Zero Biasing in One and Higher Dimensions, and Applications [*‡]

Larry Goldstein[§] and Gesine Reinert[¶]

June 30, 2004

Abstract

Given any mean zero, finite variance σ^2 random variable W, there exists a unique distribution on a variable W^* such that $EWf(W) = \sigma^2 Ef'(W^*)$ for all absolutely continuous functions f for which these expectations exist. This distributional 'zero bias' transformation of W to W^*, of which the normal is the unique fixed point, was introduced in [9] to obtain bounds in normal approximations. After providing some background on the zero bias transformation in one dimension, we extend its definition to higher dimension and illustrate with an application to the normal approximation of sums of vectors obtained by simple random sampling.

1 Introduction

In [14], Stein proved the following fundamental characterization of the univariate normal distribution: $Z \sim \mathcal{N}(0, \sigma^2)$ if and only if

$$EZf(Z) = \sigma^2 Ef'(Z) \tag{1}$$

for all absolutely continuous f with $E|f'(Z)| < \infty$. Normal approximations through the use of identity (1), first provided by Stein [14], have since been obtained by other authors using variations in abundance (see e.g. [15], [5], [13] and references therein). In [9] the authors introduced and studied the zero bias transformation in one dimension; further development is continued in [7] and [8]. Here the authors extend the study of zero biasing to $\mathbf{R}^p$, and illustrate with an application.

If W is a mean zero variance σ^2 variable then generally $EWf(W)$ does not equal $\sigma^2 Ef'(W)$, the normal being the unique distribution satisfying the Stein characterization (1). Asking that identity (1) be satisfied by some transformation of the W distribution leads to the following definition.

[*]AMS 2000 subject classifications. Primary 60F05, 62D05

[†]Key words and phrases: distributional transformation, zero bias, size bias, sampling, multivariate normal approximation.

[‡]This work was partially completed while the authors were visiting the Institute for Mathematical Sciences, National University of Singapore, in 2003. The visit was supported by the Institute.

[§]Department of Mathematics, University of Southern California, Los Angeles CA 90089-2532, USA

[¶]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK, supported in part by EPSRC grant no. GR/R52183/01

Definition 1.1 *For a mean zero, finite variance σ^2 random variable W, we say that W^* has the W-zero biased distribution if*

$$EWf(W) = \sigma^2 Ef'(W^*), \tag{2}$$

for all absolutely continuous functions f for which the expectations above exist.

In [9] the distribution W^* was shown to exist for all mean zero, finite variance W. In particular, W^* is always absolutely continuous, and one can verify that

$$p^*(w) = \sigma^{-2} E[W\mathbf{1}(W > w)] \tag{3}$$

is a density function for a distribution which satisfies (2). Though Definition 1.1 is stated in terms of random variables, it actually defines a transformation on a class of distributions, and we will use the language interchangeably.

The normal distribution being the unique fixed point of the zero bias transformation, one can guess that when the transformation maps a variable W to a W^* which is close by, then W itself is close to being a fixed point, and must therefore be close to normal. Equation (5) indicates one way in which this intuition can be quantified.

Let W be a mean zero, variance 1 random variable, Z a standard normal and $Nh = Eh(Z)$ for a given test function h. Based on the characterization (1), Stein's method obtains information on $Eh(W) - Nh$, the distance from W to the normal on a given test function h, by solving the differential equation

$$f'(w) - wf(w) = h(w) - Nh \tag{4}$$

for f and considering the expectation of the left hand side. For instance, when W and W^* are jointly defined, from (4) and Definition 1.1 we obtain

$$|Eh(W) - Nh| = |E\left[f'(W) - Wf(W)\right]| = |E\left[f'(W) - f'(W^*)\right]|.$$

By [15] for h absolutely continuous we have $||f''|| \leq 2||h'||$ and hence, with $||\cdot||$ the supremum norm,

$$|Eh(W) - Nh| \leq 2||h'||E|W - W^*|; \tag{5}$$

in particular, when W and W^* are close in the sense that $E|W - W^*|$ can be made small, $Eh(W)$ will be close to Nh.

Proposition 2.1 in Section 2 below, first shown in [9], gives the following zero bias explanation of the Central Limit Theorem, that is, why a good normal approximation can be expected when W is a sum of many comparable mean zero independent variables with finite variances: such a sum can be zero biased by choosing a single summand with probability proportional to its variance, and replacing it by one from that summand's zero biased distribution. Hence, if the terms in the sum W are roughly comparable, W and the W^* so constructed differ in only one like summand out of many, and W and W^* are close.

Definition 1.1, and Proposition 2.1, for zero biasing are parallel to the well known definition, and key property, of the W size biased distribution W^s, which exists for any nonnegative W with finite mean μ, and is characterized by

$$EWf(W) = \mu Ef(W^s) \tag{6}$$

for all functions f for which these expectations exist. In particular, (2) is (6) with variance replacing mean, and f' replacing f. Moreover, the prescription for size biasing a sum of non-negative independent variables is nearly the same as the one give in Proposition 2.1; in particular, choose a summand with probability proportional to its mean and replace it by one from that summand's size biased distribution. Due to the similarity between zero and size biasing, the transformation in Definition 1.1 was so coined as it offered a parallel of size biasing for mean zero variables, hence, zero biasing. For use of size biasing for normal approximation, see [11] and [8]; for families of distributional transformations of which both zero and size biasing are special cases, see [10].

In [9], after introducing the zero bias transformation, the authors apply it to show that smooth function rates of n^{-1} obtain for simple random sampling under certain higher order moment conditions. In [7] zero biasing is used to provide bounds to the normal for hierarchical sequences, and in [8] for normal approximation in combinatorial central limit theorems with random permutations having distribution constant over cycle type.

