Patchwork Distributions

Soumyadip Ghosh and Shane G. Henderson

Abstract Patchwork distributions are a class of distributions for use in simulation that can be used to model finite-dimensional random vectors with given marginal distributions and dependence properties. They are an extension of the previously developed chessboard distributions. We show how patchwork distributions can be selected to match several user-specified properties of the joint distribution. In constructing a patchwork distribution, one must solve a linear program that is potentially large. We develop results that shed light on the size of the linear program that one must solve. These results suggest that patchwork distributions should only be used to model random vectors with low dimension, say less than or equal to 5.

1 Introduction

Is there a part of stochastic simulation that George Fishman has not contributed to? If so, it is well hidden! His breadth of work, when multiplied by the length of time that he has been a major force, give rise to a very large area of contributions to the field. (Our apologies to the dimensional analysts that are trying to make sense of the last sentence.) So it is indeed an honour and a privilege to contribute to this volume in George Fishman's honour. Our contribution is in the area of input modeling. The one-dimensional case is well understood—see, for example, Fishman (2001, Chapter 10). But when we turn to higher dimensions, the situation is far less satisfactory.

The key issue is *statistical dependence* and its impact on performance measures. Indeed, much effort has been devoted to this problem in recent times. Recent applications include generating test problems for numerical algorithms (Hill and Reilly 2000), cost analysis (Lurie and Goldberg 1998), crop insurance pricing (Nelson 2004), and arrival process modeling (Avramidis et al. 2004). There are many classes of distributions that can be used to model (finite-dimensional) random vectors with given properties. For surveys, see Devroye (1986), Johnson (1987), and

S. Ghosh (✉)
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA
e-mail: ghoshs@us.ibm.com

C. Alexopoulos et al. (eds.), *Advancing the Frontiers of Simulation: A Festschrift in Honor of George Samuel Fishman*, International Series in Operations Research & Management Science 133, DOI 10.1007/b110059_4,

Biller and Ghosh (2004, 2006), and for a discussion of the modeling process, see Henderson (2005). In this paper we develop a new class of distributions of random vectors that we call *patchwork distributions*.

Relative to other methods, patchwork distributions have the desirable property that they afford considerable flexibility to the modeler in their ability to match properties of the distribution of the random vector. In particular, they can simultaneously match the marginal distributions of the components of the random vector, the covariance matrix, and the probability that the random vector lies within certain regions. This flexibility comes at a price: As we shall see, it can be computationally difficult to construct the desired distribution when the random vector has moderate to large dimension. Therefore, practically speaking, patchwork distributions are limited to low dimensions, say 5 or less.

Patchwork distributions are an extension of a certain class of distributions known as chessboard distributions (Ghosh and Henderson 2001, 2002), or "piecewise-uniform copulae" (Mackenzie 1994). They are constructed as follows. One assumes that the desired random vector $X = (X_1, \ldots, X_d)$ is given by $(F_1^{-1}(U_1), \ldots, F_d^{-1}(U_d))$, where F_i is the desired marginal distribution function of X_i, and U_i is a uniform random variable on $(0, 1)$. The problem then reduces to constructing the joint distribution of $(U_1, \ldots, U_d)$ on $(0, 1)^d$. (The joint distribution of $U = (U_1, \ldots, U_d)$ is known as a *copula*, and this approach of first generating U, and then constructing X from U using inversion is very common. The term "copula" was coined in Sklar 1959. See, e.g., Nelsen 1999 for background on copulas.) We break the unit cube $(0, 1)^d$ down into a grid of cells. Each cell is a hypercube with faces aligned with the axes. The conditional distribution of U given that it lies in one of the cells has a distribution with uniform marginals, which can vary from cell to cell. We call the distribution of X a *patchwork distribution*, and the distribution of the corresponding uniform random vector U a *patchwork copula*.

It is useful to allow cell-specific base copulas. For example, this allows patchwork distributions to match the extreme behaviour of X when all components are likely to move jointly. See, e.g., the discussion of tail dependence in Biller (2009).

Patchwork distributions generalize chessboard distributions, which have conditional (joint) uniform distributions, given that they lie in a fixed cell. The conditional distributions in the cells are fixed in advance heuristically, using any prior information about the joint distribution, and one then determines the vector giving the probabilities that U lies in each of the cells. This probability vector can be found by solving a certain linear program, the constraints of which reflect the desired properties of the joint distribution.

A natural property of a joint distribution that one might attempt to match is the correlation matrix, where the correlations could be Pearson product-moment, Spearman rank, or Kendall's tau correlations. We focus on Spearman rank correlations. As we will see, some rank correlation matrices are easier to match than others, in the sense that the size of the linear program that needs to be solved depends on the correlation matrix. Therefore, the computational effort required to construct a patchwork distribution is related to the correlation matrix.

The primary contributions of this paper are:

1. a characterization of the set of correlation matrices that patchwork distributions can match,
2. a theoretical analysis that relates the computational effort (size of the linear program) to a given correlation matrix, and
3. a computational study to shed further light on the results of the theoretical analysis.

The remainder of this chapter is organized as follows. In Section 2 we define patchwork distributions more carefully, and describe how one can select them to match the desired properties of X. We also describe how they can be generated. Then, in Section 3, we extend known results on the modeling power of chessboard distributions to patchwork distributions. Sections 4 and 5 contain, respectively, our theoretical and computational results on the size of the linear programs that must be solved to match a given correlation matrix. Finally, we offer some concluding remarks in Section 6.

2 Patchwork Distributions

In this section we define patchwork distributions, explain how they are constructed, and describe how to generate samples from them. For notational simplicity we mostly confine the discussion to 3-dimensional random vectors, but analogous results hold for $d \geq 2$ dimensions. Many of these results were proved for the special case of chessboard distributions in Ghosh and Henderson (2002). We give proofs of some of the extensions, even though they are usually similar to the special case of chessboard distributions, because they are helpful in understanding the structure of patchwork distributions.

We say that $X = (X_1, X_2, X_3)$ has a *patchwork distribution* if

$$X \stackrel{D}{=} (F_i^{-1}(U_i) : i = 1, 2, 3),$$

where F_i is the marginal distribution function of X_i, $i = 1, 2, 3$, and the distribution of $U = (U_1, U_2, U_3)$ is a *patchwork copula*, as described below.

Let $n \geq 1$, and divide $(0, 1]^3$ into a grid of n^3 equal-sized cubes (cells) with sides parallel to the coordinate axes. Let the cells be given by $C(j_1, j_2, j_3)$, with $j_1, j_2, j_3 = 1, \ldots, n$, so that

$$C(j_1, j_2, j_3) = \left\{(x_1, x_2, x_3) : \frac{j_i - 1}{n} < x_i \leq \frac{j_i}{n}, i = 1, 2, 3\right\}.$$

Conditional on lying in cell $C(j_1, j_2, j_3)$, U follows an appropriately scaled and translated version of a copula $\mathcal{C}(j_1, j_2, j_3)$, which can vary by cell. We call this copula the (j_1, j_2, j_3) *base* copula. To be more precise, let

$$(Z(j_1, j_2, j_3) : 1 \leq j_1, j_2, j_3 \leq n)$$

be independent random vectors where $Z(j_1, j_2, j_3)$ is distributed as $\mathcal{C}(j_1, j_2, j_3)$. Then, conditional on being in cell $C(j_1, j_2, j_3)$, the vector U is defined by

$$U_i = \frac{Z_i(j_1, j_2, j_3)}{n} + \frac{j_i - 1}{n}, \quad i = 1, 2, 3. \tag{1}$$

