one-dimensional case. Assume $\delta \sim \mathcal{N}(0, \sigma^2)$, $f_{\tilde{X}}(x_0) > 0$ for a given x_0 and denote the standard Gaussian density function by φ. Then, substituting $\frac{1}{\sigma}\varphi\left(\frac{\cdot}{\sigma}\right)$ for f_δ in (1), and using the fact that $x = -\varphi^{(1)}(x)/\varphi(x)$,

$$\begin{aligned} E\left(X|\tilde{X} = x_0\right) &= \left\{\sigma^2 \int \left[\tfrac{\partial}{\partial x}\varphi\left(\tfrac{x-y}{\sigma}\right)|_{x=x_0}\right] f_X(y)dy \right. \\ &\left. + \int x_0 \varphi\left(\tfrac{x_0-y}{\sigma}\right) f_X(y)dy\right\} / \int \varphi\left(\tfrac{x_0-y}{\sigma}\right) f_X(y)dy \qquad (2) \\ &= x_0 + \sigma^2 \tfrac{\partial}{\partial x} f_{\tilde{X}}(x)|_{x=x_0} / f_{\tilde{X}}(x_0). \end{aligned}$$

Under the additional assumption $X \sim \mathcal{N}(\mu, \tau^2)$, we have $\tilde{X} \sim \mathcal{N}(\mu, \sigma^2 + \tau^2)$. Substituting $\frac{1}{(\tau^2+\sigma^2)^{1/2}}\varphi\left(\frac{\cdot}{(\tau^2+\sigma^2)^{1/2}}\right)$ for $f_{\tilde{X}}$ in Eq. 2 then produces the classical regression-to-the-mean formula

$$E\left(X|\tilde{X} = x_0\right) = \frac{\sigma^2}{\sigma^2 + \tau^2}\mu + \frac{\tau^2}{\sigma^2 + \tau^2} x_0. \tag{3}$$

Both Eq. 1 and 2 reveal that regression-to-the-mean corresponds to shrinkage towards the mean; in Eq. 2, this becomes shrinkage to the mode, rather, as $\frac{\partial}{\partial x} f_{\tilde{X}}(x)|_{x=x_0}/f_{\tilde{X}}(x_0) = 0$ at a mode of the density $f_{\tilde{X}}$.

Extending Eq. 2 to the p-dimensional case, one finds analogously

$$E\left(X|\tilde{X} = x_0\right) = x_0 + V\frac{\nabla f_{\tilde{X}}(x)|_{x=x_0}}{f_{\tilde{X}}(x_0)}. \tag{4}$$

Here $V = cov\,(\delta)$ is the $p \times p$ covariance matrix of the contaminating errors δ, which are assumed p-variate normal, $\delta \sim \mathcal{N}_p(0, V)$, and $\nabla f_{\tilde{X}}(x) = \left(\frac{\partial f_{\tilde{X}}}{\partial x_1}(x), ..., \frac{\partial f_{\tilde{X}}}{\partial x_p}(x)\right)^T$ is the gradient of the p-dimensional pdf $f_{\tilde{X}}$.

The Nonparametric Case

The general regression-to-the-mean formula (Eq. 1) is not applicable in practice when neither f_δ nor f_X are contained in a parametric class; indeed it is easily seen that these components are then unidentifiable. The derivation of Eq. 2-4 is tied to the feature that the Gaussian pdf is the unique solution of the differential equation $g^{(1)}(x)/g(x) = -x$.

The following basic assumptions are made.

(A1) The p-dimensional ($p \geq 1$) measurements that are observed for n subjects are generated as follows:

$$\tilde{X}_i = X_i + \delta_i, \quad 1 \leq i \leq n,$$

where the uncontaminated unobservable data X_i are independently and identically distributed (i.i.d.) with pdf f_X, and the measurement errors δ_i are i.i.d. with pdf

$$f_\delta(x) = |V_n|^{-1/2}\psi(V_n^{-1/2}x) \quad \text{for} \quad x \in \Re^p, \tag{5}$$

where ψ is an unknown pdf and V_n is a sequence of covariance matrices $V = V_n = (v_{kl})_{1\leq k,\, l\leq p}$ of full rank, with $\|V_n\| \to 0$, where $\|V_n\| = (\sum_{1\leq k,\, l\leq p} {v_{kl}}^2)^{1/2}$ and $|V|$ denotes the determinant of V. Moreover, X_i and δ_i are independent for all i. For the case $p = 1$, we set $V_n = (\sigma_n) = \sigma$. The $\tilde{X}_i$ are i.i.d. with pdf $f_{\tilde{X}}$.

(A2) At a given point x_0 in the interior of the support of f_X, such that $f_X(x_0) > 0$, the pdf's ψ and f_X are twice continuously differentiable, and ψ satisfies the moment conditions ($p = 1$)

$\int \psi(x)x dx = 0, \qquad \int \psi(x)x^2 dx = \mu_2 = 1,$

$\int \psi(x)x^3 dx = \mu_3, \quad \mu_3 < \infty,$

and for $p > 1$, ψ satisfies

$\int \psi(x) x_j dx = 0, \qquad \int \psi(x) x_j x_k dx = \delta_{jk}$

($\delta_{jk} = 0$ for $j \neq k$, $\delta_{jk} = 1$ for $j = k$), and all third order moments are bounded.