Here the authors generalize Definition 1.1 to $\mathbf{R}^p$ based on the Stein characterization that $\mathbf{Z} \in \mathbf{R}^p$ is a multivariate normal $\mathcal{N}(0, \Sigma)$ if and only if

$$E\sum_{i=1}^{p} Z_i f_i(\mathbf{Z}) = E\sum_{i,j=1}^{p} \sigma_{ij} f_{ij}(\mathbf{Z}) \quad \text{all smooth } f : \mathbf{R}^p \to \mathbf{R},$$

where f_i and f_{ij} denote the first and second partial derivatives of f, respectively.

Definition 1.2 *Let Γ be an arbitrary index set and let $\mathbf{X} = \{X_\gamma : \gamma \in \Gamma\}$ be a collection of mean zero random variables with covariances $EX_\alpha X_\beta = \sigma_{\alpha\beta}$. For pairs α, β with $\sigma_{\alpha\beta} \neq 0$, we say that the collection of variables $\mathbf{X}^{\alpha,\beta} = \{X_\gamma^{\alpha,\beta} : \gamma \in \Gamma\}$ has the $\mathbf{X}$-zero biased distribution in coordinates α, β if for all finite $I \subset \Gamma$,*

$$E\sum_{\beta\in I} X_\beta f_\beta(\mathbf{X}) = E\sum_{\alpha\in I}\sum_{\beta\in I} \sigma_{\alpha\beta} f_{\alpha,\beta}(\mathbf{X}^{\alpha,\beta}) \tag{7}$$

for all twice differentiable functions f for which the above expectations exist.

For $\gamma \in \Gamma$ and a smooth function g, setting $f(\mathbf{X}) = g(X_\gamma)$ we have $f_\beta(\mathbf{X}) = g'(X_\gamma)\mathbf{1}(\gamma = \beta)$, $f_{\alpha,\beta}(\mathbf{X}) = g''(X_\gamma)\mathbf{1}(\gamma = \alpha = \beta)$ and (7) reduces to

$$EX_\gamma g'(X_\gamma) = \sigma_\gamma^2 E g''(X_\gamma^{\gamma,\gamma}),$$

showing that $X_\gamma^{\gamma,\gamma}$ has the X_γ zero bias distribution. More generally, if the collection $\mathbf{X}^{\alpha,\beta}$ satisfies Definition 1.2 for variables indexed over Γ, then the restriction of the same variables indexed over a subset of Γ satisfies Definition 1.2 over that subset. In particular, when Γ is finite we need only verify the definition for $I = \Gamma$.

Modifying the Stein normal characterization to yield an identity such as (2) which applies to a large class of distributions is an approach also taken in [4]. Rather than changing the distribution of X to satisfy the right hand side of (2), in [4] the existence of a function w is postulated such that $EXf(X) = \sigma^2 Ew(X)f'(X)$. Based on an idea in [2], the use of the w function is extended in [3] to a multivariate case for independent mean zero variables $X_1, \ldots, X_p$ with finite variances σ_i^2 under the condition that the given variables have density $p_i(x), i = 1, \ldots, p$. In this case, one can define for each i the w_i-function via $\sigma_i^2 w_i(x_i)p_i(x_i) =$

$E\{X_i \mathbf{1}(X_i > x_i)\}$; we recognize $w_i(x)$ as the Radon-Nikodyn derivative of the zero bias density (3) of X_i^* with respect to the density of X_i. Following this approach, in [3], the covariance identity $\mathrm{Cov}(\sum_{i=1}^p X_i, g(\mathbf{X})) = \sum_{i=1}^p \sigma_i^2 E[w_i(\mathbf{X}) g_i(\mathbf{X})]$ is obtained for smooth functions $g : \mathbf{R}^p \to \mathbf{R}$. In [12] the relationship between the w-function approach and the zero-bias coupling is exploited. Whereas our interests here lie in normal approximation and the associated couplings, the emphasis in [3] and in [12] is to derive variance lower bounds.

In Section 2 we list some of the known properties of the zero bias transformation in one dimension. Proposition 2.2, whose short zero bias proof provides a test function type bound in the Central Limit Theorem for a sum of independent random variables in $\mathbf{R}$, is generalized to higher dimension by Theorem 3.1 in Section 3. These two results are based on the principle that the proximity of a variate $\mathbf{W}$ to the normal can be measured by the proximity of $\mathbf{W}$ to its zero biased version. Proposition 2.1, proving how a zero bias coupling can be generated for the sum of independent variables by replacing one summand, is generalized to higher dimension by Theorem 3.2. For this extension, we introduce the notion of *constant sign covariance*. The type of specification given by Proposition 2.3 to generate zero bias couplings in the non-independent case in $\mathbf{R}$ is used in Theorem 4.1 in Section 4 to construct vectors with the zero bias distribution of a vector whose dependent coordinates satisfy certain exchangeability conditions. Finally, putting the bounds and the construction together, in Theorem 4.2 we obtain bounds in a multivariate central limit theorem for sums of vectors obtained by simple random sampling.

2 Zero Biasing in One Dimension

The key property of the zero bias transformation which illuminates its use for normal approximation is the following fact from [9], that the zero biased distribution W^* for a sum W of independent variables can be constructed by choosing a single variable with probability proportional to its variance, and replacing that variable with one from its own zero bias distribution.

Proposition 2.1 *Let $X_1, \ldots, X_n$ be independent, mean zero random variables with finite variances $\sigma_1^2, \ldots, \sigma_n^2$ and set*

$$W = \sum_{i=1}^n X_i.$$

Then with I a random index independent of $X_1, \ldots, X_n$ having distribution

$$P(I = i) = \frac{\sigma_i^2}{\sum_{j=1}^n \sigma_j^2},$$

and X_i^* having the X_i size bias distribution, independent of $X_j, j \neq i$ and of I, the variable

$$W^* = W - X_I + X_I^*$$

has the W-zero biased distribution.