We allow the mass of each cell to vary. Let

$$q(j_1, j_2, j_3) = P(U \in C(j_1, j_2, j_3))$$

be the mass assigned to cell $C(j_1, j_2, j_3)$. We require that the $q(j_1, j_2, j_3)$s satisfy

$$\begin{aligned}
\sum_{j_2, j_3=1}^{n} q(j_1, j_2, j_3) &= P\left(U_1 \in \left(\frac{j_1 - 1}{n}, \frac{j_1}{n}\right]\right) = \frac{1}{n}, \quad \forall j_1 = 1, \ldots, n, \\
\sum_{j_1, j_3=1}^{n} q(j_1, j_2, j_3) &= P\left(U_2 \in \left(\frac{j_2 - 1}{n}, \frac{j_2}{n}\right]\right) = \frac{1}{n}, \quad \forall j_2 = 1, \ldots, n, \\
\sum_{j_1, j_2=1}^{n} q(j_1, j_2, j_3) &= P\left(U_3 \in \left(\frac{j_3 - 1}{n}, \frac{j_3}{n}\right]\right) = \frac{1}{n}, \quad \forall j_3 = 1, \ldots, n, \\
q(j_1, j_2, j_3) &\geq 0 \quad \forall j_1, j_2, j_3 = 1, \ldots, n.
\end{aligned} \tag{2}$$

We call the distribution of U, as constructed above, a *patchwork copula*. Theorem 1 below proves that the distribution is indeed a copula, and therefore that X has the desired marginal distributions.

Theorem 1 *If U is constructed as above, with cell probabilities q satisfying the constraints* (2)*, then U has uniform marginals. Consequently, X has the desired marginals.*

Proof Let the marginal distribution function of U_i be denoted by $G_i(\cdot)$. We show that $G_1(x) = x$ for $x \in (0, 1]$, and the proof for dimensions 2 and 3 is exactly the same. We rely on the conditional relationship (1). For any $x \in (i - 1, i]/n$, we have that

$$\begin{aligned}
G_1(x) &= \sum_{j_1 \leq i-1} \sum_{j_2, j_3=1}^{n} q(j_1, j_2, j_3) + \\
&\qquad \sum_{j_2, j_3=1}^{n} P\big((i-1)/n < U_1 \leq x | U \in C(i, j_2, j_3)\big) q(i, j_2, j_3) \\
&= \frac{i-1}{n} + \sum_{j_2, j_3=1}^{n} P\big(0 < Z_1(i, j_2, j_3) \leq n(x - (i-1)/n)\big) q(i, j_2, j_3)
\end{aligned}$$

$$= \frac{i-1}{n} + n\left(x - \frac{i-1}{n}\right) \sum_{j_2, j_3=1}^{n} q(i, j_2, j_3)$$
$$= \frac{i-1}{n} + x - \frac{i-1}{n} = x$$

as required. Hence, U has uniform marginals. Since X_i is obtained from U_i via the probability integral transform, it follows that X_i has the desired marginal distribution, $i = 1, 2, 3$. □

Remark 1 Chessboard copulas, as introduced in Ghosh and Henderson (2002), are patchwork copulas where all the base copulas are derived from independent uniform random variables. They coincide with the "piecewise-uniform copulae" developed by Mackenzie (1994). Cloned distributions (Johnson and Kotz 2004) are bivariate patchwork copulas where the base copula is the same for all cells, and all cells have the same probability.

The constraints (2) are sufficient to ensure that U has uniform marginals. So long as those constraints hold, we are then free to choose the cell probabilities to match other desired properties of the joint distribution. Covariance is one such property.

We believe that for non-Gaussian marginals, it is usually more appropriate to use rank covariance than product-moment covariance as a measure of dependence. Recall that the rank covariance between two random variables X_1 and X_2 with distribution functions F_1 and F_2 respectively is given by $E[F_1(X_1)F_2(X_2)] - E[F_1(X_1)]E[F_2(X_2)]$. Our preference for rank covariance over product-moment covariance stems from the facts that rank covariance is always well defined, irrespective of whether the X_is have finite second moments or not, and that rank covariance is invariant to strictly increasing transformations of the random variables. In the case where F_i is continuous, $F_i(X_i)$ is uniformly distributed on (0, 1]. Indeed, if X_1 and X_2 are components of a patchwork random vector with continuous marginal distribution functions, then the rank covariance between X_1 and X_2 equals the product-moment covariance between U_1 and U_2, from which X was constructed. Hence, we can reduce a study of rank covariance of patchwork distributions with arbitrary continuous marginals to one of product-moment covariance of uniform random variables on (0, 1] (rank and product-moment covariances coincide for uniform marginals).

Remark 2 When some of the marginal distributions are not continuous, this convenient relationship does not hold, and one must then attempt to match the desired correlations using more-complicated methods; see Avramidis et al. (2009).

We need an expression for the product-moment covariance of two components of a patchwork copula. Let U be distributed according to the patchwork copula. Then

$$\begin{aligned}\Sigma_{12}^{U} = \mathrm{Cov}(U_1, U_2) &= E[U_1U_2] - 1/4 \\ &= \sum_{j_1, j_2, j_3} q(j_1, j_2, j_3) E[U_1U_2 | U \in C(j_1, j_2, j_3)] - 1/4 \\ &= \sum_{j_1, j_2, j_3} q(j_1, j_2, j_3)\mu_{12}(j_1, j_2, j_3) - 1/4, \end{aligned} \tag{3}$$

where

$$\mu_{12}(j_1, j_2, j_3) = E[U_1 U_2 | U \in C(j_1, j_2, j_3)].$$

The μ terms are constants that depend on the base copulas, but not on q. It follows that the covariance between U_i and U_j is a linear function of q, for each $i, j = 1, \ldots, n$.

Suppose now that we want to match the true covariance matrix Σ^U to a desired covariance matrix Σ. The diagonal terms are all equal to $1/12$, and covariance matrices are symmetric, so we can measure the error $r(\Sigma^U, \Sigma)$ as

$$r(\Sigma^U, \Sigma) = \sum_{1 \le i < j \le 3} \left| \Sigma^U_{ij} - \Sigma_{ij} \right|.$$

It immediately follows that we can attempt to match Σ using the linear program

$$\min \quad \sum_{i=1}^{2} \sum_{j=i+1}^{3} (z^+_{ij} + z^-_{ij}) \tag{4}$$

$$\text{subject to} \quad \Sigma^U_{ij} - \Sigma_{ij} = z^+_{ij} - z^-_{ij}, \ i = 1,2 \text{ and } j = i + 1 \text{ to } 3$$

$$z^+_{ij} \ge 0, z^-_{ij} \ge 0, \text{ together with constraints (2) and (3).}$$

Remark 3 The linear program (4) is always feasible since $q(j_1, j_2, j_3) = n^{-3}$, for all j_1, j_2, j_3, is feasible. Also, the objective function is bounded below by 0, so an optimal solution exists. If the optimal objective value is 0, then we exactly match the desired properties.

Clemen et al. (2000) discussed a number of other properties that one might elicit from users about joint distributions and therefore want to match. Several of these are easily matched when X has a patchwork distribution. Ghosh and Henderson (2001) described how to match such properties using chessboard distributions, and the methods extend to patchwork distributions. For example, probabilities of the form $P(X \in A)$ for various regions A can be expressed as linear functions of q, and therefore can be matched using linear programming as above. The set A does not have to be rectilinear. Similarly, conditional fractiles can be matched using linear programming and concordance probabilities can be matched using quadratic programming.

