Population Obfuscation: A Masking Problem and Some Solutions*

Michael Frey[1], Adam Wunderlich[1], Randy Hoover[2], Kyle Caudle[2], Lucas Koepke[1], David Newton[1]

[1]National Institute of Standards and Technology, Boulder, CO
[2]South Dakotas School of Mines and Technology, Rapid City, SD

Abstract

Sample obfuscation is the widely studied, challenging problem of providing access to a data sample while guarding aspects of its privacy. Sample obfuscation can take different forms, including masking or redaction to protect sample variables or anonymization or the methodology of differential privacy to secure individuals' data records. This work extends the notion of sample obfuscation to obfuscation of populations. Population obfuscation aims to protect information and features of a whole statistical population of data, the population being represented by an algorithm, formula, model, or sampling plan from which users can synthesize or otherwise access unlimited numbers of data records. Canonical sample masking can be extended to allow masking generally of functions of sample variables. With this extension we present a conceptual framework for population masking, with elementary examples of both canonical and general population masking. Three procedures are outlined for masking a population, one based on transfer learning, one on data augmentation, and one on optimal transport. We also introduce the idea of inherent population masking and offer a simple class of examples in which it occurs.

Key words: obfuscation, data privacy, masking, transfer learning, optimal transport

1 Introduction

Data managers are often charged to share data samples that have proprietary or sensitive elements—data entries, variables, or individuals' whole records—whose privacy must be maintained. Meeting these conflicting goals of data access and privacy is a challenging sample obfuscation problem that has been broadly studied from a variety of perspectives [1, 2, 3, 4].

Sample obfuscation is any technique intended to preserve the privacy of sample data. Figure 1 shows samples of size n in standard format with columns representing variables V_k and rows containing data record entries x_{jk} for units U_j. Sample obfuscation can be aimed variously to directly protect units, variables, or entries.

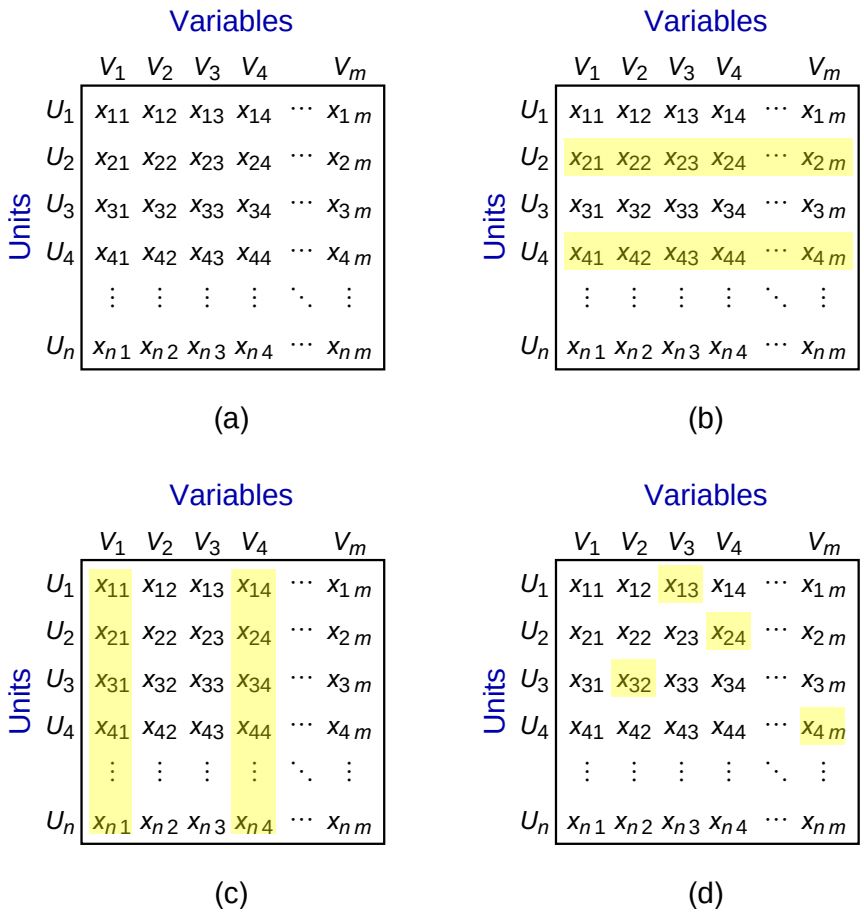

Figure 1: (a) A sample in standard format. (b) A set of sample units (individuals) chosen for obfuscation. (c) A set of sample variables chosen for obfuscation. (d) A set of sample entries chosen for obfuscation.