Proof: For any smooth f with compact support,

$$EWf(W) = \sum_{i=1}^n EX_i f(W)$$

$$= \sum_{i=1}^{n} EX_i f(X_i + \sum_{t \neq i} X_t) = \sum_{i=1}^{n} \sigma_i^2 Ef'(X_i^* + \sum_{t \neq i} X_t) \quad (8)$$

$$= \sigma^2 \sum_{i=1}^{n} \frac{\sigma_i^2}{\sigma^2} Ef'(W - X_i + X_i^*) = \sigma^2 Ef'(W - X_I + X_I^*) = \sigma^2 Ef'(W^*),$$

where we have used independence of X_i and $X_t, t \neq i$ in (8). Now extend from smooth f to absolutely continuous f where expectations in (2) exist by standard limiting arguments. ■

Proposition 2.1 leads directly to the following simple proof of the Central Limit Theorem with a smooth function bound on the rate under a third moment assumption.

Proposition 2.2 *Let $X_1, \ldots, X_n$ be independent mean zero variables with variances $\sigma_1^2, \ldots, \sigma_n^2$ and finite third moments, and let $W = (X_1 + \ldots + X_n)/\sigma$ where $\sigma^2 = \sigma_1^2 + \ldots + \sigma_n^2$. Then for all absolutely continuous test functions h,*

$$|Eh(W) - Nh| \leq \frac{2||h'||}{\sigma^3} \sum_{i=1}^{n} E\left(|X_i|\sigma_i^2 + \frac{1}{2}|X_i|^3\right), \quad (9)$$

so in particular, when the variables have variance 1 and common third absolute moment,

$$|Eh(W) - Nh| \leq \frac{3||h'||E|X_1|^3}{\sqrt{n}}. \quad (10)$$

Proof: Proposition 2.1 and the fact that $(cW)^* = cW^*$ for any constant $c \neq 0$ gives

$$E|W - W^*| = \frac{1}{\sigma} E|X_I - X_I^*| = \frac{1}{\sigma^3} \sum_{i=1}^{n} E|X_i - X_i^*|\sigma_i^2 \quad \text{since} \quad P(I = i) = \sigma_i^2/\sigma^2. \quad (11)$$

Letting $f(x) = x^2 \text{sgn}(x)$ and $f'(x) = 2|x|$ in equation (2) we derive $\sigma_i^2 E|X_i^*| = E|X_i|^3/2$ and so

$$E|X_i - X_i^*| \leq E\left(|X_i| + |X_i^*|\right) \leq E\left(|X_i| + \frac{1}{2\sigma_i^2}|X_i|^3\right). \quad (12)$$

Using (12) in (11) and applying (5) proves (9). By Hölder's inequality when $EX_i^2 = 1$, we have $E|X_i| \leq 1 \leq E|X_i|^3$ from which follows (10). ■

The Wasserstein distance inequality

$$d(W, Z) \leq 2d(W, W^*) \quad (13)$$

is almost immediate from (5) and was proved in [7]. The distance d is defined by

$$d(X, Y) = \inf_{h \in \mathcal{L}} |Eh(X) - Eh(Y)| \quad \text{where} \quad \mathcal{L} = \{h : |h(x) - h(y)| \leq |x - y|\},$$

and also has the dual representation

$$d(X, Y) = \inf E|X - Y|$$

where the infimum is over all joint distributions on (X, Y) with given marginals; the infimum is achieved for variables on $\mathbf{R}$. Choosing W, W^* to achieve $d(W, W^*)$ on the right hand side of (5) and then taking supremum over $h \in \mathcal{L}$ gives (13).

For constants $\alpha_1, \ldots, \alpha_n$, when

$$W = \sum_{i=1}^{n} \frac{\alpha_i}{\lambda} X_i \tag{14}$$

with $X_1, \ldots, X_n$ i.i.d. mean zero variance 1 and $\sum_i \alpha_i^2 = \lambda^2$, we have from Proposition 2.1, with $P(I = i) = \alpha_i^2/\lambda^2$, that

$$|W^* - W| = \frac{\alpha_I}{\lambda}|X_I^* - X_I| = \sum_{i=1}^{n} \frac{\alpha_i}{\lambda}|X_i^* - X_i|\mathbf{1}(I = i).$$

Taking $(X_i^*, X_i) =_d (X^*, X)$ to achieve the infimum $d(X, X^*)$ gives an instance of (W^*, W) on a joint space, and so

$$d(W, W^*) \le E|W^* - W| = \sum_{i=1}^{n} \frac{|\alpha_i|^3}{\lambda^3} E|X_i^* - X_i| = \varphi d(X, X^*) \tag{15}$$

where

$$\varphi = \frac{\sum_{i=1}^{n} |\alpha_i|^3}{(\sum_{i=1}^{n} \alpha_i^2)^{3/2}}.$$

It is easily checked that $\varphi < 1$ unless all but one of the α_i values are zero, and so (15) makes precise an intuition that in some sense a sum of n independent variables, even for $n = 2$, has a distribution closer to the normal than that of its summands; iteration yields a contraction mapping type proof of the Central Limit Theorem. The equally weighted case $\alpha_i = 1$ gives $\lambda = \sqrt{n}$ and $\varphi = 1/\sqrt{n}$, and now applying (13) we obtain

$$d(W, Z) \le 2d(W, W^*) \le 2n^{-1/2} d(X, X^*) \to 0 \quad \text{as } n \to \infty,$$

a simple proof of the Central Limit Theorem in the independent case, with a bound in Wasserstein distance. Bounds in normal convergence when iterating non-linear 'averaging' functions was obtained in [7] using linear approximations of the form (14).