It is relatively straightforward to generate random vectors that have a patchwork distribution. The basic procedure consists of the following steps:

1. Generate the (random) index (J_1, J_2, J_3) of the cell $C(J_1, J_2, J_3)$ containing the uniform random vector U from the discrete distribution formed by the $q(j_1, j_2, j_3)$s. With some preprocessing, it is possible to do this in constant

time using, e.g., the alias method (Walker 1977). The following description is adapted from Law and Kelton (2000):

- **Alias Method Setup:** Two arrays are calculated from the qs. The array $(AC_j : j = 1, \ldots, n^3)$ contains *cutoff values*, and the array $(AA_j : j = 1, \ldots, n^3)$ contains *aliases*. These arrays can be computed as follows; see Kronmal and Peterson (1979).

  a. Set $AC_j = n^3 q(j_1, j_2, j_3)\ \forall j = 1, \ldots, n^3$, where the index j represents the cell $C(j_1, j_2, j_3)$ via $j = (j_1 - 1)n^2 + (j_2 - 1)n + j_3$.
  b. Define sets $TG = \{j : AC_j \geq 1\}$ and $TS = \{j : AC_j < 1\}$.
  c. Do the following steps until TS becomes empty:

     i. Remove an element k from TG and remove an element m from TS.
     ii. Set $AA_m = k$ and replace AC_k by $AC_k - 1 + AC_m$.
     iii. If $AC_k < 1$, put k into TS; otherwise put k back in TG.

- **Generating Cells:** Once the arrays AA and AC have been calculated, the random cell index (J_1, J_2, J_3) (and equivalently the index J) can be generated as follows:

  a. Generate I from the discrete uniform distribution over $\{1, \ldots, n^3\}$ and $U \sim U(0, 1)$ independent of I.
  b. If $U \leq AC_I$, return $J = I$. Otherwise, return $J = AA_I$.

2. Generate U conditional on $U \in C(J_1, J_2, J_3)$ via (1). Here we need to be able to generate random vectors from the base copula $\mathcal{C}(J_1, J_2, J_3)$, but since we can select the base copula, this should present little difficulty.
3. Generate the components of X via $X_i = F_i^{-1}(U_i)$.

If q is an extreme-point solution to the d-dimensional version of the linear program (4), then there are on the order of nd strictly positive cell probabilities. The exact number of positive values depends on the number of equality constraints in the LP and the degree to which the extreme-point solution is degenerate. On the other hand, there are n^d cells. Therefore, for large d, the fraction of cells receiving positive mass is quite small.

The fact that nd is small relative to n^d can be viewed as an advantage with respect to variate generation since it reduces the setup time required to implement the alias method. However, it can also be viewed as a disadvantage. As the dimension d increases, the fraction of cells receiving positive probabilities is vanishingly small. This means that the set of values that the random vector X can assume is somewhat limited, and so the distributions take a nonintuitive form. As more constraints are added due to the need to match more distributional properties, the problem severity is reduced, but it still remains. Mackenzie (1994) avoids this problem by maximizing the entropy of the discrete distribution q. In this case, all of the cells receive positive probability. However, the problem of maximizing the entropy of q subject to linear constraints is a convex optimization problem that is more difficult to solve than the LPs discussed above. A computationally attractive alternative is to place a lower

bound on the cell probabilities. We do not discuss this issue further here as it would lead us too far afield.

3 Modeling Power

In this section we focus on matching covariance matrices. We say that a covariance matrix Σ is *feasible* if a copula exists with that covariance matrix. Let Ω denote the set of feasible covariance matrices. (We suppress the dependence on dimension d.) We view Ω as a subset of the vector space $\mathbb{R}^{d(d-1)/2}$ equipped with the usual inner product, because each $\Sigma \in \Omega$ is symmetric, the elements on the diagonal are all equal to $1/12$, and there are $d(d-1)/2$ elements above the diagonal. (It is therefore also a subset of $[-1/12, 1/12]^{d(d-1)/2}$.) In what follows, the notation Σ will represent both the actual covariance matrix and its vector form in Ω.

One might expect that Ω corresponds with the set of symmetric positive semidefinite matrices with diagonal elements equal to $1/12$. For $d \leq 3$, this is correct (Joe 1997), but for $d > 3$ it is not known whether this is the case or not. It *is* known that in any dimension d, Ω is convex, closed, and full-dimensional (Ghosh and Henderson 2002).

We are now ready to state some results about the modeling power of patchwork distributions. The proofs of these results are, in general, similar to the corresponding results for chessboard distributions (Ghosh and Henderson 2002) and so, for the most part, are omitted. We start with the following lemma, the proof of which is needed later in this paper.

Lemma 1 *Suppose that $\Sigma \in \Omega$. Then the optimal objective value of the linear program* (4) *is at most $d(d-1)/n$.*

Proof Since $\Sigma \in \Omega$, there exists a random vector V with uniform marginals and covariance matrix Σ. We modify the distribution of V as follows. We keep the total mass within each cell constant, but we modify the distribution of V within each cell to conform with the corresponding base copula. This process yields a patchwork copula corresponding to a random vector U, say. The cell probabilities (the qs) for U (which are the same as those for V) constitute a feasible solution to the linear program (4). Furthermore, we can bound the differences in the covariance matrices of V and U, as detailed below. These bounds translate into a bound on the objective value of the solution q. Since the optimal solution of the linear program can do no worse, we obtain a bound on the optimal objective value, thereby proving the result.

For now, assume that $d = 3$. Let $q(j_1, j_2, j_3) = P(V \in C(j_1, j_2, j_3))$ and note that

$$\begin{aligned}
&\mathrm{Cov}(U_1, U_2) - \Sigma_{12} \\
&= E[U_1 U_2] - E[V_1 V_2] \\
&= \sum_{j_1, j_2, j_3 = 1}^{n} \big(\mu_{12}(j_1, j_2, j_3) - E[V_1 V_2 | V \in C(j_1, j_2, j_3)]\big) q(j_1, j_2, j_3). \qquad (5)
\end{aligned}$$

But

$$\frac{j_1-1}{n}\frac{j_2-1}{n} \le E[V_1V_2|V \in C(j_1, j_2, j_3)] \le \frac{j_1}{n}\frac{j_2}{n}. \tag{6}$$

Combining (5) with (6) we see that

$$\begin{aligned}&\mathrm{Cov}(U_1, U_2) - \Sigma_{12}\\ &\le \sum_{j_1,j_2,j_3=1}^{n} q(j_1, j_2, j_3)\left(\mu_{12}(j_1, j_2, j_3) - \frac{(j_1-1)(j_2-1)}{n^2}\right)\end{aligned} \tag{7}$$

and

$$\mathrm{Cov}(U_1, U_2) - \Sigma_{12} \ge \sum_{j_1,j_2,j_3=1}^{n} q(j_1, j_2, j_3)\left(\mu_{12}(j_1, j_2, j_3) - \frac{j_1 j_2}{n^2}\right). \tag{8}$$

These bounds will prove useful later, but for now we obtain more-explicit bounds. The bounds (6) also apply to $\mu_{12}(j_1, j_2, j_3)$ and so from (7),

$$\begin{aligned}\mathrm{Cov}(U_1, U_2) - \Sigma_{12} &\le \sum_{j_1,j_2,j_3=1}^{n} q(j_1, j_2, j_3)\left(\frac{j_1 j_2}{n^2} - \frac{(j_1-1)(j_2-1)}{n^2}\right)\\ &= n^{-2}\sum_{j_1,j_2,j_3=1}^{n} q(j_1, j_2, j_3)(j_1 + j_2 - 1)\\ &\le n^{-2}\sum_{j_1,j_2,j_3=1}^{n} q(j_1, j_2, j_3)(2n-1)\\ &= \frac{2n-1}{n^2}.\end{aligned}$$

A lower bound follows similarly, so that

$$|\mathrm{Cov}(U_1, U_2) - \Sigma_{12}| \le \frac{2n-1}{n^2}. \tag{9}$$

The bound (9) was derived assuming $d = 3$, but the same argument and bound hold in higher dimensions. Hence, if Σ^U denotes the covariance matrix of U, we have that

$$r(\Sigma^U, \Sigma) \le \frac{d(d-1)}{2}\frac{2n-1}{n^2}$$

and the result follows. □

We can now state the main result of this section. Let A° and ∂A denote the interior and boundary of a set A.