For example, differential privacy and anonymization are both intended to secure the privacy of unit data records, redaction and masking are common techniques for protecting variables [5], and individual entries can be protected by tokenization [6]. Figure 1 illustrates these different possible targets of sample obfuscation.

Beyond the obfuscation of finite samples of data lies the prospect of obfuscating whole data populations of unlimited size. Increasingly sophisticated generative data models are being built for research, simulation, training, and system testing. These generative models, implemented by an algorithm, formula, or machine learning architecture, allow users to synthesize at will an unlimited number of data records. Managers of these generative models are challenged to make the models available to disparate users while securing proprietary or sensitive elements of the population represented by the generative model. Obfuscating an entire population of unlimited size, rather than a fixed finite sample, presents data managers with a new challenge, distinct from sample obfuscation.

Population obfuscation aims to protect the information and features of a whole

statistical population of data. We focus on continuous populations[1] and distinguish these populations from samples by size; samples are finite in size, continuous populations are infinite. Then, the generative mechanism for a population can be conceptualized as random drawings from a sample with an infinite number of units (rows). Viewing the population as a sample with an infinity of units, we see that unit privacy is inherently protected; in any finite number of random drawings from the continuous population, any finite set of population units has zero probability of being drawn and observed. Similarly, any finite set of entries, no matter the entries' natures or locations in the population, has zero probability of being drawn. The privacy of individual units and entries in a continuous population is inherently protected by the population's infinite size; population variables are not similarly protected, even when the number of variables is infinite. Population obfuscation is fundamentally different in this sense from sample obfuscation.

The remainder of the paper is organized into five sections. Section 2 reviews sample masking, with examples of both canonical and general sample masking. This review sets the stage for population masking introduced in Sect. 3. There, the insights of Sect. 2 are used to propose a mathematical framework for population masking. This framework is explored in some simple examples and a population masking problem is presented. Section 4 describes three approaches to this masking problem, and Sect. 5 offers a time series example in which the population is inherently masked by the nature of the time series and the variables chosen for masking. Section 6 summarizes this work and offers some final remarks.

2 Sample masking

Unit and entry privacy are inherently protected in a continuous population, so the aim of population obfuscation is to secure population variables. Masking is a prominent means to securing variables in samples, and the purpose of this work is to extend the idea of sample masking to populations. Variables explicitly present in the sample are termed canonical variables, and in the literature sample masking means masking one or more of these canonical variables [7, 8, 9, 10]. We term this form of masking canonical masking and introduce general masking to refer to masking variables which are non-trivial functions of the sample's canonical variables. This section reviews both forms of sample masking, canonical and general. This sets the stage in the following section for a precise formulation of population masking.

We illustrate canonical masking using an example from human biometry. The data for this example, shown in Fig. 2(a), consist of ten units (adult male individuals) and three variables, height H, weight W, and age A. We wish to secure W in the sample but otherwise make the data for (H, A) fully available. Here W is called a marked variable (marked for masking) and H and A are the set of free variables. Masking W means that we are willing to share the individuals' joint distribution of H and A in the sample. In particular, we are willing to share the association between H and A. We are also willing to share the sample's marginal distribution

[1] A continuous population is a population whose (multivariate) distribution is continuous. Such populations are necessarily (uncountably) infinite in size. Moreover, any finite set of members of a continuous population has zero probability of being observed.

	H (m)	W (kg)	A (yr)
U_1	1.71	73	31
U_2	1.74	77	42
U_3	1.64	81	78
U_4	1.69	80	56
U_5	1.94	96	34
U_6	2.01	95	23
U_7	1.78	84	45
U_8	1.65	72	57
U_9	1.90	87	39
U_{10}	1.81	78	27

(a)

	H (m)	W^*(kg)	A (yr)
U_1	1.71	80	31
U_2	1.74	72	42
U_3	1.64	77	78
U_4	1.69	78	56
U_5	1.94	96	34
U_6	2.01	73	23
U_7	1.78	87	45
U_8	1.65	81	57
U_9	1.90	84	39
U_{10}	1.81	95	27

(b)

Figure 2: Canonical sample masking. (a) Data sample with weight variable W (highlighted) marked for masking. (b) Data sample with weight masked by sampling without replacement from W.

of W. Masking means that we do not want values of W to be traceable back to the individuals from which they originated; equivalently, we are saying that (H, A) identifies individuals, and we do not want to reveal the association between W and (H, A).