As first shown in [9], for W in general, a sum of dependent variables for example, a construction of W and W^* on a joint space can be accomplished using Proposition 2.3 and the exchangeable pair of Stein [15].

Proposition 2.3 *Let W', W'' be an exchangeable pair with distribution $dF(w', w'')$ satisfying the linearity condition*

$$E(W''|W') = (1 - \lambda)W' \quad \textit{for some } \lambda \in (0, 1). \tag{16}$$

If $\hat{W}', \hat{W}''$ has distribution

$$d\hat{F}(\hat{w}', \hat{w}'') = \frac{(\hat{W}' - \hat{W}'')^2 dF(\hat{w}', \hat{w}'')}{E(W' - W'')^2},$$

and $U \sim \mathcal{U}[0, 1]$ is independent of $\hat{W}', \hat{W}''$, then the variable

$$W^* = U\hat{W}' + (1 - U)\hat{W}'' \quad \textit{has the } W' \textit{ zero bias distribution.}$$

Proof: It suffices to verify (2) for smooth f with compact support. Letting $\sigma^2 = \mathrm{Var}(W')$, we have

$$\begin{aligned} \sigma^2 E f'(U\hat{W}' + (1-U)\hat{W}'') &= \sigma^2 E\left(\frac{f(\hat{W}') - f(\hat{W}'')}{\hat{W}' - \hat{W}''}\right) \\ &= \sigma^2 E\left(\frac{(f(W') - f(W''))(W' - W'')}{E(W' - W'')^2}\right). \end{aligned}$$

Now using (16) to see that $EW''f(W') = (1-\lambda)EW'f(W')$ and $E(W' - W'')^2 = 2\lambda\sigma^2$, expanding the expression above yields

$$\sigma^2 E\left(\frac{W'f(W') - W''f(W') - W'f(W'') + W''f(W'')}{E(W' - W'')^2}\right) = \frac{2\lambda\sigma^2 EW'f(W')}{E(W' - W'')^2} = EW'f(W').$$

■

This construction is the basis for deriving the Kolmogorov supremum norm type bounds to the normal distribution function for the combinatorial central limit theorem application in [8].

3 Multivariate Zero Biasing

Using Definition 1.2 we now generalize some of the properties of univariate zero biasing given in Section 2 to higher dimension. In particular, Theorem 3.1 is a multidimensional version of the bound (5), Theorem 3.2 a generalization of Proposition 2.1, and Theorem 4.1 in Section 4 an extension of Proposition 2.3.

Given a vector $\mathbf{a}$ in R^p, let $||\mathbf{a}|| = \max_{1\le i\le p} |a_i|$. Given a $p \times p$ matrix $A = (a_{ij})$ we set $||A|| = \max_{1\le i,j\le p} |a_{ij}|$, and more generally for any array, $||\cdot||$ will denote its maximal absolute value. For an array of functions arbitrarily indexed, say $A(\mathbf{w}) = \{a_\alpha(\mathbf{w})\}_{\alpha\in\mathcal{A}}$, $||A|| = \sup_{\mathbf{w}} \sup_\alpha |a_\alpha(\mathbf{w})|$. For a smooth function $h : \mathbf{R}^p \to \mathbf{R}$ we let ∇h or Dh denote the vector of first partial derivatives of h, D^2h the Hessian matrix of second order partial derivatives and in general $D^k h$ is the array of k^{th} order partial derivatives of h.

Theorem 3.1 *Let $\mathbf{W}$ be a mean zero random vector in $\mathbf{R}^p$ with positive definite covariance matrix $\Sigma = (s_{\alpha,\beta})$. Suppose that for all $s_{\alpha,\beta} \neq 0$, we have vectors $\mathbf{W}^{\alpha,\beta}$ with the $\mathbf{W}$ zero biased distribution in coordinates α, β. Then for a three times differentiable test function $h : \mathbf{R}^p \to \mathbf{R}$, $\mathbf{Z}$ a standard normal vector in $\mathbf{R}^p$ and $Nh = Eh(\mathbf{Z})$,*

$$|Eh(\Sigma^{-1/2}\mathbf{W}) - Nh| \le \frac{p^4}{3}||\Sigma^{-1/2}||^3||D^3h|| \sum_{\alpha,\beta=1}^{p} |s_{\alpha\beta}|\, E||\mathbf{W} - \mathbf{W}^{\alpha,\beta}||.$$

Proof: Following [6], and [1] (see also [11]), a solution f to the differential equation

$$\mathrm{tr}\Sigma D^2 f(\mathbf{W}) - \mathbf{W}\cdot\nabla f(\mathbf{W}) = h(\Sigma^{-1/2}\mathbf{W}) - Nh$$

exists and satisfies

$$\left|\frac{\partial^k}{\prod_{j=1}^k \partial w_{\alpha_j}} f(\mathbf{w})\right| \le \frac{p^k}{k}||\Sigma^{-1/2}||^k||D^k h|| \quad k = 1,2,3. \tag{17}$$

Applying Definition 1.2,

$$\begin{aligned} |Eh(\Sigma^{-1/2}\mathbf{W}) - Nh| &= |E\left(\mathrm{tr}\Sigma D^2 f(\mathbf{W}) - \mathbf{W}\cdot\nabla f(\mathbf{W})\right)| \\ &= |E\sum_{\alpha,\beta=1}^{p} s_{\alpha\beta}(f_{\alpha,\beta}(\mathbf{W}) - f_{\alpha,\beta}(\mathbf{W}^{\alpha,\beta}))| \\ &= |E\sum_{\alpha,\beta=1}^{p} s_{\alpha\beta}\nabla f_{\alpha,\beta}(\xi_{\alpha,\beta})\cdot(\mathbf{W} - \mathbf{W}^{\alpha,\beta})|, \end{aligned}$$

where $\xi_{\alpha,\beta}$ is on the line segment between $\mathbf{W}$ and $\mathbf{W}^{\alpha,\beta}$. The proof is completed by applying the triangle inequality and the bound (17). ∎