Theorem 2 *Patchwork distributions can get arbitrarily close to any $\Sigma \in \Omega$ and can exactly match any $\Sigma \in \Omega^\circ$ (for sufficiently large n), but cannot necessarily exactly match $\Sigma \in \partial\Omega$. Furthermore, $\Sigma \notin \Omega$ iff the optimal objective value of the linear program* (4) *exceeds $d(d-1)/n$ for some $n \geq 1$.*

Proof This result is proved using the bounds given in Lemma 1. Most of the proof (specifically the $\Sigma \in \Omega^\circ$ and the $\Sigma \notin \Omega$ parts) is very similar to corresponding results in Ghosh and Henderson (2002) and so is omitted. All that needs to be shown is that a patchwork distribution may, or may not, be able to exactly match matrices on the boundary $\partial\Omega$. Ghosh and Henderson (2002) showed that chessboard distributions, which are a special case of patchwork distributions, cannot exactly match *any* matrix on the boundary for a finite n. It remains to show that *some* patchwork distributions can match *some* boundary covariance matrix exactly. This is trivially true, for instance, when the base copulas of the patchwork distribution all have a covariance $\Sigma^b \in \partial\Omega$. Then one can exactly match the boundary covariance matrix Σ^b with $n = 1$ using this base copula. □

Theorem 2 establishes that patchwork distributions can exactly match feasible covariance matrices lying in the interior of Ω, and that infeasible covariance matrices can be proved to be infeasible in finite time by the linear program, but that little can be concluded for covariance matrices lying on the boundary of Ω. The boundary matrices are feasible because Ω is closed, but why are they difficult to match?

Theorem 3 below shows that the joint distribution of a copula with a covariance matrix that lies on the boundary of Ω is a rather strange creature! Recall that any copula F can be decomposed into a singular part F_s and an absolutely continuous part F_{ac} with respect to Lebesgue measure restricted to $(0, 1]^3$. (This is simply the Lebesgue Decomposition; e.g., see Billingsley 1995, p. 414.) Thus, $F = F_{ac} + F_s$. Moreover, the absolutely continuous part has a density f_{ac} with respect to Lebesgue measure.

Theorem 3 *Suppose that f_{ac} is defined as above for a distribution F with covariance matrix $\Sigma \in \partial\Omega$. Then, there cannot exist an open set G such that*

$$f_{ac}(x) \geq \phi > 0 \quad \text{a.e. in } G. \tag{10}$$

(Recall that a property holds *almost everywhere (a.e.)* on the set G if it is true for all $x \in G$ except on a subset of Lebesgue measure 0.)

Proof For notational ease we give a proof in the 3-dimensional case. The general case is virtually identical. Suppose such a G exists. We assume, without loss of generality, that $f_{ac}(x) \geq \phi > 0$ for *all* $x \in G$ and not just a.e. (If not, just redefine f_{ac} on the set of measure 0.) Now, we can choose an open ball $B(x, \epsilon)$ within G and an open cubical region C with faces aligned with the axes within $B(x, \epsilon)$ such that the interior of C is non-empty. Split f_{ac} into two parts, f_C and $f_{\bar{C}}$, defined as:

$$f_C(x) = \begin{cases} \phi & x \in C \\ 0 & \text{elsewhere} \end{cases} \quad \text{and} \quad f_{\bar{C}}(x) = \begin{cases} f_{ac}(x) - \phi & x \in C \\ f_{ac}(x) & \text{elsewhere} \end{cases}.$$

Let u and v be the endpoints that define C, so that

$$C = \{(x_1, x_2, x_3) \in (0, 1]^3 : u_i < x_i \leq v_i,\ i = 1, 2, 3\}.$$

Divide the region C into 4 (equal size) subregions,

$$C_{ab} = \left\{(x_1, x_2, x_3) \in C : u_1 + (a-1)\frac{v_1 - u_1}{2} < x_1 \leq u_1 + a\frac{v_1 - u_1}{2}, \right.$$
$$\left. u_2 + (b-1)\frac{v_2 - u_2}{2} < x_2 \leq u_2 + b\frac{v_2 - u_2}{2} \right\},$$

for $1 \leq a, b \leq 2$.

Define a new distribution H from F as follows. The singular parts H_s and F_s coincide, as do the $h_{\bar{C}}$ and $f_{\bar{C}}$ parts, respectively, of the absolutely continuous density. The density h_C takes the value 2ϕ on C_{11} and C_{22}, and 0 on C_{12} and C_{21}. Then it is straightforward to show that H has uniform marginals, that the (1, 2)th covariance is strictly increased, and that the other covariances remain unchanged. Alternatively, if h_C takes the value 0 on C_{11} and C_{22}, and 2ϕ on C_{12} and C_{21}, then the covariance strictly decreases.

The argument above could be repeated for each pair of components. Convexity of Ω then implies that Σ must lie in the interior Ω° which is a contradiction, and the proof is complete. □

One consequence of Theorem 3 is that we cannot hope to exactly match covariance matrices on the boundary of Ω if we use a base copula which has a density component that satisfies (10) for some set G. This gives another explanation for why chessboard distributions cannot match covariance matrices on the boundary of Ω.

We have already seen a singular base copula that can exactly match a covariance matrix on the boundary of Ω. We might ask whether a base copula exists that can match all matrices on the boundary of Ω. We do not have a complete answer to this question, but we will shed further light on it in Section 4.

In summary, the import of the results in this section is that patchwork distributions

- can prove that a given covariance matrix is infeasible in finite time,
- can arbitrarily closely approximate any feasible covariance matrix,
- can exactly match any feasible covariance matrix in the interior of the set of feasible covariance matrices, but
- *might not* exactly match any covariance matrix on the boundary of the set of feasible covariance matrices.

4 Modeling Effort: Theoretical Results

In order to use a patchwork distribution, we need to perform the setup steps outlined in Section 2. The main computational bottleneck there is the solution of the linear programming problem. The time required to solve a linear program typically increases with the size of the linear program, which in turn depends on the discretization level n. So it is of interest to see how large n needs to be to match a given covariance matrix for a fixed dimension d of the random vector, and that is the subject of this section and the next one. Here we focus on theoretical analysis, while the next section performs some computational experiments.

We limit ourselves to the case where the patchwork distribution uses the same base copula in all cells, since this makes the arguments more elegant. Let $\Omega^n(\Sigma^b)$ represent the set of covariance matrices that can be matched by patchwork distributions of size n with a base copula that has Σ^b as its covariance matrix. (In many contexts, the argument Σ^b will be clear and hence shall be dropped.) The set Ω^n shares many of the properties of Ω, namely that it is non-empty, closed, convex, and full-dimensional in $\mathbb{R}^{d(d-1)/2}$ (Ghosh and Henderson 2002). We have shown in Theorem 2 that patchwork distributions can achieve any feasible covariance matrix in the interior of Ω for some finite n. Thus, in a sense the sequence $\{\Omega^n, n \geq 1\}$ grows to cover the whole of Ω as $n \to \infty$; we shall establish this rigorously and provide bounds on the rate of convergence in terms of n. Our results show that, roughly speaking, the set Ω^n is smaller than Ω by a factor that is somewhere between $(1 - \kappa_1/n)$ and $(1 - \kappa_2/n^2)$ for some constants κ_1 and κ_2. In order to state these results precisely we need some definitions.

Let $B(x, \epsilon) = \{y : \|x - y\|_2 < \epsilon\}$ be the (open) $\epsilon-$ball centered at x, defined under the l_2 metric on the space $\mathbb{R}^{m(d)}$, where $m(d) = d(d-1)/2$. The ball $B(0, 1)$, the unit open ball centered at the origin, is denoted by B. Thus, $B(x, \epsilon) = x + \epsilon B$, where the notation vM denotes the set $\{vx : x \in M\}$ for any scalar v, and $y + M = \{y + x : x \in M\}$.