To mask W in Fig. 2(a) we create a new data set, still with all ten units and all three variables. We transfer the data for H and A without change into the new data set. For the weight variable W^* in the masked data set, though, we randomly sample without replacement from W in the original sample to fill the entries for W^* for the ten individuals. The result is shown in Fig. 2(b). The masked data set has intact both the marginal joint distribution of (H, A) and the marginal distribution of W. Masking breaks the association of the masked variable with the rest of the sample distribution; whatever association W and (H, A) had in the original data sample, W^* and (H, A) will tend to be approximately independent in the masked sample. In Fig. 2, for example, the correlation between H and W in the original sample is 0.86; between H and W^* in the masked sample, it is 0.06.

Body weight was masked in the biometry example in Fig. 2 using sampling without replacement[2]. Instead, sampling with replacement could have been used; any sampling scheme is allowed that preserves the marginal distribution of W while rendering W independent of the free variables in the sample.[3] Also, only weight W was marked for masking. In general, any subset of sample variables can be marked for masking. Masking multiple variables is accomplished by sampling the marked variables according to some scheme that preserves their joint distribution while rendering them independent of the remaining free sample variables. If the set

[2]Masking is often called swapping in cases where sampling without replacement is used.

[3]In practice, because of the random sampling involved in masking, we would not expect either perfect preservation of W's distribution or complete independence of W and (H, A).

of marked variables and the set of free variables happen to be already independent of one another in the original sample, we say the sample is inherently masked with respect to the chosen marked variables.

The biometry sample in Fig. 2 in which body weight W was marked for masking is an example of canonical masking; the variables H, W, and A in the originally given sample are called canonical variables and the marked variable W is one of them. We may want to mask variable(s) that are not present in the given sample but, instead, are derived from them. For example, suppose we want to mask body mass index $B = W/H^2$, while sharing gross size $S = HW$ and age A. The variable B is a function of W and H, and masking variable(s) determined by a function of the sample's canonical variables is called general masking. Canonical masking is a special case of general masking in which the derived marked variables and the derived free variables trivially belong to the set of canonical variables.

The process of masking body mass index B in the biometry example is shown in Fig. 3. In Step 1 in Fig. 3, we transform the sample's canonical variables to make explicit the variable B that we want to mask and the free variables S and A that we want to share. This transformation g is called a marking map because it identifies (marks) the variables to be masked and those to be shared. In Step 2 we mask B in the transformed sample just as we would in canonical masking. Then, in Step 3 we transform the masked result back into a sample with variables H, W, and A. This three-step process yields a sample with the original canonical variables H, W, and A but in which B is now close to independent of (S, A), as indicated by the correlations in Table 1. Note that the entries for H and W in Fig. 3 have changed from the original sample (a) to the masked sample (d). The purpose of masking is not to preserve individual entries; its goal is to preserve the distributions of both the marked and shared variables while rendering the two distributions independent.

Original sample	Masked sample
$\text{corr}(B, S) = -.46$	$\text{corr}(B^*, S) = -.01$
$\text{corr}(B, A) = .94$	$\text{corr}(B^*, A) = .15$

Table 1: Correlations with the marked variable B in the original sample and in the masked sample. The association between B and (S, A) is reduced to close to zero by masking.

In our example of general sample masking in Fig. 3, the map $g = (g_x, g_y)$ with

$$g_x : (H, W, A) \to B, \quad g_y : (H, W, A) \to (S, A)$$

is invertible, allowing us in Step 3 of Fig. 3 to recover the original sample's canonical variables. Other marking maps g exist that could be used instead; for example,