Given a collection of vectors $\mathbf{X}_i = (X_{1,i}, \ldots, X_{p,i}) \in \mathbf{R}^p, i = 1, \ldots, n$ let

$$\mathbf{W} = \sum_{i=1}^{n} \mathbf{X}_i, \quad \text{and set} \quad \sigma_{jl,i} = \mathrm{Cov}(X_{j,i}, X_{l,i}).$$

In order the generalize Proposition 2.1 to higher dimension, we make the following

Definition 3.1 *The collection of vectors $\mathbf{X}_i = (X_{1,i}, \ldots, X_{p,i}) \in \mathbf{R}^p, i = 1, \ldots, n$ has constant sign covariance if the sign of $\sigma_{jl,i}$ does not depend on i.*

Example 3.1 *If each vector in the collection $(\mathbf{X}_i)_{i=1,\ldots,n}$ has uncorrelated components then trivially the collection has constant sign covariance.*

Example 3.2 *If the components of each vector $\mathbf{X}_i \in \mathbf{R}^p, i = 1, \ldots, n$ is obtained by a simple random sample of size p from a population with characteristics $\mathcal{A}_i$, $|\mathcal{A}_i| = N_i > p$, satisfying*

$$\sum_{a\in\mathcal{A}_i} a = 0,$$

then

$$\sigma_{j,i}^2 = \frac{1}{N_i}\sum_{a\in\mathcal{A}_i} a^2 \quad \textit{and} \quad \sigma_{jl,i} = \frac{-1}{N_i(N_i-1)}\sum_{a\in\mathcal{A}_i} a^2 \quad \textit{for } j \neq l,$$

and hence the collection $(\mathbf{X}_i)_{i=1,\ldots,n}$ has constant sign covariance.

When a collection of vectors $(\mathbf{X}_i)_{i=1,\ldots,n}$ has constant sign covariance then for every $j, l = 1, \ldots, p$ we may define a probability distribution on indices $i = 1, \ldots, n$ by

$$P(I^{jl} = i) = \frac{\sigma_{jl,i}}{s_{jl}}, \quad \text{with} \quad s_{jl} = \sum_{i=1}^{n} \sigma_{jl,i}, \tag{18}$$

and in addition, if for $i \neq i'$ we have $\mathrm{Cov}(\mathbf{X}_i, \mathbf{X}_{i'}) = 0$, then

$$s_{jl} = \mathrm{Cov}(W_j, W_l) \quad \text{where} \quad W_j = \sum_{i=1}^{n} X_{j,i}.$$

Theorem 3.2 *Let $(\mathbf{X}_i)_{i=1,\ldots,n}$ be a collection of mean zero constant sign covariance vectors, independent over i, and for each $i = 1, \ldots, n$ and $j, l = 1, \ldots, p$ suppose that $\mathbf{X}_i^{jl}$ has the $\mathbf{X}_i$ zero biased distribution in coordinates j, l and is independent of $\mathbf{X}_j$ for $j \neq i$. Then with I^{jl} having distribution (18) and independent of all other variables*

$$\mathbf{W}^{jl} = \mathbf{W} - \mathbf{X}_{I^{jl}} + \mathbf{X}_{I^{jl}}^{jl} = \mathbf{X}_{I^{jl}}^{jl} + \sum_{t \neq I^{jl}} \mathbf{X}_t$$

has the $\mathbf{W}$ zero biased distribution in coordinates jl. In particular, when $\mathbf{W}$ has positive definite covariance matrix $\Sigma = (s_{jl})$, then for any three times differentiable test function $h : \mathbf{R}^p \to \mathbf{R}$, and $Nh = Eh(\mathbf{Z})$ for a standard normal vector $\mathbf{Z} \in \mathbf{R}^p$,

$$|Eh(\Sigma^{-1/2}\mathbf{W}) - Nh| \leq \frac{p^4}{3} ||\Sigma^{-1/2}||^3 ||D^3 h|| \sum_{j,l=1}^{p} |s_{jl}| \, E||\mathbf{X}_{I^{jl}} - \mathbf{X}_{I^{jl}}^{jl}||.$$

Proof: It suffices to consider a smooth function f with compact support, for which

$$\begin{aligned}
& E\sum_{j=1}^{p} W_j f_j(\mathbf{W}) = E\sum_{j=1}^{p}\sum_{i=1}^{n} X_{j,i} f_j(\mathbf{W}) = E\sum_{i=1}^{n}\sum_{j=1}^{p} X_{j,i} f_j(\mathbf{X}_i + \sum_{t \neq i} \mathbf{X}_t) \\
= & \sum_{i=1}^{n} E\sum_{j=1}^{p} X_{j,i} f_j(\mathbf{X}_i + \sum_{t \neq i} \mathbf{X}_t) = \sum_{i=1}^{n} E\left\{ \sum_{j=1}^{p}\sum_{l=1}^{p} \sigma_{jl,i} f_{jl}(\mathbf{X}_i^{jl} + \sum_{t \neq i} \mathbf{X}_t) \right\} \\
= & E\sum_{j=1}^{p}\sum_{l=1}^{p} s_{jl} \sum_{i=1}^{n} \frac{\sigma_{jl,i}}{s_{jl}} f_{jl}(\mathbf{X}_i^{jl} + \sum_{t \neq i} \mathbf{X}_t) \\
= & E\sum_{j=1}^{p}\sum_{l=1}^{p} s_{jl} \sum_{i=1}^{n} P(I^{jl} = i) f_{jl}(\mathbf{X}_i^{jl} + \sum_{t \neq i} \mathbf{X}_t) = E\sum_{j=1}^{p}\sum_{l=1}^{p} s_{jl} f_{jl}(\mathbf{W}^{jl}).
\end{aligned}$$