We note that in the case of repeated measurements per subject,

$$\tilde{X}_{ij} = X_i + \delta_{ij}, \quad 1 \leq i \leq n, \quad 1 \leq j \leq m, \tag{6}$$

assuming all δ_{ij} and (X_i, δ_{ij}) are independent, one may work with averages

$$\tilde{X}_{i.} = X_{i.} + \delta_{i.}, \tag{7}$$

where $\delta_{i.} = \frac{1}{m} \sum_{j=1}^{n} \delta_{ij}$, and analogously for $\tilde{X}_i$, X_i. Then, for $p = 1$, Eq. 5 is replaced by

$$f_{\delta.}(x) = \frac{m^{1/2}}{\sigma} \, \psi(\frac{m^{1/2}}{\sigma} x) \tag{8}$$

for fixed m (and analogously for $p > 1$). If the number of repeated measurements is large, we may consider the case $m = m(n) \to \infty$ as $n \to \infty$, where

$$f_{\delta.}(x) = f_{\delta.,n}(x) = \frac{1}{\sigma_{m(n)}} \, \psi_n(\frac{x}{\sigma_{m(n)}}) \tag{9}$$

for $\sigma_{m(n)} = \sigma / m(n)^{1/2}$, with ψ replaced by ψ_n, satisfying the moment properties as in (A2); this case is covered, as long as ψ_n and its first order derivatives are uniformly bounded for all n.

For simplicity, we develop the following argument for the case $p = 1$; the extension to $p > 1$ is straightforward. The central observation under (A1) and (A2) is the following argument: From Eq. 1,

$$\begin{aligned} E\left(X | \tilde{X} = x_0\right) &= \int y \tfrac{1}{\sigma} \psi\left(\tfrac{x_0 - y}{\sigma}\right) f_X(y) dy / \int \tfrac{1}{\sigma} \psi\left(\tfrac{x_0 - y}{\sigma}\right) f_X(y) dy \\ &= \int \left(x_0 - \sigma z\right) \psi(z) f_X(x_0 - \sigma z) dz / \int \psi(z) f_X(x_0 - \sigma z) dz \\ &= x_0 + \sigma \int \{-z\psi(z)\} f_X(x_0 - \sigma z) dz / \int \psi(z) f_X(x_0 - \sigma z) dz, \end{aligned} \tag{10}$$

and for the denominator

$$\int \psi(z) f_X(x_0 - \sigma z) dz = -f_{\tilde{X}}(x_0).$$

Let $\mu_j = \int \psi(x) x^j dx$ for $j \geq 1$. Combining a Taylor expansion with the moment conditions (A2) and observing that, since ψ is a pdf, $\int \psi^{(1)}(z) dz = 0$, $\int \psi^{(1)}(z) z dz = -\int \psi(z) dz = -1$, $\int \psi^{(1)}(z) z^2 dz = -2 \int \psi(z) z dz = 0$, and $\int \psi^{(1)}(z) z^3 dz = -3\mu_2$, we find

$$\int \{\psi^{(1)}(z) - (-z\psi(z))\} f_X(x_0 - \sigma z) dz = \frac{\sigma^2}{2} \mu_3 f_X^{(2)}(x_0) + o(\sigma^2). \tag{11}$$

We note that in the Gaussian case, where $\psi = \varphi$, the term on the l.h.s. of Eq. 11 vanishes, as then $\psi^{(1)}(z) = -z\psi(z)$. In case the contaminating errors have a symmetric pdf, or more generally whenever $\mu_3 = 0$, and the pdfs are three times continuously differentiable, the Taylor expansion can be carried one step further to yield

$$\int \{\psi^{(1)}(z) - (-z\psi(z))\} f_X(x_0 - \sigma z)dz \quad = \tfrac{\sigma^3}{6}[3\mu_2 - \mu_4] f_X^{(3)}(x_0) + o(\sigma^3). \tag{12}$$

Likewise, the difference in Eq. 11, 12 can be made of even smaller order by requiring additional moments to be equal to those of a Gaussian error distribution. Finally,

$$\begin{aligned} \tfrac{\partial}{\partial x} f_{\tilde{X}}(x)|_{x=x_0} \quad &= \int \tfrac{1}{\sigma^2}\psi^{(1)}(\tfrac{x_0-y}{\sigma}) f_X(y)dy \\ &= -\tfrac{1}{\sigma}\int \psi^{(1)}(z) f_X(x_0 - \sigma z)dz. \end{aligned} \tag{13}$$

Combining Eq. 10, 11 and 13,

$$\begin{aligned} E\left(X|\tilde{X} = x_0\right) \quad &= x_0 + \tfrac{\sigma}{f_{\tilde{X}}(x_0)}\int z\psi(z) f_X(x_0 - \sigma z)dz \\ &= x_0 + \sigma^2 \tfrac{f^{(1)}_{\tilde{X}}(x_0)}{f_{\tilde{X}}(x_0)} + \tfrac{1}{2}\sigma^3\mu_3 \tfrac{f^{(2)}_{\tilde{X}}(x_0)}{f_{\tilde{X}}(x_0)} + o(\sigma^3), \end{aligned} \tag{14}$$

and if $\mu_3 = 0$, the leading remainder term is $\sigma^4[3\mu_3 - \mu_4] f^{(3)}_{\tilde{X}}(x_0)/6 f_{\tilde{X}}(x_0)$. Finally, for the multivariate case the same arguments lead to the following extension of Eq. 14,

$$E\left(X|\tilde{X} = x_0\right) = x_0 + V\frac{\nabla f_{\tilde{X}}(x_0)}{f_{\tilde{X}}(x_0)} + O(V^{3/2}). \tag{15}$$

Local Sample Means for Nonparametric Regression-to the Mean

The concept of local moments and local sample moments is related to the data sharpening ideas proposed in (13) and was formulated in (14). The special case of a local sample mean is used implicitly in “mean update” mode finding algorithms (15,16) and provides an attractive device for implementing nonparametric regression-to-the-mean.