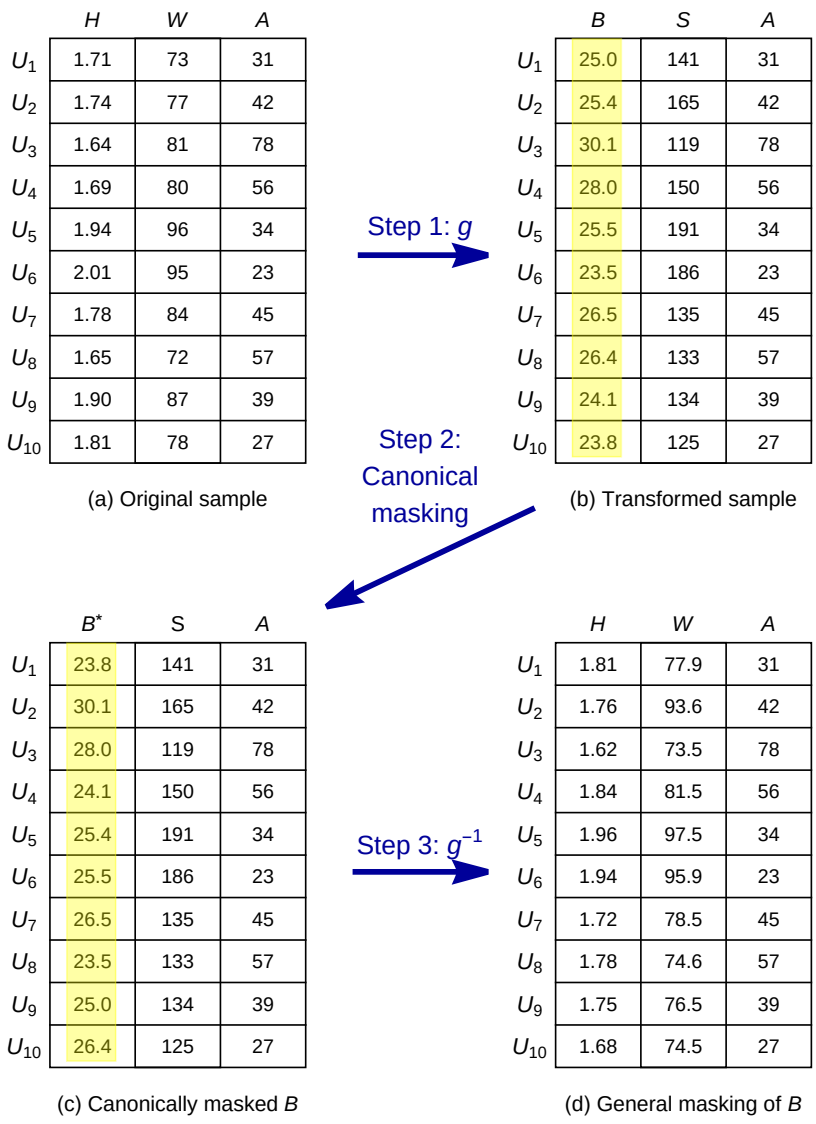

Figure 3: General sample masking. Step 1 transforms the original sample (a) according to an invertible map g which makes explicit in the transformed sample (b) the variable B (highlighted) to be masked and the variables S and A to be shared. The variable B is canonically masked in Step 2, yielding sample (c). Step 3 uses g^{-1} to arrive at the desired result (d) in which the data set has its original form in terms of H, W, and A, but now with B masked relative to the variables S and A to be shared.

general masking with $g' = (g'_x, g'_y)$ with

$$g'_x : (H, W, A) \to \frac{1}{B}, \qquad g'_y : (H, W, A) \to (S^3, \sqrt{A})$$

achieves the same result as with g. On the other hand, the map $g'' = (g''_x, g''_y)$ with

$$g''_x : (H, W, A) \to B, \qquad g''_y : (H, W, A) \to (H, A)$$

is invertible, but its effect is different; g separates B and (S, A) while g'' separates B and (H, A). A marking map's key requirements are that it be invertible and that it identify both the variables to be masked and the free variables to be shared.

3 Framework for population masking

We propose in this section a mathematical framework for population masking. Within this framework we state a population masking problem for study. This problem is non-trivial, yet practically solvable with sufficient resources.

3.1 Definitions

Marked population: A population with one or more variables marked for masking is called a marked population $(\vec{M}, \mathcal{M})$, where $\mathcal{M}$ is the population distribution and $\vec{M} \sim \mathcal{M}$ is a generic member of the marked population. The marked variables in $(\vec{M}, \mathcal{M})$ can be canonical variables, originally present in $(\vec{M}, \mathcal{M})$ or, more generally, they can be functions of the population's canonical variables. A population $(\vec{M}, \mathcal{M})$ is defined to be marked if a map g exists such that

$$g(\vec{M}) = \begin{pmatrix} g_x(\vec{M}) \\ g_y(\vec{M}) \end{pmatrix} = \begin{pmatrix} \vec{X} \\ \vec{Y} \end{pmatrix}. \tag{1}$$

The map g is called the marking map for $(\vec{M}, \mathcal{M})$; g identifies the variables $\vec{X} = g_x(\vec{M})$ in $(\vec{M}, \mathcal{M})$ that are to be masked and the variables $\vec{Y} = g_y(\vec{M})$ to be shared. The population marking is canonical when g is just a reordering of the components of $\vec{M}$ so that its first components $g_x(\vec{M})$ are the marked variables and its other components $g_y(\vec{M})$ are the free variables. The marking map for a population with a given set of marked variables is not unique; g_x and g_y, though, must be such that g is invertible. For general marking maps g, the random vectors $\vec{X} = g_x(\vec{M})$ and $\vec{Y} = g_y(\vec{M})$ are typically dependent. When they are independent, we say the marked population $(\vec{M}, \mathcal{M})$ is inherently masked with respect to the marked variables in $\vec{X}$.