The second assertion follows directly from Theorem 3.1, completing the proof. ∎

4 Construction using Exchangeable Pairs

Theorem 4.1 demonstrates how to construct the zero biased vectors $\mathbf{X}^{jl} \in \mathbf{R}^p, j, l = 1, \ldots, p$ for a mean zero vector $\mathbf{X} \in \mathbf{R}^p$ with components satisfying conditions similar to those imposed to prove Proposition 2.1 and Theorem 2.1 in [9]; we note (X_j', X_j'') in (19) is an embedding of an univariate exchangeable pair in a multidimensional vector.

Theorem 4.1 *Let $\mathbf{X}' = (X_1', X_2', \ldots, X_p')$ be a vector of mean zero finite variance variables with $Var(X_j') = \sigma^2 > 0$ and $EX_j'X_l' = \kappa$ for $j \neq l$, $\sigma^2 > \kappa$. Assume that for every j there exists X_j'' such that*

$$\mathbf{X}_j^{'\,''} = (X_1', \ldots, X_j', X_j'', \ldots, X_p') \;\; = \;\; {}_d(X_1', \ldots, X_j'', X_j', \ldots, X_p'); \tag{19}$$

let $dF_j^{'\,''}(\mathbf{x})$ denote the distribution of $\mathbf{X}_j^{'\,''}$. By (19) for all $j \neq l$,

$$E(X_j'' X_l') = \kappa; \tag{20}$$

assume that (20) holds for $j = l$ as well. In addition, assume that for some λ,

$$E(X_j''|\mathbf{X}') = \lambda Y' \quad \textit{where} \quad Y' = \sum_{m=1}^{p} X_m'. \tag{21}$$

By the foregoing, the positive quantity

$$v^2 = E(X_j' - X_j'')^2 = 2(\sigma^2 - \kappa) \tag{22}$$

does not depend on j, and we may consider the $p+1$ vector

$$\hat{\mathbf{X}}_j^{'''} = (\hat{X}_1', \ldots, \hat{X}_j', \hat{X}_j'', \ldots, \hat{X}_p')$$

with distribution

$$d\hat{F}_j^{'''}(\hat{\mathbf{x}}) = \frac{(\hat{X}_j' - \hat{X}_j'')^2}{v^2} dF_j^{'''}(\hat{X}_1, \ldots, \hat{X}_j', \hat{X}_j'', \ldots \hat{X}_p), \tag{23}$$

and $\hat{\mathbf{X}}_j'$ and $\hat{\mathbf{X}}_j''$ the p-vectors obtained by removing $\hat{X}_j''$ and $\hat{X}_j'$ from $\hat{\mathbf{X}}_j^{'''}$, respectively.

Then, with U_j a uniform $\mathcal{U}(0,1)$ variable independent of $\{\hat{\mathbf{X}}_j', \hat{\mathbf{X}}_j''\}$, and

$$\mathbf{X}^{jj} = U_j \hat{\mathbf{X}}_j' + (1 - U_j)\hat{\mathbf{X}}_j'', \tag{24}$$

with V_{jl} Bernoulli random variables $P(V_{jl} = 1) = P(V_{jl} = 0) = 1/2$ independent of $\{\mathbf{X}^{jj}, \mathbf{X}^{ll}\}$, the collection

$$\mathbf{X}^{jl} \;=\; V_{jl}\mathbf{X}^{jj} + (1 - V_{jl})\mathbf{X}^{ll}, \quad j, l = 1, \ldots, p$$

has the $\mathbf{X}$ zero bias distribution in coordinates j, l.

Since $0 \le \mathrm{Var}(X_j'' + Y') = (p+1)(\sigma^2 + p\kappa)$, it follows $p\kappa \ge -\sigma^2$ and hence $\mathrm{Var}(Y') = p(\sigma^2 + (p-1)\kappa) \ge -p\kappa$. Therefore $\kappa < 0$ implies $\mathrm{Var}(Y') > 0$; clearly if $\kappa \ge 0$ then $\mathrm{Var}(Y') > 0$, directly. Hence the denominator of the ratio in (25), relating λ and κ, is always strictly positive, and these two values share the same sign.

Remark 4.1 *When $|X_j'| \le M$ for $j = 1, \ldots, p$ then $||\mathbf{X}'|| \le M$ giving in turn $||\mathbf{X}_j^{'''}|| \le M$, and then by the construction in Theorem 4.1, that $||\hat{\mathbf{X}}_j^{'''}||, ||\hat{\mathbf{X}}_j'||, ||\hat{\mathbf{X}}_j''||$ and finally $||\mathbf{X}^{jl}||$ are all bounded by M. That the support of a variate and that of its zero bias distribution do not coincide in general, however, is easy to see even in $\mathbf{R}$; if X has the discrete uniform distribution on the two values -1 and 1, then the X zero-bias distribution is the continuous uniform $\mathcal{U}[-1, 1]$.*