The starting point is a random variable Z with twice continuously differentiable density f_Z. Given an arbitrary point $x_0 \in \Re^p, \quad x_0 = (x_{01}, ..., x_{0p})'$, and choosing a sequence of window widths $\gamma = \gamma_n > 0$, define a sequence of local neighborhoods

$$S = S_n = \prod_{j=1}^{p} [x_{0j} - \gamma,\, x_{0j} + \gamma].$$

The local mean at x_0 is defined as $\mu_z = (\mu_{z_1}, ..., \mu_{z_p})'$, with

$$\mu_{z_j} = \lim_{\gamma \to 0} \frac{1}{\gamma^2} E\{(Z - x_0)^{e_j} | Z \in S\}, \quad j = 1, ..., p, \tag{16}$$

where in $e_j = (0, ..., 1, ..., 0)'$ the 1 occurs in the j-th position. According to (14),

$$\mu_{z_j} = \frac{1}{3} D^{e_j} f_Z(x_0)/f_Z(x_0). \tag{17}$$

The empirical counterpart to these local means are the local sample means. Given an i.i.d. sample $(Z_1, ..., Z_n)$ of $\Re^p$-valued random variables with pdf f_Z, where $Z_i = (Z_{i1}, ..., Z_{ip})'$, the local sample mean is $\mu_Z = (\mu_{Z_1}, ..., \mu_{Z_p})'$, where

$$\hat{\mu}_{Z_j} = \frac{1}{\gamma^2} \sum_{i=1}^{n} (Z_{ij} - x_{0j})\, 1_S(Z_i) / \sum_{i=1}^{n} 1_S(Z_i), \quad j = 1, ..., p, \tag{18}$$

and $\gamma = \gamma_n > 0$ is a sequence with $\gamma \to 0$ as $n \to \infty$. This is the sample mean found from the data falling into the local neighborhood $S(x_0)$, standardized by γ^{-2}. By (14), Eq. (3.4) and (3.8),

$$\hat{\mu}_Z = \frac{1}{3} \frac{\nabla f_Z(x_0)}{f_Z(x_0)} + O_p\left(\left(n\gamma^{2+p} \right)^{-1/2} \right), \tag{19}$$

motivating the connection to nonparametric regression-to-the mean as in Eq. 15.

Usually the covariance matrix V of the contaminating errors δ is unknown and can be estimated via the sample covariance matrix

$$\hat{V} = (\frac{1}{n} \sum_{k=1}^{m_i} \frac{1}{m_i} (\tilde{X}_{ikr} - \tilde{X}_{i.r})(\tilde{X}_{iks} - \tilde{X}_{i.s}))_{rs}, \quad 1 \leq r, s \leq p, \tag{20}$$

given a contaminated sample with repeated measurements, $(\tilde{X}_{ik1}, ..., \tilde{X}_{ikp})'$, $1 \leq i \leq n$, $1 \leq k \leq m_i$, and $\tilde{X}_{i.r} = \frac{1}{m_i} \sum_{k=1}^{m_i} X_{ikr}$, where $m_i \geq 2$, $1 \leq r \leq p$.

We note that consistency $\hat{V} = V(1 + o_p(1))$ holds as long as $\sum_{i=1}^{n} m_i \to \infty$, $n \to \infty$. Then the estimate

$$\hat{E}(X | \tilde{X} = x_0) = x_0 + 3\hat{V}\hat{\mu}_{\tilde{X}} \tag{21}$$

satisfies

$$\hat{E}(X | \tilde{X} = x_0) = E(X | \tilde{X} = x_0)(1 + o_p(1)), \tag{22}$$

as long as $\gamma \to 0$, $\sigma \to 0$ and $n\gamma^{2+p} \to \infty$.

The following additional regularity conditions are needed for asymptotic results.

(A3) As $n \to \infty$, $\gamma \to 0$, $n\gamma^{2+p} \to \infty$, and for a $\lambda \geq 0$, $n\gamma^{2+p+4} \to \lambda^2$.

(A4) It holds that $V = V_n = \sigma_n^2 V_0$ for a fixed covariance matrix V_0 with trace$(V_0) = p$ and a sequence $\sigma^2 = \sigma_n^2 \to 0$ as $n \to \infty$. Here, V_0 is the covariance matrix associated with the error pdf ψ defined in (A2).