We call any compact, convex set with a non-empty interior a *convex body*. The *Minkowski subtraction* set operation on two convex bodies M and N can be defined (Schneider 1993, Chapter 3) as

$$M \sim N \stackrel{\Delta}{=} \{x \in M : x + N \subset M\}.$$

A convex body E is said to be *centered* if it contains the origin as an interior point. Sangwine-Yager (1988) defines, for an $\epsilon > 0$, the ϵth *relative inner parallel body* of a convex body M with respect to a centered convex body E to be $M \sim \epsilon E$.

The families of sets $\mathcal{U}(\epsilon, \Sigma^b)$ and $\mathcal{L}(\epsilon)$ are indexed by ϵ and defined as

$$\mathcal{U}(\epsilon) \stackrel{\Delta}{=} \Omega \sim \epsilon\Omega,$$

$$\mathcal{U}(\epsilon, \Sigma^b) \stackrel{\Delta}{=} \mathcal{U}(\epsilon) + \epsilon\Sigma^b, \quad \text{and} \tag{11}$$

$$\mathcal{L}(\epsilon) \stackrel{\Delta}{=} \Omega \sim \epsilon B. \tag{12}$$

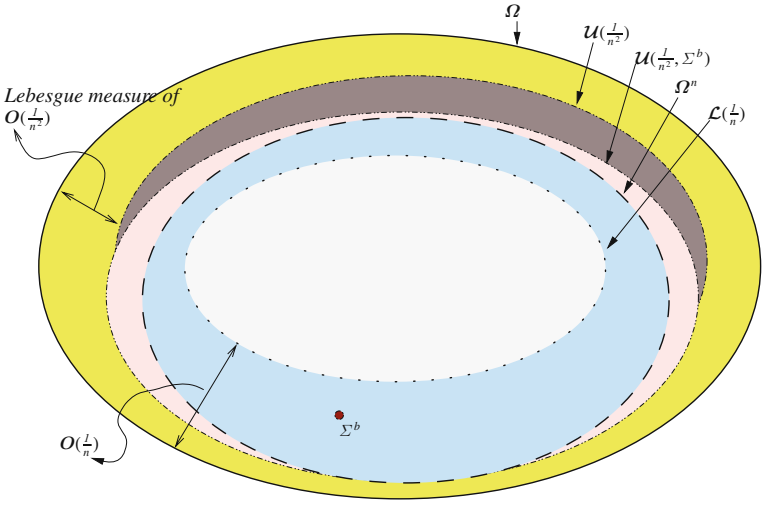

Fig. 1 The sets $\mathcal{U}(\frac{1}{n^2}), \mathcal{U}(\frac{1}{n^2}, \Sigma^b)$, and $\mathcal{L}(\frac{1}{n})$

These definitions are illustrated in Fig. 1.

A matrix z belongs to $\mathcal{U}(\epsilon) \subset \Omega$ if the set $z + \epsilon\Omega$ belongs to Ω. The set $\mathcal{U}(\epsilon)$ has a non-empty interior for all $0 < \epsilon < 1$. The set $\mathcal{U}(\epsilon) + \epsilon\Sigma^b$ is simply the set $\mathcal{U}(\epsilon)$ translated by the matrix $\epsilon\Sigma^b$. Similarly, a matrix z belongs to $\mathcal{L}(\epsilon)$ if the $\epsilon-$ball $B(z, \epsilon) \subset \Omega$. This has a simple interpretation, in that $\mathcal{L}(\epsilon)$ is the subset of points in Ω that are at least an l_2-distance ϵ away from the boundary $\partial\Omega$. Again, the sets $\mathcal{L}(\epsilon)$ can be empty for large ϵ, but are non-empty for sufficiently small $\epsilon > 0$. Note that the lower-bound sets $\mathcal{L}(\epsilon)$ are defined independent of the base covariance Σ^b.

We are now ready to state the main result of this section.

Theorem 4 *Let $\ell = \sqrt{m(d)}$. Then,*

$$a) \qquad \Omega^n(\Sigma^b) \subseteq \mathcal{U}\left(\frac{1}{n^2}, \Sigma^b\right), \text{ and}$$

$$b) \qquad \mathcal{L}\left(\frac{2\ell}{n}\right) \subseteq \Omega^n(\Sigma^b).$$

Theorem 4 establishes that the "gap" between $\Omega^n(\Sigma^b)$ and Ω has a width that is somewhere between $O(n^{-1})$ and $O(n^{-2})$. The following corollary uses that result to obtain bounds on the volume of the set $\Omega^n(\Sigma^b)$ relative to that of Ω. Let $\mathbb{L}$ represent Lebesgue measure on the real vector space $\mathbb{R}^{m(d)}$.

Corollary 1 *There is a constant $K(d)$ that depends on the dimension d such that*

$$\mathbb{L}(\Omega) - \frac{K(d)}{n} \leq \mathbb{L}(\Omega^n(\Sigma^b)) \leq \left(1 + \frac{1}{n^2}\right)^{-m(d)} \mathbb{L}(\Omega).$$

Corollary 1 formalizes the rather imprecise statement we made earlier about the rate at which Ω^n approaches Ω. The rate at which patchwork distributions can cover the set Ω of feasible covariance matrices is at least of the order $1 - K(d)n^{-1}$, but can be no faster than a factor of the order $(1 + n^{-2})^{-m(d)}$ which, in turn, is of the order $1 - m(d)n^{-2}$ when n is large. These results are illustrated in Fig. 1.

We now turn to proving these results.

Proof of Theorem 4(a) For notational ease we prove the result for $d = 3$. The case $d > 3$ is proved similarly. We establish the result by showing that a certain operation on any n-sized patchwork distribution having a covariance matrix $\Sigma \in \Omega^n$ constructs a new distribution with a new covariance matrix in Ω. One can obtain an upper bound on the distance between these matrices, which then gives the result.

Let $\{q(\cdot, \cdot, \cdot)\}$ represent the solution to the LP (4) that exactly matches a covariance matrix $\Sigma \in \Omega^n$. Then

$$\begin{aligned}\Sigma_{12} &= E[U_1U_2] - E[U_1]E[U_2] \\ &= \sum_{j_1,j_2,j_3=1}^{n} E[U_1U_2|U \in C(j_1, j_2, j_3)]q(j_1, j_2, j_3) - \frac{1}{4}. \end{aligned} \tag{13}$$

Let $Z = (Z_1, Z_2, Z_3)$ be a random vector distributed according to the base copula, and let $\Sigma^b \in \Omega$ be its covariance matrix. Since $E[Z_i] = 1/2$, $i = 1, 2, 3$, we see that

$$\begin{aligned}E[U_1U_2|U \in C(j_1, j_2, j_3)] &= E\left[\left(\frac{Z_1}{n} + \frac{j_1 - 1}{n}\right)\left(\frac{Z_2}{n} + \frac{j_2 - 1}{n}\right)\right] \\ &= \frac{E[Z_1Z_2]}{n^2} + \frac{j_1 + j_2 - 2}{2n^2} + \frac{(j_1 - 1)(j_2 - 1)}{n^2} \\ &= \frac{E[Z_1Z_2]}{n^2} + t(j_1, j_2), \end{aligned} \tag{14}$$

where $t(j_1, j_2)$ is a function only of j_1, j_2, and n.