Proof of Theorem 4.1 Multiplying the conditional expectation in (21) by Y' and taking expectation we obtain

$$EX_j''Y' = \lambda Var Y' = \lambda p(\sigma^2 + (p-1)\kappa).$$

But from (20) we also have that $EX_j''Y' = \kappa p$, and equating it follows that

$$\lambda = \frac{\kappa}{\sigma^2 + (p-1)\kappa}. \tag{25}$$

Using (23) and (24), with f any smooth function and, suppressing j, letting $\mathbf{X}'$ and $\mathbf{X}''$ denote the vector obtained by removing X_j'' and X_j' from $\mathbf{X}_j^{'''}$ respectively,

$$\begin{aligned} Ef_j(\mathbf{X}^{jj}) &= Ef_j(U_j\hat{\mathbf{X}}_j' + (1-U_j)\hat{\mathbf{X}}_j'') \\ &= E\left(\frac{f(\hat{\mathbf{X}}_j') - f(\hat{\mathbf{X}}_j'')}{\hat{X}_j' - \hat{X}_j''}\right) \\ &= \frac{1}{v^2}E(X_j' - X_j'')(f(\mathbf{X}') - f(\mathbf{X}'')) \\ &= \frac{2}{v^2}E\left(X_j'f(\mathbf{X}') - X_j''f(\mathbf{X}')\right) \\ &= \frac{2}{v^2}E\left(X_j'f(\mathbf{X}') - \lambda Y'f(\mathbf{X}')\right) \quad \text{by (21).} \end{aligned}$$

Now, taking expectation over V_{jl} and noting that $f_{jl} = f_{lj}$, we have

$$\begin{aligned} & E\sum_{j=1}^p\sum_{l=1}^p \sigma_{jl}f_{jl}(\mathbf{X}^{jl}) = \sigma^2\sum_{j=1}^p Ef_{jj}(\mathbf{X}^{jj}) + \frac{\kappa}{2}\sum_{j=1}^p\sum_{l\neq j}\{Ef_{lj}(\mathbf{X}^{jj}) + Ef_{jl}(\mathbf{X}^{ll})\} \\ = \;& \frac{2\sigma^2}{v^2}\sum_{j=1}^p E\left(X_j'f_j(\mathbf{X}') - \lambda Y'f_j(\mathbf{X}')\right) + \frac{2\kappa}{v^2}\sum_{j=1}^p\sum_{l\neq j} E\left(X_j'f_l(\mathbf{X}') - \lambda Y'f_l(\mathbf{X}')\right) \\ = \;& \frac{2\sigma^2}{v^2}\sum_{j=1}^p E\left(X_j'f_j(\mathbf{X}') - \lambda Y'f_j(\mathbf{X}')\right) \\ & + \frac{2\kappa}{v^2}\sum_{j=1}^p E\left\{\sum_{l=1}^p X_j'f_l(\mathbf{X}') - X_j'f_j(\mathbf{X}') - \lambda\sum_{l=1}^p Y'f_l(\mathbf{X}') + \lambda Y'f_j(\mathbf{X}')\right\} \\ = \;& \frac{2(\sigma^2-\kappa)}{v^2}\sum_{j=1}^p E\left(X_j'f_j(\mathbf{X}') - \lambda Y'f_j(\mathbf{X}')\right) \\ & + \frac{2\kappa}{v^2}\sum_{j=1}^p E\left\{\sum_{l=1}^p X_j'f_l(\mathbf{X}') - \lambda\sum_{l=1}^p Y'f_l(\mathbf{X}')\right\}. \end{aligned}$$

Employing (22) for the first term, and letting $\mathrm{div}f(x) = \sum_l f_l(x)$, this expression can be written

$$\begin{aligned} & \sum_{j=1}^p EX_j'f_j(\mathbf{X}') - \lambda EY'\mathrm{div}f(\mathbf{X}') + \frac{2\kappa}{v^2}E\left(Y'\mathrm{div}f(\mathbf{X}') - \lambda pY'\mathrm{div}f(\mathbf{X}')\right) \\ = \;& \sum_{j=1}^p EX_j'f_j(\mathbf{X}') + \left(\frac{-\lambda v^2 + 2\kappa(1-\lambda p)}{v^2}\right)EY'\mathrm{div}f(\mathbf{X}') \\ = \;& E\sum_{j=1}^p X_j'f_j(\mathbf{X}'), \end{aligned}$$

since by (22) and (25),

$$-\lambda v^2 + 2\kappa(1-\lambda p) = 2\left(-\lambda(\sigma^2-\kappa) + \kappa(1-\lambda p)\right) = 2(\kappa - \lambda(\sigma^2 + (p-1)\kappa)) = 0. \quad \blacksquare$$

Example 4.1 *Independent Variables. It is not difficult to see directly from Definition 1.2 that a vector* $\mathbf{X}$ *of independent random variables can be zero biased in coordinate* j *by replacing* X_j *by a variable* X_j^* *having the* X_j *zero biased distribution, independent of* $X_l, l \neq j$*; this construction is equivalent to the special case of Theorem 4.1 when taking* X_j'' *in (19) to be an independent copy of* X_j'*. In particular, in this case*

$$||\mathbf{X} - \mathbf{X}^{jj}|| = |X_j - X_j^*|.$$

From this observation and calculations parallel to those in Proposition 2.2 we obtain the following corollary of Theorem 3.2.

Corollary 4.1 *Let* $(\mathbf{X}_i)_{i=1\ldots,n}$ *be an independent collection of mean zero random vectors in* $\mathbf{R}^p$ *whose coordinates* $X_{j,i}, j = 1, \ldots, p$ *are independent with variance* $\sigma_{jj,i}$ *and finite third absolute moments. Then for*

$$\mathbf{W} = \sum_{i=1}^n \mathbf{X}_i \quad \textit{and} \quad s_{jj} = \sum_{i=1}^n \sigma_{jj,i},$$

and any three times differentiable test function h*,*

$$|Eh(\Sigma^{-1/2}\mathbf{W}) - Nh| \leq \frac{p^4}{3}(\min_{1\leq j\leq p} s_{jj})^{-3/2}||D^3h|| \sum_{j=1}^p \sum_{i=1}^n E\left(\sigma_{jj,i}^2|X_{j,i}| + \frac{1}{2}|X_{j,i}|^3\right),$$

and when $X_{j,i}, i = 1, \ldots, n$ *are identically distributed with variance 1 for all* $j = 1, \ldots, p$*,*

$$|Eh(\Sigma^{-1/2}\mathbf{W}) - Nh| \leq \frac{p^4}{2\sqrt{n}}||D^3h|| \sum_{j=1}^p E|X_{j,1}|^3.$$

In our next example we consider vectors having non-independent components.