(A5) As $n \to \infty$, $\quad (n\gamma^{2+p})^{1/2}\sigma \to 0, \quad \sigma/\gamma \to 0.$

We then obtain, using local sample means of Eq. 18 and estimates $\hat{V}$ of Eq. 20, the following main result on asymptotic normality and consistency of the shrinkage estimates in Eq. 21:

Theorem 4.1 Under (A1)-(A5), as $n \to \infty$,

$$(n\gamma^{2+p})^{1/2}\hat{V}^{-1}\{\hat{E}(X|\tilde{X} = x_0) - E(X|\tilde{X} = x_0)\} \to \mathcal{N}(\lambda B, \Sigma) \quad \text{in distribution,} \tag{23}$$

where $B = (\beta_1, ..., \beta_p)'$,

$$\beta_j = \tfrac{1}{10} D^{3e_j} f_X(x_0) - \tfrac{1}{2} D^{e_j} f_X(x_0) \sum_{l=1}^{p} \tfrac{D^{2e_l} f_X(x_0)}{f_X(x_0)} + \tfrac{1}{6} \sum_{l=1, l\neq j}^{p} D^{e_j + 2e_l} f_X(x_0), \tag{24}$$
$$j = 1, ..., p,$$

and

$$\Sigma = (\sigma_{kl}), \quad \sigma_{kl} = (3 \times 2^{-p} f_X(x_0)\delta_{kl}), \; 1 \leq k, l \leq p. \tag{25}$$

In the one-dimensional case ($p = 1$), this simplifies to

$$\beta_1 = \frac{1}{10} f_X^{(3)}(x_0) - \frac{1}{2} \frac{f_X^{(1)}(x_0) f_X^{(2)}(x_0)}{f_X(x_0)}, \quad \Sigma = (\frac{3}{2} f_X(x_0)).$$

Simulation Results

To illustrate the advantage of nonparametric regression-to-the-mean in Eq. 21, we compare it with the Gaussian analog. If $X \sim N(\mu_X, \Sigma)$, $\delta \sim N(0, V)$, $\tilde{X} = X + \delta$, with X, δ independent, the extension of Eq. 3 to the multivariate case is

$$E(X|\tilde{X} = x_0) = \Sigma(\Sigma + V)^{-1} x_0 + V(\Sigma + V)^{-1} \mu_X \tag{26}$$

A total of 300 observations were generated from the $(\frac{1}{2}, \frac{1}{2})$-mixture of two bivariate normal distributions with means $(-1, -1)$ and $(1, 1)$ and common covariance matrix $\frac{1}{8} I$, where I stands for the identity matrix. Samples were then contaminated by adding Gaussian noise with zero mean and covariance matrix $V = \frac{1}{4} I$.

Parametric and nonparametric regression-to-the-mean estimates, assuming that V is known, while μ_X is estimated through the sample mean of the observed $\tilde{X}_i$, are presented in Figure 1 for a typical simulation run. Circles represent the generated uncontaminated data and arrows point from the original data to the contaminated data, which correspond to the tips of the arrows. The graphical results clearly indicate that the nonparametric procedure tracks the original uncontaminated data well, whereas the parametric procedure shrinks the data towards the origin, which is the wrong strategy for these non-normal data.

As a measure of accuracy in recovering the original uncontaminated data, we computed the average sum of squared differences between original uncontaminated data and regression-to-the-mean estimates for the Gaussian method of Eq. 26 and the nonparametric method of Eq. 21 over 500 Monte Carlo samples under the above specifications. The resulting average squared error measures for the Gaussian and nonparametric procedures were 414.44 and 60.19, respectively, indicating an almost seven-fold improvement for nonparametric relative to Gaussian regression-to-the-mean in this example.

Application to Repeated Blood Sugar Measurements

Blood sugar measurements are a common tool in diabetes testing. In a glucose tolerance test, glucose level in blood is measured after a period of fasting (Fasting Glucose measurement) and again one hour after giving the subject a defined dose of glucose (Postprandial Glucose measurement). Pregnant women are prone to develop subclinical or manifest diabetes and establishing the distribution of blood glucose levels under fasting and after a dose of glucose is therefore of interest.

O'Sullivan and Mahan (17) collected data on $n = 52$ pregnant women whose blood glucose levels (fasting and postprandial) were measured during three subsequent pregnancies, thus establishing a series of repeated bivariate measurements with three repetitions ($m = 3, p = 2$); see also (18), p. 211. In a pre-processing step, the data were standardized by subtracting the mean and dividing by the standard deviation for each of the two variables Fasting Glucose (mean 72.9 mg/100ml, st.dev. 6.05) and Postprandial Glucose (mean 107.8 mg/100ml, st.dev. 18.65) separately. Subsequently, 52 bivariate sample means $\tilde{X}_{i.}$ were obtained by averaging over the three repeated measurements for each subject. These data are shown as open circles in Figure 2.

Applying Eq. 19-21 with window width $\gamma = 1.4$ and sample covariance matrix $\hat{V} = (\hat{v}_{ij})$, $\hat{v}_{11} = .531$, $\hat{v}_{22} = .415$, and $\hat{v}_{12} = \hat{v}_{21} = .107$, we obtain the predictions $\hat{E}\left(X_i|\tilde{X}_{i.}\right)$. The arrows in Figure 2 show the displacement from observed to predicted values, the latter corresponding to the tips of the arrows.

Moving from the original observations to the predictions has a data sharpening effect. This can be seen quite clearly from Parzen-Rosenblatt nonparametric kernel density estimates of the bivariate density, comparing the density of the original observations (upper panel) with that of the predictors (lower panel) in Figure 3.

Concluding Remarks

We have generalized the regression-to-the-mean paradigm to a nonparametric situation, where both the nature of the target distribution of given observations as well as that of the contaminating errors are unknown. It is shown that in this fairly general situation regression-to-the-mean corresponds to shrinkage towards the mode of the distribution. We propose a straightforward estimation scheme for the shrinkage factor based on local sample means. Thus a connection emerges between nonparametric regression-to-the-mean with data shrinkage ideas and the mean update algorithm which has been used previously for mode finding and cluster analysis.