Masked population: A population $(\vec{T}, \mathcal{T})$ is a masking of a marked population $(\vec{M}, \mathcal{M})$ with marking map g if there exist random vectors $\vec{H}$ and $\vec{Y}$ such that[4]

$$\vec{T} \stackrel{d}{=} g^{-1} \begin{pmatrix} \vec{H} \\ \vec{Y} \end{pmatrix}, \tag{2}$$

where $\vec{H}$ and $\vec{X} = g_x(\vec{M})$ have the same distribution and where $\vec{H}$ and $\vec{Y} = g_y(\vec{M})$ are independent. A masked population $(\vec{T}, \mathcal{T})$ is equivalently the counterpart a marked population $(\vec{M}, \mathcal{M})$ with marking map g if

$$\vec{T} \stackrel{d}{=} g^{-1} \begin{pmatrix} g_x(\vec{M}) \\ g_y(\vec{M}') \end{pmatrix}, \tag{3}$$

[4]The notation $\stackrel{d}{=}$ means equal in distribution.

where $\vec{M}$ and $\vec{M}'$ are two members drawn independently from the marked population $(\vec{M}, \mathcal{M})$. We call a map $h : \vec{M} \to \vec{T}$ that maps the marked population $(\vec{M}, \mathcal{M})$ to the target population $(\vec{T}, \mathcal{T})$ a masking map. Like the marking map g, the masking map h for $(\vec{M}, \mathcal{M})$ is not unique.

3.2 A population masking example

To illustrate population masking's definitions (2) and (3), we present an example of a trivariate marked population $(\vec{M}, \mathcal{M})$. Suppose $\vec{M}$ is uniformly distributed within the unit cube

$$Q = \{(m_1, m_2, m_3) : 0 \leq m_1 \leq 1, 0 \leq m_2 \leq 1, 0 \leq m_3 \leq 1\}.$$

For $\vec{m} = (m_1, m_2, m_3) \in Q$, radial distance of $\vec{m}$ from the origin is

$$r = \sqrt{m_1^2 + m_2^2 + m_3^2},$$

and its associated azimuthal and polar angles are

$$\theta = \text{Tan}^{-1} \frac{m_2}{m_1}, \quad \phi = \text{Tan}^{-1} \frac{m_3}{\sqrt{m_1^2 + m_2^2}}.$$

Suppose we want to mask radial distance R of members of $(\vec{M}, \mathcal{M})$ but are otherwise willing to share the joint distribution of the angles Θ, Φ. This is an example of general masking because R is a nontrivial function of the population's canonical variables M_1, M_2, and M_3. An invertible marking map (1) for this purpose has

$$g_x(\vec{M}) = \sqrt{M_1^2 + M_2^2 + M_3^2}, \quad g_y(\vec{M}) = \begin{pmatrix} \text{Tan}^{-1} \frac{M_2}{M_1} \\ \text{Tan}^{-1} \frac{M_3}{\sqrt{M_1^2 + M_2^2}} \end{pmatrix},$$

where $\vec{M} = (M_1, M_2, M_3)$. Define the random vector

$$\vec{S} = \begin{pmatrix} (M_1^2 + M_2^2 + M_3^2)^{1/2} \\ \text{Tan}^{-1} \frac{M_2}{M_1} \\ \text{Tan}^{-1} \frac{M_3}{\sqrt{M_1^2 + M_2^2}} \end{pmatrix},$$

where $\vec{M} \sim \mathcal{M}$, and let $\mathcal{S}$ be the distribution of $\vec{S}$. Corresponding to $\mathcal{S}$ is the distribution $\mathcal{S}^o$ defined by the random vector

$$\vec{S}^o = \begin{pmatrix} (M_1'^2 + M_2'^2 + M_3'^2)^{1/2} \\ \text{Tan}^{-1} \frac{M_2}{M_1} \\ \text{Tan}^{-1} \frac{M_3}{\sqrt{M_1^2 + M_2^2}} \end{pmatrix},$$