Example 4.2 *Simple Random Sampling. Let* $\mathbf{X}' \in \mathbf{R}^p$ *be a vector whose values are obtained by taking a simple random sample of size* p *from a population having characteristics* $\mathcal{A}, |\mathcal{A}| = N > p$*, with* $\sum_{a\in\mathcal{A}} a = 0$*. Taking one additional observation* X_j'' *we form an enlarged vector that satisfies (19). In the notation of Theorem 4.1,*

$$E(X_j'X_l'') = \frac{-1}{N(N-1)} \sum_{a\in\mathcal{A}} a^2, \tag{26}$$

and (20) is satisfied with κ *the value (26), and*

$$E(X_j''|\mathbf{X}') = \frac{-1}{N-p} Y',$$

so (21) is satisfied with $\lambda = -1/(N-p)$*. Hence the hypotheses of Theorem 4.1 hold, and the construction so given can be used to produce a collection with the* $\mathbf{X}'$ *zero biased distribution.*

We now arrive at

Theorem 4.2 *Let $(\mathbf{X}_i)_{i=1\ldots,n}$ be an independent collection of mean zero random vectors in $\mathbf{R}^p$ obtained by simple random sampling as in Example 3.2. Suppose $|a| \leq M$ for all $a \in \mathcal{A}_i, i = 1, \ldots, n$, and let*

$$\mathbf{W} = \sum_{i=1}^{n} \mathbf{X}_i \quad \textit{and} \quad s_{jl} = Cov(W_j, W_l).$$

Then for a three times differentiable test function h,

$$|Eh(\Sigma^{-1/2}\mathbf{W}) - Nh| \leq \frac{2}{3} M p^4 ||\Sigma^{-1/2}||^3 ||D^3 h|| \sum_{j,l=1}^{p} |s_{jl}|.$$

Proof: As shown in Example 3.2, the collection of vectors have constant sign covariance, and hence Theorem 3.2 applies. Using the construction given Theorem 4.1 and the bound from Remark 4.1 we have

$$||\mathbf{X}_{I_{jl}} - \mathbf{X}^{jl}_{I_{jl}}|| \leq 2M,$$

and the conclusion follows. ∎

In typical situations, Σ and s_{jl} will have order n, resulting in a bound of the correct order, $n^{-1/2}$.

Acknowledgement. The authors would like to thank the organizers of the program 'Stein's method and applications: A program in honor of Charles Stein' and the Institute of Mathematical Sciences in Singapore for their most generous hospitality and an excellent meeting. Also many thanks to an anonymous referee for very helpful remarks.

Bibliography

1. Barbour, A.D. (1990). Stein's method for diffusion approximations, *Probability Theory and Related Fields* **84** 297-322.

2. Cacoullos, T. (1982). On upper and lower bounds for the variance of a function of a random variable. *Annals of Probability*, **10**, 799-809.

3. Cacoullos, T., and Papathanasiou, V. (1992). *Lower variance bounds and a new proof of the central limit theorem. Journal of Multivariate Analysis*, **43**, 173-184.

4. Cacoullos, T., Papathanasiou, V., and Utev, S. (1994). Variational inequalities with examples and an application to the central limit theorem. *Annals of Probability*, **22**, 1607-1618.

5. Chen, L.H.Y., and Shao, Q. (2004). Stein's method and normal approximation. Tutorial notes for the Singapore workshop on Stein's method, August 2003.

6. Götze, F. (1991). On the rate of convergence in the multivariate CLT. *Annals of Probability*, **19**, 724-739.

7. Goldstein, L. (2004). Normal approximation for hierarchical structures. To appear, *Annals of Applied Probability*

8. Goldstein, L. (2004). Berry Esseen bounds for combinatorial central limit theorems and pattern occurrences, using zero and size biasing. *Preprint.*

9. Goldstein, L., and Reinert, G. (1997). Stein's Method and the Zero Bias Transformation with Application to Simple Random Sampling *Annals of Applied Probability*, **7**, 935-952.

10. Goldstein, L., and Reinert, G. (2003). Distributional transformations, orthogonal polynomials, and Stein characterizations. *Preprint.*

11. Goldstein, L., and Rinott, Y. (1996). On multivariate normal approximations by Stein's method and size bias couplings. *Journal of Applied Probability* **33**, 1-17.

12. Papadatos, N., and Papathanasiou, V. (2001). Unified variance bounds and a Stein-type identity. In *Probability and Statistical Models with Applications*, Charalambides, Ch. A., Koutras, M. V., and Balakrishnan, N. (eds.), Chapman and Hall/CRC, New York, 87-100.

13. Raič, M. (2003). Normal approximations by Stein's method. In *Proceedings of the Seventh Young Statisticians Meeting*, Mvrar, A. (ed.), Metodološki zveski, 21, Ljubljana, FDV, 71-97.

14. Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. Proc. Sixth Berkeley Symp. Math. Statist. Probab. **2** 583-602, Univ. California Press, Berkeley.

15. Stein, C. (1986). Approximate Computation of Expectations. IMS, Hayward, CA.