Open questions concern choice of smoothing parameters. A plug-in approach could be based on estimating the unknown quantities in the asymptotic distribution provided in Eq. 23-25, and bootstrap methods based on residuals are another option. Procedures for more elaborate designs where nonparametric regression-to-the-mean would be incorporated into more complex models involving comparison of means, analysis of variance, or regression components are also of interest, as is the estimation of the contaminating errors and their distribution from the "residuals" $\hat{E}(X|\tilde{X}) - \tilde{X}$.

Appendix: Proof of Theorem 4.1

We first establish the following result on multivariate asymptotic normality of local sample means, computed from random samples $(X_1, ..., X_n)$ with pdf f_X.

Theorem A.1 For vectors of local sample means $\hat{\mu} = (\hat{\mu}_1, ..., \hat{\mu}_p)'$ of Eq. 18 and $\mu = (\mu_1, ..., \mu_p)'$, $\mu_j = D^{e_j} f_X(x_0)/3f_X(x_0)$ of Eq. 17, it holds under (A1)-(A3) that

$$(n\gamma^{2+p})^{1/2}\{\hat{\mu} - \mu\} \rightarrow \mathcal{N}_p(\lambda\tilde{B}, \tilde{\Sigma}) \quad \text{in distribution,} \tag{A.1}$$

where $\tilde{B} = B/3$ (see Eq. 24) and $\tilde{\Sigma} = \Sigma/9$ (see Eq. 25).

Proof. Extending an argument of (14), p. 105, consider random variables

$$U_j = (X - x_0)^{e_j} 1_S(X) - \gamma^2 \mu_j(x_0) 1_S(X).$$

By third order Taylor expansion of f_X, $EU_j = \gamma^4 |S| \tilde{\beta}_j + o(\gamma^4|S|)$, $\quad EU_j^2 = \gamma^2 |S| f_X(x_0)/3 + o(\gamma^2|S|)$ and $EU_jU_k = O(\gamma^4|S|)$ for $j \neq k$. Defining random variables

$$W_{inj} = \frac{1}{n|S|\sigma^2}(X_i - x_0)^{e_j} 1_S(X_i) - \gamma^2 \mu_j(x_0) 1_S(X_i)$$

and using fixed constants $\alpha_1, ..., \alpha_p$, we find that

$$E\{(n\gamma^{p+2})^{1/2} \sum_{i=1}^{n} \sum_{j=1}^{p} \alpha_i W_{inj}\} \to \lambda^{1/2} \sum_{j=1}^{p} \alpha_j \tilde{\beta}_j \quad (n \to \infty)$$

and

$$\mathrm{var}\{(n\gamma^{p+2})^{1/2} \sum_{i=1}^{n} \sum_{j=1}^{p} \alpha_i W_{inj}\} \to 2^{-p} \sum_{j=1}^{p} \alpha_j^2 f_X(x_0)/3 \quad (n \to \infty).$$

Applying the Cramér-Wold device and Slutsky's theorem completes the proof.

Proof of Theorem 4.1. Observing Eq. 15, 19, 21, (A4), (A5) and the consistency of $\hat{V}$,

$$(n\gamma^{p+2})^{1/2}\hat{V}^{-1}(\hat{E}(X|\tilde{X} = x_0) - E(X|\tilde{X} = x_0)) = 3\hat{V}^{-1}V[(n\gamma^{p+2})^{1/2}(V^{-1}\hat{V}\hat{\mu}_{\tilde{X}} - \mu_{\tilde{X}})] + O(\sigma^4(n\gamma^{p+2})^{1/2}),$$

is seen to have the same limiting distribution as $3(n\gamma^{p+2})^{1/2}(\hat{\mu}_{\tilde{X}} - \mu_{\tilde{X}})$. Therefore, Theorem 4.1 is a direct consequence of Theorem A.1 once we establish the following two results:

$$(n\gamma^{p+2})^{1/2})(\mu_{\tilde{X}_j} - \mu_{X_j}) \xrightarrow{p} 0, \quad j = 1, ..., p, \tag{A.2}$$

and

$$(n\gamma^{p+2})^{1/2}\hat{\mu}_{\tilde{X}_j} - \hat{\mu}_{X_j}) \xrightarrow{p} 0, \quad j = 1, ..., p. \tag{A.3}$$

The moment conditions for ψ (see (A2)) in the multivariate case are, with constants β_α and ζ_α,

$$\int \psi(y) y^\alpha dy \quad = 1, \alpha = 0; \qquad = 0, |\alpha| = 1; \qquad = \beta_\alpha, |\alpha| = 2;$$

and this leads to (see (19), ch. 6, and (20))

$$\int D^{e_j}\psi(y) y^\alpha dy \quad = 0, \text{ for } \alpha = 0, |\alpha| = 1 \text{ and } \alpha \neq e_j, \, |\alpha| = 2;$$

$$= -1, \; \alpha = e_j; \qquad = \zeta_\alpha, \; |\alpha| = 3.$$

Using these moment conditions in second order Taylor expansions,

$$f_{\tilde{X}}(x_0) - f_X(x_0) = \sigma^2 \sum_{|\alpha|=2} \beta_\alpha D^\alpha f_X(x_0) + o(\sigma^2), \tag{A.4}$$

$$D^{e_j} f_{\tilde{X}}(x_0) - D^{e_j} f_X(x_0) = \sigma^2 \sum_{|\alpha|=3} \zeta_\alpha D^\alpha f_X(x_0) + o(\sigma^2), \tag{A.5}$$

whence (A.2) follows by (A5).