Suppose now that we replace the base copula in each cell with another copula represented by the random vector Z'. The result is still a valid patchwork copula because of Theorem 1, and represents the distribution of a random vector U', say. If Σ' is the covariance matrix of U', then

$$\Sigma'_{12} = \sum_{j_1,j_2,j_3=1}^{n} \left(\frac{E[Z'_1Z'_2]}{n^2} + t(j_1, j_2)\right) q(j_1, j_2, j_3) - \frac{1}{4}. \tag{15}$$

Let $\Sigma^{b'}$ be the covariance matrix of Z'. The net change in covariance due to the replacement operation is, from (13), (14), and (15),

$$\Sigma'_{12} - \Sigma_{12} = \sum_{j_1,j_2,j_3=1}^{n} \frac{1}{n^2}(E[Z'_1Z'_2] - E[Z_1Z_2])\, q(j_1, j_2, j_3)$$

$$= \frac{1}{n^2}(\Sigma_{12}^{b'} - \Sigma_{12}^{b}). \tag{16}$$

Equation (16) holds for every component of the covariance matrix. Hence,

$$\Sigma' = \Sigma + \frac{1}{n^2}(\Sigma^{b'} - \Sigma^{b}),$$

and is contained in Ω. We can choose $\Sigma^{b'} \in \Omega$ arbitrarily. Thus,

$$\left(\Sigma + \frac{\Omega}{n^2}\right) - \frac{\Sigma^b}{n^2} \subset \Omega,$$

and we have established that for any $\Sigma \in \Omega^n$, $\Sigma \in \mathcal{U}(n^{-2}, \Sigma^b)$. This gives the result. □

This result is tight in a certain sense. Consider the case where chessboard copulae of size n are used to match a perfectly correlated uniform random vector with pairwise covariances all equal to $1/12$. (This target covariance matrix belongs to $\partial\Omega$.) A chessboard copula can be constructed by equally distributing its mass on the diagonal cells, and all pairwise covariances of this copula are equal to $1/12 - 1/12n^2$. If we perform the transformation described in the proof above, where $\Sigma^{b'}$ is the covariance matrix of a perfectly correlated uniform random vector (so all entries in the covariance matrix are equal to $1/12$), then we obtain the distribution of the perfectly correlated uniform random vector as a result. Thus, we see that Ω^n can have some points in common with the boundary of $\mathcal{U}(n^{-2}, \Sigma^b)$.

We now prove the second part of Theorem 4. First, recall that all norms in a real vector space are equivalent; see, for example, Golub and Van Loan (1996, p. 53). Indeed, for any $x \in \mathbb{R}^{m(d)}$,

$$\|x\|_\infty \;\leq\; \|x\|_2 \;\leq \ell\; \|x\|_\infty. \tag{17}$$

Proof of Theorem 4(b) The result is trivial if $\mathcal{L}(2\ell/n)$ is empty, so assume it is nonempty. The proof of Lemma 1 derived a bound on the optimal objective function of the linear program (4). Specifically, if $\Sigma \in \Omega$ denotes a target covariance matrix and Σ^n is an optimal solution to the linear program then, from (9),

$$|\Sigma(i, j) - \Sigma^n(i, j)| \leq \frac{2}{n} \quad \forall\ 1 \leq i < j \leq 3. \tag{18}$$

Equation (18) shows that we can get within $l_\infty-$distance $2/n$ from any $\Sigma \in \Omega$ using patchwork distributions. From (17), we then have that $\Sigma^n \in B(\Sigma, 2\ell/n)$. Hence, in particular, for any $\Sigma \in \partial\Omega$, we can pick a matrix $\Sigma^n \in \Omega^n$ such that $\Sigma^n \in B(\Sigma, 2\ell/n)$.

Now, suppose the assertion in the theorem is false, and there exists a $\Lambda \in \mathcal{L}(2\ell/n)$ that does not belong to Ω^n. Since Ω^n is convex, the celebrated Separating

Hyperplane Theorem (see, e.g., Luenberger 1969, Theorem 3, Section 5.12) gives us a hyperplane $\mathcal{H}$ through Λ that separates the point Λ from Ω^n.

Consider a line $\mathcal{N}$ passing through Λ that is orthogonal to the hyperplane $\mathcal{H}$. Busemann (1958, Chapter 1) tells us that since Λ is in the interior of Ω, this line intersects the boundary $\partial\Omega$ of the convex set Ω at exactly two points, say Σ^1 and Σ^2. By definition, the point $\Lambda \in \mathcal{L}(2\ell/n)$ does not belong to either of the sets $B(\Sigma^i, 2\ell/n)$, $i = 1, 2$. Thus, $\mathcal{H}$ separates each of the sets $B(\Sigma^i, 2\ell/n)$, $i = 1, 2$, from Λ. Moreover, the sets lie on opposite sides of $\mathcal{H}$, since $\Lambda \in \Omega^\circ$. Thus, at least one ball is separated from Ω^n by the hyperplane $\mathcal{H}$. But this contradicts the earlier observation that one can always choose a point that belongs to Ω^n from each ball $B(\Sigma^i, 2\ell/n)$, $i = 1, 2$. This completes the proof. □

In order to prove Corollary 1, we need the following result. Brannen (1997, Theorem 1) quotes a lower bound from Sangwine-Yager (1988) for the Lebesgue measure of a relative inner parallel body $M \sim \epsilon E$. That result establishes that

$$\mathbb{L}(M \sim \epsilon E) \geq \mathbb{L}(M) - \epsilon S(M; E) + R(m(d), \epsilon),$$

where $S(M; E)$ represents the *relative surface area* of M with respect to E, and the function $R(m(d), \epsilon)$ is nonnegative. They also give conditions under which $S(M; E)$ is finite and positive, and these are satisfied by using the sets Ω and B in the definition of $\mathcal{L}(\epsilon)$ as M and E respectively. Thus, if $\epsilon < 1$, then

$$\mathbb{L}(\mathcal{L}(\epsilon)) \geq \mathbb{L}(\Omega) - k(d)\epsilon \tag{19}$$

for some positive constant $k(d)$ that possibly depends on the dimension $m(d)$ of the sets.

Proof of Corollary 1 From (19), for n large enough that $2\ell/n < 1$,

$$\mathbb{L}(\Omega) - k(d)\left(\frac{2\ell}{n}\right) \leq \mathbb{L}\left(\mathcal{L}\left(\frac{2\ell}{n}\right)\right),$$

where $k(d)$ is a positive value that depends on d. This equation, along with Theorem 4(b), gives the lower bound in the statement of the result with $K(d) = 2k(d)\ell$.

For the upper bound, first note that $\mathcal{U}(n^{-2}, \Sigma^b)$ is a translation of the set $\mathcal{U}(n^{-2})$, and so both sets have the same Lebesgue measure. Also, if $\Lambda \in \mathcal{U}(n^{-2})$, then, by definition, $\Lambda + n^{-2}\Omega \subseteq \Omega$. In particular, $\Lambda + n^{-2}\Lambda \in \Omega$, i.e., $\Lambda \in (1 + n^{-2})^{-1}\Omega$. Hence,

$$\mathcal{U}(n^{-2}) \subseteq (1 + n^{-2})^{-1}\Omega.$$

The Lebesgue measure of the linearly scaled set $(1 + n^{-2})^{-1}\Omega$ is given by $(1 + n^{-2})^{-m(d)}\mathbb{L}(\Omega)$ (see Billingsley 1995, Theorem 12.2). This, along with

Theorem 4(a), establishes the upper bound on the Lebesgue measure of Ω^n and we are done. □

We conclude this section by showing that for any fixed choice of base copula, there will exist covariance matrices in Ω that cannot be exactly matched, no matter how n is chosen. This result shows that it is pointless to attempt to identify a "powerful" base copula that matches all of Ω for some finite n.

Proposition 1 *For any fixed base copula, there exists a covariance matrix $\Sigma \in \Omega$ that cannot be exactly matched for any n.*

Proof On the contrary, suppose that such a base copula exists, and let Σ^b be its covariance matrix. Consider a line $\mathcal{N}$ through Σ^b and the origin. (If Σ^b is equal to the origin, then pick an arbitrary line through the origin.) Since Ω is compact, convex, and the origin is in its interior, this line intersects $\partial\Omega$ at two points. Follow the line from Σ^b through the origin until you reach one of those points. Call that point $\bar{\Sigma}$. By the supposition, a Σ^b-based patchwork copula, of size n, say, can exactly match $\bar{\Sigma}$. Then, by the argument establishing Theorem 4(a), a Σ' of value

$$\Sigma' = \bar{\Sigma} + \frac{\bar{\Sigma} - \Sigma^b}{n^2}$$

can also be achieved by replacing the base copula in each cell with a copula that has covariance matrix $\bar{\Sigma}$. The matrix Σ' is, however, outside Ω, and we have the desired contradiction. □

Whether patchwork copulas with a bounded number of base copulas can match all of Ω for some finite n is an open problem. We conjecture that this is impossible.