We next discuss the denominators of $\hat{\mu}_X$ and $\hat{\mu}_{\tilde{X}}$. Abbreviating $\rho_n = (n\gamma^{p+2})^{1/2}$, we find, based on the kernel density estimator with uniform kernel and window S, denoting the indicator function by $I(\cdot)$,

$$\rho_n(\hat{f}_{\tilde{X}}(x_0) - \hat{f}_X(x_0)) = \frac{\rho_n}{n|S|} \sum_{i=1}^{n} \{I(\tilde{X}_i \in S, X_i \notin S) - I(\tilde{X}_i \notin S, X_i \in S)\},$$

and since by (A.4), $EI(\tilde{X}_i \in S) - EI(X_i \in S) = O(|S|\sigma^2)$, we arrive at

$$E[\rho_n(\hat{f}_{\tilde{X}}(x_0) - \hat{f}_X(x_0))] = O(\rho_n \sigma^2). \tag{A.6}$$

Note that due to (A5),

$$\int\limits_{u\in S} f_\delta(u - x_0)du = \int\limits_{u\in S} \tfrac{1}{\sigma^p}\psi((\sigma^2 V_0)^{-1/2}(u - x_0))du \longrightarrow 1, \quad \text{as} \quad n \to \infty,$$

which implies

$$\begin{aligned} E(I(\tilde{X}_i \in S) - & \quad I(X_i \in S))^2 \\ = & \int\limits_{u\in S} f_{\tilde{X}}(u)du + \int\limits_{u\in S} f_X(u)du - 2\int\limits_{u\in S}\int\limits_{v\in S} f_\delta(u-v) f_X(v)dudv \\ = & \left[|S|\{f_{\tilde{X}}(x_0) + f_X(x_0)\} - 2\int\limits_{u\in S} f_\delta(u-x_0)\{|S| f_X(x_0)\} \right](1 + o(1)) \\ = & \; O(|S|\sigma^2), \end{aligned}$$

again using (A.4). We conclude $\text{var}(\rho_n(\hat{f}_{\tilde{X}}(x_0) - \hat{f}_X(x_0))) = O(\rho_n^2 \sigma^2/n|S|)$, whence, with (A.6),

$$\rho_n \left\{ \frac{1}{n|S|} \sum_{i=1}^{n} \left[I(X_i \in S) - I(\tilde{X}_i \in S) \right] \right\} \xrightarrow{p} 0 \tag{A.7}$$

Regarding the numerator, the terms to consider are

$$T_n = \frac{\rho_n}{n|S|} \left\{ \sum_{i=1}^{n} (\tilde{X}_{ij} - x_{0j}) I(X_i \in S) - \sum_{i=1}^{n} (\tilde{X}_{ij} - x_{0j}) I(\tilde{X}_i \in S) \right\}, \tag{A.8}$$

and the terms that include x_{0j} are handled in the same way as the denominator, using (A.7). Since $\tilde{X}_{ij} = X_{ij} + \delta_{ij}$, it therefore remains to consider

$$\frac{\rho_n}{n|S|} \left\{ \sum_{i=1}^{n} \tilde{X}_{ij} [I(X_i \in S) - I(\tilde{X}_i \in S)] - \sum_{i=1}^{n} \delta_{ij} I(X_i \in S) \right\} = \text{I} + \text{II}.$$

The same argument as for the denominator and additional Cauchy-Schwarz bounds lead to $E\text{I} \to 0$, $E\text{I}^2 \to 0$ and therefore $\text{I} \xrightarrow{p} 0$. For II, note that

$$E\,\text{II} = -\frac{\rho_n}{n|S|} \sum_{i=1}^{n} E\delta_{ij} EI(X_{ij} \in S) = 0,$$

as X_{ij} and δ_{ij} are independent. Furthermore, $E(\delta_{ij} I(X_{ij} \in S))^2 = O(\sigma^2|S|)$ leads to $\text{var(II)} = O(\gamma^2\sigma^2) = o(1)$, according to (A5). Therefore, $T_n \xrightarrow{p} 0$, and (A.3) follows, concluding the proof.

Acknowledgments

This research was supported in part by NSF grants DMS-9971602, DMS-0204869 and 0079430. We are grateful for the helpful and detailed comments of two reviewers.

1. Galton, F. (1886) *J. Anthropological Institute* **15**, 246-263.
2. James, K.E. (1973) *Biometrics* **29**, 121-130.
3. Pitts, S.R. & Adams, R.P. (1998) *Annals of Emergency Medicine* **31**, 214-218.
4. Bland, J.M. & Altman, D.G. (1994) *British Medical Journal* **309**, 780.
5. Yudken, P.L. & Stratton, I.M. (1996) *Lancet* **347**, 241-243.
6. Davis, C.E. (1976) *Am. J. Epidemiology* **104**, 1163-1190.
7. Das, P. & Mulder, P.G.H. (1983) *Statistica Neerlandica* **37**, 493-497.
8. Beath, K.J. & Dobson, A.J. (1991) *Biometrika* **78**, 431-435.
9. Chen, S. & Cox, C. (1992) *Biometrics* **48**, 593-598.
10. Chen, S., Cox, C. & Cui, L. (1998) *Biometrics* **54**, 939-947.
11. Abramson, I. (1988) *J. Am. Statistical Association* **83**, 1073-1077.
12. Haff, L.R. (1991) *Annals of Statistics* **19**, 1163-1190.
13. Choi, E. & Hall, P. (1999) *Biometrika* **86**, 941-947.
14. Müller, H.G. & Yan, X. (2001) *J. Multivariate Analysis*, **76**, 90-109.
15. Funkunaga, K. & Hostetler, L.D. (1975) *IEEE Trans. Information Theory* **21**, 32-40.
16. Fwu, C., Tapia, R.A. & Thompson, J.R. (1981) In: Proceedings of the 26th conference of the Design of Experiments in Army Research Development and Testing, pp. 309-326.
17. O'Sullivan, J.B. & Mahan, C.M. (1966) *American Journal of Clinical Nutrition* **19**, 345-351.
18. Andrews, D.F. & Herzberg, A.M. (1985) *Data.* Springer, New York
19. Müller, H.G. (1988) *Nonparametric Regression Analysis for Longitudinal Data.* Springer, New York.

20. Müller, H.G. & Stadtmüller, U. (1999) *J. Royal Statistical Society B* **61**, 439-458.

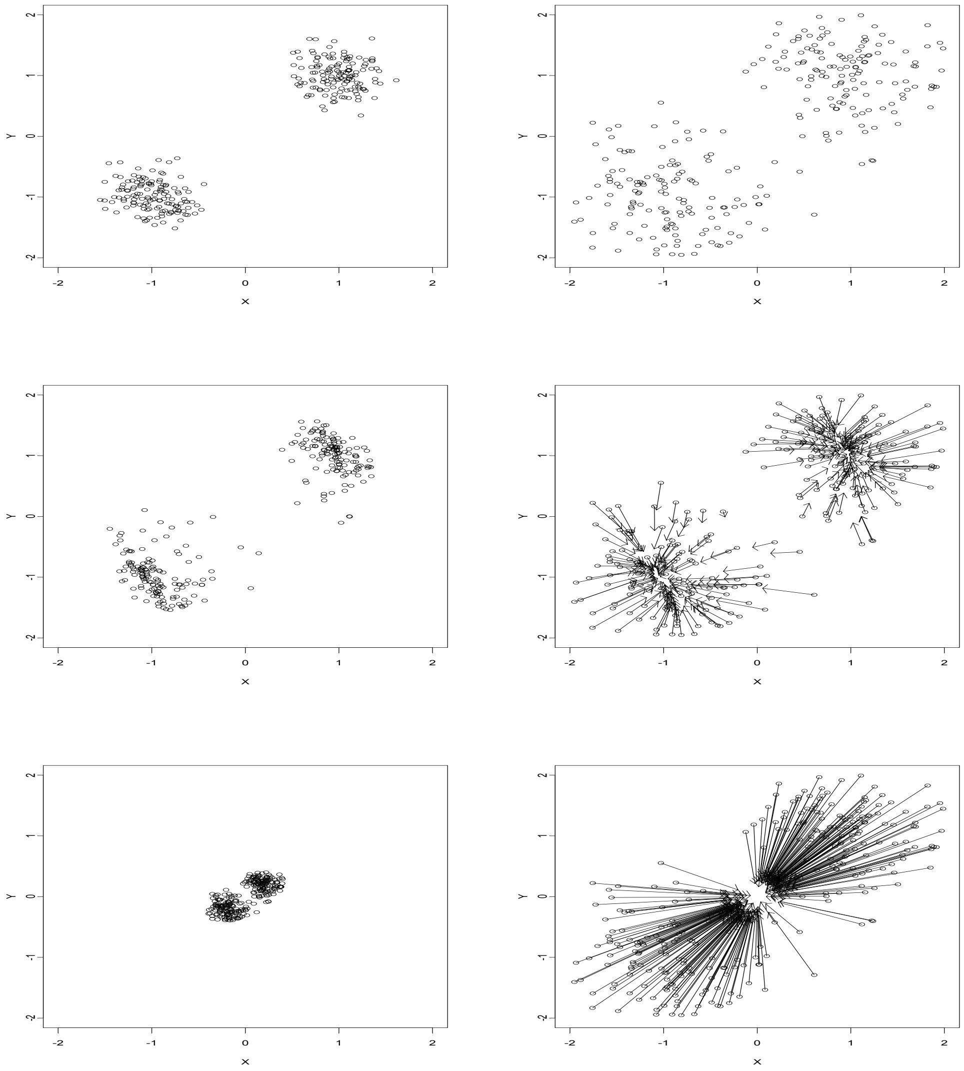

Figure 1: Sample of size 300 from a mixture of bivariate normal distributions (top row left), contaminated sample (top row right), nonparametic-regression-to-the-mean using Eq. 21 (middle row left), arrows pointing from contaminated to predicted observations (middle row right) and Gaussian estimates using Eq. 26 (bottom row left) with corresponding arrows (bottom row right). Only data falling into the window $[-2, 2] \times [-2, 2]$ are shown.