5 Modeling Effort: Computational Results

Corollary 1 proves that patchwork distributions with discretization level n can match covariance matrices that are a distance r from the boundary set $\partial\Omega$, where the order of r lies somewhere between n^{-1} and n^{-2}. In this section we describe a computational study that sheds further light on this rate for the special case of chessboard distributions.

Let $\mathcal{S}$ be a collection of $d(d-1)/2-$dimensional vectors that represent the off-diagonal elements of covariance matrices Σ in $\mathbb{R}^{d\times d}$. Consider rays from the origin through each of these vectors. We determine the rate at which each ray is "covered" by chessboard distributions as the discretization level n grows. We populate the set $\mathcal{S}$ by sampling uniformly from the set of all positive semidefinite matrices of size $d \times d$. (Ghosh and Henderson 2003 provides such a sampler.) This allows us to test whether the rate varies in different regions of Ω.

The origin is in the strict interior of the set of all positive semidefinite (PSD) matrices, which implies that there is a finite maximum value $r^{sd}(\Sigma) > 0$ such

that $r\Sigma$ is positive semidefinite for all $r^{sd}(\Sigma) \geq r > 0$. We compute $r^{sd}(\Sigma)$ by formulating and solving a semidefinite program. The set of feasible covariance matrices Ω also contains the origin in its strict interior, and so a finite maximum $r^*(\Sigma)$ exists such that $r\Sigma$ is a feasible covariance matrix for uniform marginals for all $r^*(\Sigma) \geq r > 0$. Finally, let $r^*(n, \Sigma)$ represent the point at which the ray $r\Sigma$ intersects the set Ω^n of all covariance matrices that chessboards of size n can exactly match. Fig. 2 illustrates these definitions.

We numerically determine $r^*(n, \Sigma)$ by solving, for each $\Sigma \in \mathcal{S}$, the LP

$$\begin{aligned} r^*(n, \Sigma) = \quad & \max r \\ \text{s.t.} \quad & r\Sigma(i, j) = \text{Cov}(X_i, X_j), \ \forall i < j, \\ & r \geq 0, \end{aligned} \tag{20}$$

along with the constraints (2) that ensure that the distribution of X is a copula. The LPs (20) are feasible ($r = 0$, $q = 1/n^d$ is a feasible solution) and terminate with a positive finite optimal solution $r^*(n, \Sigma)$. The rate of change in $r^*(n, \Sigma)$ along rays $r\Sigma$, $r > 0$, $\forall \Sigma \in \mathcal{S}$, provides an indication of the rate at which the set Ω^n covers Ω. To see why, recall that two convex closed bodies that differ in their sizes by a small $\epsilon > 0$ also differ in volume (Lebesgue measure) by the same order. (We used this result in the proof of Corollary 1.)

For the $d = 3$ case, the set Ω is known to coincide with the set of all PSD matrices (e.g., Joe 1997) and thus $r^*(\Sigma) = r^{sd}(\Sigma)$, which facilitates the calculation of the exact coverage rate. Figure 3 plots the relative difference $(r^{sd}(\Sigma) - r^*(n, \Sigma))/r^{sd}(\Sigma)$ against n in the log-log scale for 20 different values of $\Sigma \in \mathcal{S}$ and n taking values up to 128. (Here, and in what follows, in the interest of notational

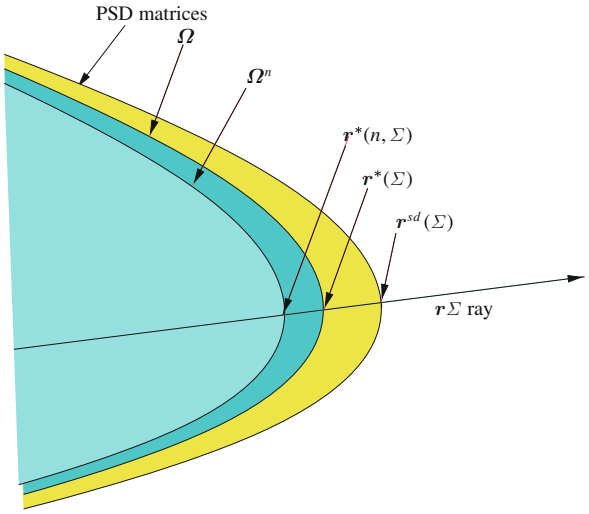

Fig. 2 The points $r^{sd}(\Sigma)$, $r^*(\Sigma)$, and $r^*(n, \Sigma)$

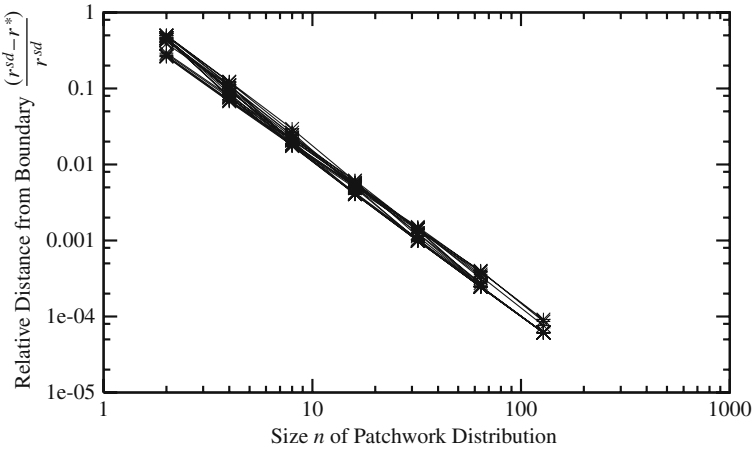

Fig. 3 Log-log plots for 20 covariance rays $r\,\Sigma$, $r > 0$, in 3 dimensions

brevity we write r^* and r^{sd} in place of $r^*(\Sigma)$ and $r^{sd}(\Sigma)$ when no confusion should arise.) Table 1 provides the estimated slopes of the curves for the 20 illustrated rays. The slope estimates are calculated from only a handful of sample points and are thus noisy, but the slopes hover close to -2 in all the cases tested. Thus, it would seem likely that the coverage rate is closer to n^{-2} than n^{-1} for $d = 3$.