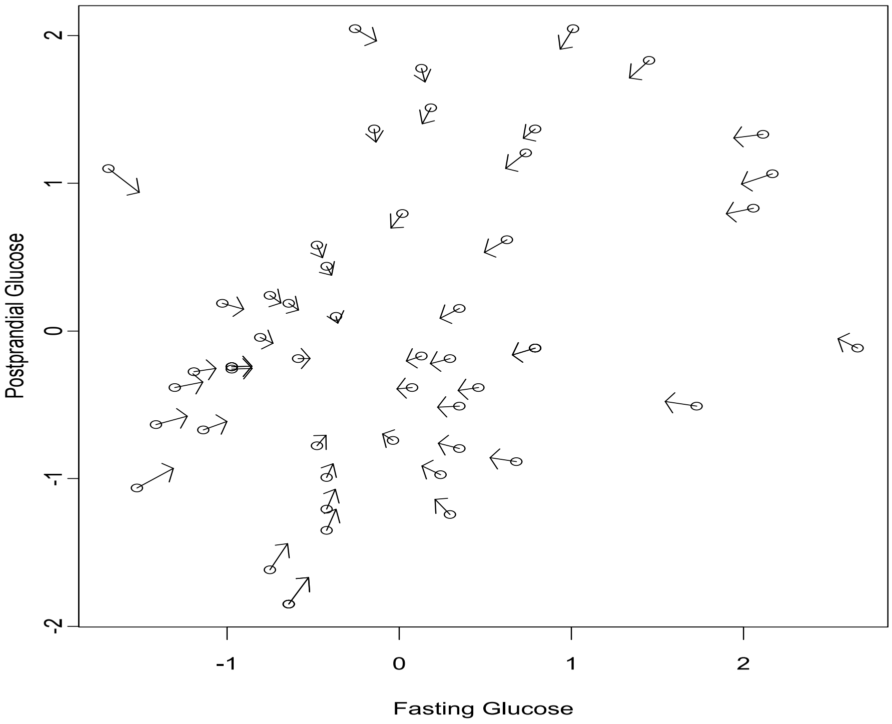

Figure 2: Bivariate nonparametric regression-to-the-mean (Eq. 21) for glucose measurements for 52 women, with repeated measurements over three pregnancies. Circles are observed sample means obtained from the three repetitions of the standardized values of (Fasting Glucose, Postprandial Glucose). Arrows point from observed to predicted values.

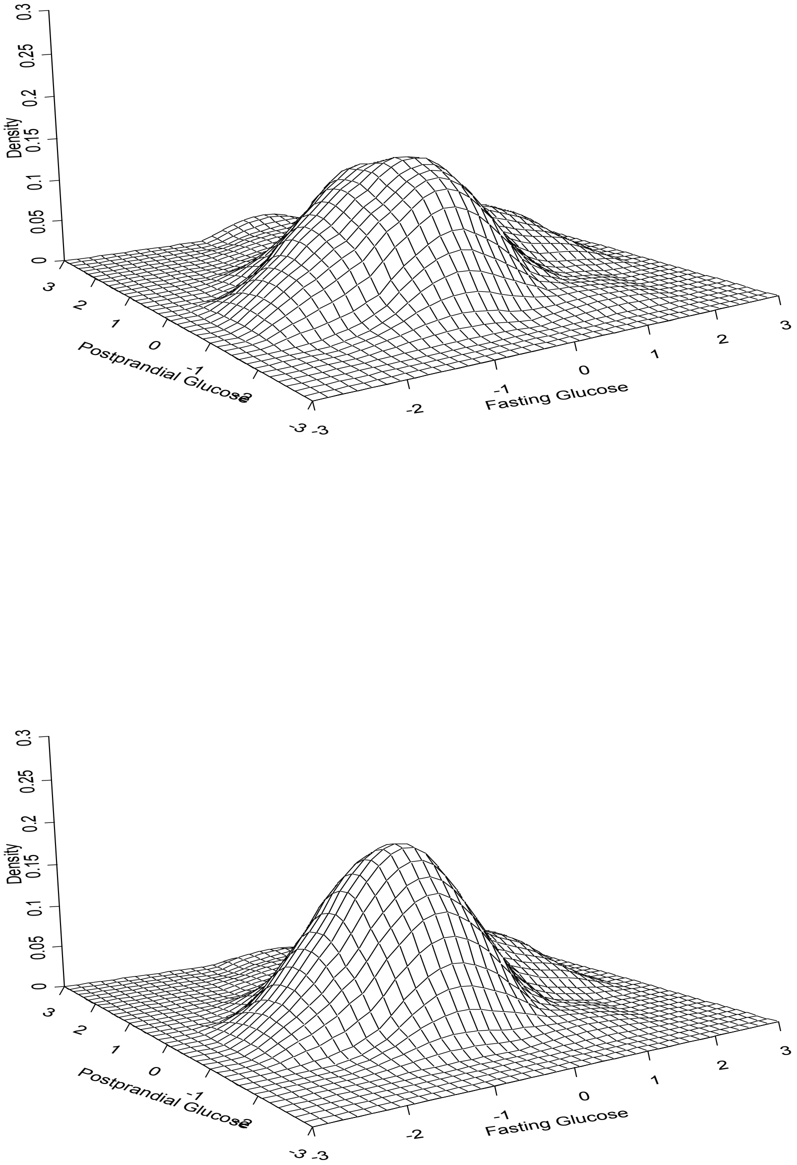

Figure 3: Bivariate kernel density estimates of the joint density of (Fasting Glucose, Postprandial Glucose) data, with bandwidth $(1, 1)$. Upper panel: Density estimate based on original observations. Lower panel: Density estimate based on predicted values after applying nonparametric-regression-to-the-mean (Eq. 21).