Table 1 Slopes of log-log plots for 20 covariance rays in 3 dimensions. All values reported to 3 decimal places

Correlations			
Σ_{12}	Σ_{13}	Σ_{23}	Calculated Slope
−0.996	−0.055	0.075	−1.902
−0.944	−0.300	−0.136	−2.162
−0.912	0.280	0.299	−2.066
−0.773	0.371	0.514	−2.175
−0.731	−0.188	0.656	−2.117
−0.613	0.427	0.664	−2.197
−0.488	−0.070	0.870	−2.136
−0.300	0.223	0.928	−2.130
−0.118	0.142	0.982	−2.035
0.198	0.256	0.946	−2.001
−0.989	−0.118	0.092	−1.900
−0.912	−0.324	−0.251	−2.136
−0.849	−0.122	0.514	−2.135
−0.731	−0.678	0.080	−2.120
−0.713	−0.634	−0.300	−2.171
−0.592	0.493	0.638	−2.200
−0.368	−0.327	0.870	−2.035
−0.236	−0.221	0.946	−1.879
−0.021	0.514	0.857	−2.166
0.336	0.360	0.870	−1.921

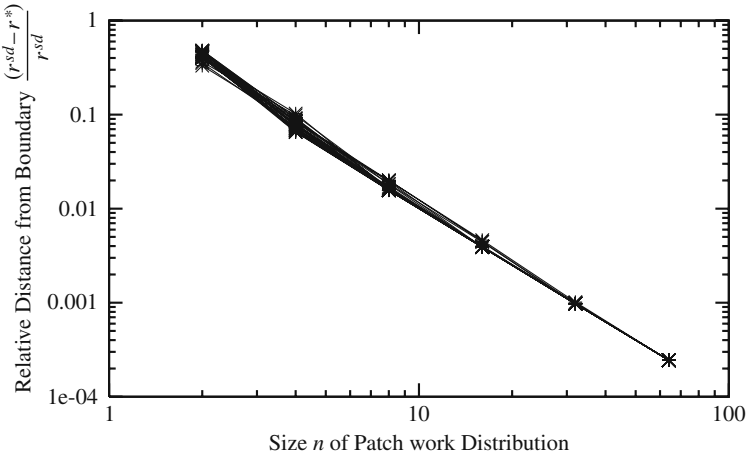

Fig. 4 log-log plots for 20 covariance rays $r\Sigma$, $r > 0$, in 4 dimensions

In the case of higher dimensions $d \geq 4$, the value r^* is not known (in general $r^* \leq r^{sd}$), and the rates of coverage calculated via the same log-scale plot as in the $d = 3$ case will only yield approximate values. Figure 4 plots the relative difference $(r^{sd} - r^*(n, \Sigma))/r^{sd}$ against n in the log-log scale for 20 different values of $\Sigma \in \mathcal{S}$ for dimension $d = 4$. If r^* were strictly less than r^{sd}, then $(r^{sd} - r^*(n, \Sigma))/r^{sd}$ would not drop linearly (in the log scale) to 0 with increasing n. No non-linearity is manifest in the range of values $n = [1, 64]$ plotted. The plots seem fairly linear, with the slopes varying within $[-2.117, -2.013]$.

Our implementation could not solve the linear programs for larger n due to numerical instability (the relative error is below 0.0004 for $n = 64$). It might be the case that the difference $(r^{sd} - r^*)$ is non-zero but too small to be detected by our tests. This indicates that for the $d = 4$ case, $r^* = r^{sd}$ is a good assumption for practical purposes, and that the rate of coverage is again closer to the upper bound in Corollary 1.

6 Conclusions

We have shown that patchwork distributions represent a flexible class of distributions that can match many desired properties of a random vector. They are primarily effective in low dimensions, say up to dimension 5. (We have solved patchwork LPs in dimension 5 for $n = 32$ using a column generation technique, but we will not report in detail on these ideas here.) The primary bottleneck in constructing patchwork distributions is solving a certain linear program, the size of which is related to the discretization level n. The discretization level required to match a given correlation matrix depends on the distance of the correlation matrix to the boundary of the set of all feasible correlation matrices. Our theoretical and computational results

give strong evidence that the set of feasible correlation matrices not matched by patchwork distributions using discretization level n diminishes at the rate n^{-2}.

Acknowledgments The authors would like to thank the editors for their superb efforts in improving the presentation of this chapter. This work was partially supported by National Science Foundation grants DMI 0400287 and CMMI 0800688.

References

Avramidis, A. N., N. Channouf, and P. L'Ecuyer. 2009. Efficient correlation matching for fitting discrete multivariate distributions with arbitrary marginals and normal copula dependence. *INFORMS Journal on Computing* 21:88–106.

Avramidis, A. N., A. Deslauriers, and P. L'Ecuyer. 2004. Modeling daily arrivals to a telephone call center. *Management Science* 50:896–908.

Biller, B. 2009. Copula-based multivariate input models for stochastic simulation. *Operations Research* forthcoming.

Biller, B., and S. Ghosh. 2004. Dependence modeling for stochastic simulation. In *Proceedings of the 2004 Winter Simulation Conference*, eds. R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, pp. 153–161. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Biller, B., and S. Ghosh. 2006. Multivariate input processes. In *Handbooks in Operations Research and Management Science: Simulation*, eds. B. L. Nelson and S. G. Henderson, pp. 123–154. Amsterdam: Elsevier Science.

Billingsley, P. 1995. *Probability and Measure*, 3rd ed. New York: Wiley.

Brannen, N. S. 1997. The Wills conjecture. *Transactions of the American Mathematical Society* 349:3977–3987.

Busemann, H. 1958. *Convex Surfaces*. Interscience Tracts in Pure and Applied Mathematics, No. 6. New York: Interscience.

Clemen, R. T., G. W. Fischer, and R. L. Winkler. 2000. Assessing dependence: Some experimental results. *Management Science* 46:1100–1115.

Devroye, L. 1986. *Non-Uniform Random Variate Generation*. New York: Springer-Verlag.

Fishman, G. S. 2001. *Discrete-event Simulation: Modeling, Programming and Analysis*. Springer Series in Operations Research. New York: Springer.

Ghosh, S., and S. G. Henderson. 2001. Chessboard distributions. In *Proceedings of the 2001 Winter Simulation Conference*, eds. B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, pp. 385–393. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Ghosh, S., and S. G. Henderson. 2002. Chessboard distributions and random vectors with specified marginals and covariance matrix. *Operations Research* 50:820–834.

Ghosh, S., and S. G. Henderson. 2003. Behaviour of the NORTA method for correlated random vector generation as the dimension increases. *ACM Transactions on Modeling and Computer Simulation* 13:276–294.

Golub, G. H., and C. F. Van Loan. 1996. *Matrix Computations*, 3rd ed. Baltimore, Maryland: The Johns Hopkins University Press.

Henderson, S. G. 2005. Should we model dependence and nonstationarity, and if so how? In *Proceedings of the 2005 Winter Simulation Conference*, eds. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, pp. 120–129. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Hill, R. R., and C. H. Reilly. 2000. The effects of coefficient correlation structure in two-dimensional knapsack problems on solution procedure performance. *Management Science* 46:302–317.

Joe, H. 1997. *Multivariate Models and Dependence Concepts*. London: Chapman & Hall.

Johnson, M. E. 1987. *Multivariate Statistical Simulation*. New York: Wiley.

Johnson, N. L., and S. Kotz. 2004. Cloning of distributions. Preprint.

Kronmal, R. A., and A. V. Peterson. 1979. On the alias method for generating random variables from a discrete distribution. *American Statistician* 33:214–218.

Law, A. M., and W. D. Kelton. 2000. *Simulation Modeling and Analysis*, 3rd ed. New York: McGraw-Hill.

Luenberger, D. G. 1969. *Optimization by Vector Space Methods*. New York: John Wiley and Sons.

Lurie, P. M., and M. S. Goldberg. 1998. An approximate method for sampling correlated random variables from partially-specified distributions. *Management Science* 44:203–218.

Mackenzie, G. R. 1994. Approximately maximum-entropy multivariate distributions with specified marginals and pairwise correlations. PhD dissertation, Department of Decision Sciences, University of Oregon, Eugene, Oregon.

Nelsen, R. B. 1999. *An Introduction to Copulas*. Lecture Notes in Statistics, vol. 139. New York: Springer-Verlag.

Nelson, B. L. 2004. Personal communication.

Sangwine-Yager, J. R. 1988. Bonnesen-style inequalities for Minkowski relative geometry. *Transactions of the American Mathematical Society* 307:373–382.

Schneider, R. 1993. *Convex Bodies: The Brunn-Minkowski Theory*. Cambridge, UK: Cambridge University Press.

Sklar, A. 1959. Fonctions de répartition à *n* dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris* 8:229–231.

Walker, A. J. 1977. An efficient method for generating discrete random variables with general distributions. *ACM Transactions on Mathematical Software* 3:253–256.