where $\vec{M}' \in \mathcal{M}$ is another member of the marked population drawn independently of $\vec{M}$. According to (3), the inverse mapping g^{-1} applied to the distribution $\mathcal{S}^o$ yields the target masked population $\mathcal{T}$. Members of $(\vec{T}, \mathcal{T})$ carry the same (marginal and joint) distributional information about angles and the same distributional information about radial distance as $(\vec{M}, \mathcal{M})$. In $(\vec{T}, \mathcal{T})$, though, the association between

angles and radial distance in $\mathcal{M}$ is no longer present. Radial distance in $\mathcal{T}$ is independent of the angles, and the radial distance of each population member $\vec{T} \sim \mathcal{T}$ is masked. Each member of $\mathcal{T}$ has a radial distance—it can be calculated from the components of $\vec{T} \sim \mathcal{T}$—but this radial distance comes as if from another independently drawn population member $\vec{T}' \sim \mathcal{T}$.

The marked variable R is masked in this example and the angles Θ and Φ are free. However, we could just as well have made R free and designated (Θ, Φ) to be masked. The result would be the same. Masking is mathematically symmetric in its treatment of the sets of marked and free variables; the effect of masking is to break the statistical association between the two sets while preserving the joint distribution of each set of variables.

3.3 A class of bivariate masking examples

An accessible class of examples of population masking is the two-dimensional case in which the marked population $(\vec{M}, \mathcal{M})$ has the bivariate normal distribution

$$\mathcal{M} = N(\vec{0}, \Sigma_{\mathcal{M}}), \quad \Sigma_{\mathcal{M}} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \rho & 0 \\ 0 & 1/\rho \end{pmatrix} \tag{4}$$

and the marking map $g : \vec{M} \to \vec{S}$ is linear, determined by a unitary rotation $\vec{S} = \mathbf{U}\vec{M}$. Here,

$$\vec{M} = \begin{pmatrix} M_1 \\ M_2 \end{pmatrix}, \quad \vec{S} = \begin{pmatrix} X \\ Y \end{pmatrix},$$

the unitary matrix $\mathbf{U}$ is

$$\mathbf{U} = \begin{pmatrix} \cos\gamma & \sin\gamma \\ -\sin\gamma & \cos\gamma \end{pmatrix},$$

and the marked variable X in $\vec{S}$ is $X = M_1 \cos\gamma + M_2 \sin\gamma$. This example of general masking has two parameters, the rotation angle γ associated with g and the elongation $\rho = \sigma_1/\sigma_2$ of the ellipse associated with the marked distribution $\mathcal{M}$. The distribution $\mathcal{S}$ of $\vec{S}$ is normal with zero mean and covariance matrix $\Sigma_{\mathcal{S}} = \mathbf{U}\Sigma_{\mathcal{M}}\mathbf{U}^\top$. Some calculation yields[5]

$$\Sigma_{\mathcal{S}} = \Sigma_{\mathcal{M}} - \nu \sin\gamma \begin{pmatrix} \sin\gamma & \cos\gamma \\ \cos\gamma & -\sin\gamma \end{pmatrix}. \tag{5}$$

The distribution $\mathcal{S}^o$ is the counterpart of $\mathcal{S}$ with the same marginal distributions as X and Y in $\mathcal{S}$, but in $\mathcal{S}^o$ these marginal distributions are independent. Therefore, $\mathcal{S}^o$ is normal with zero mean and, from (5), covariance matrix

$$\Sigma_{\mathcal{S}^o} = \Sigma_{\mathcal{M}} - \nu \sin^2\gamma \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

The target masked distribution $\mathcal{T}$ corresponding to $\mathcal{M}$ is normal with zero mean and covariance matrix $\Sigma_{\mathcal{T}} = \mathbf{U}^\top \Sigma_{\mathcal{S}^o} \mathbf{U}$, so

$$\mathcal{T} = N(\vec{0}, \Sigma_{\mathcal{T}}), \quad \Sigma_{\mathcal{T}} = \frac{1}{2}\Sigma_{\mathcal{M}} + \frac{\mu}{4}\mathbf{I} + \frac{\nu}{4} \begin{pmatrix} \cos 4\gamma & \sin 4\gamma \\ \sin 4\gamma & -\cos 4\gamma \end{pmatrix}. \tag{6}$$

[5] We use the notations $\mu = \rho + 1/\rho$ and $\nu = \rho - 1/\rho$.

The distributions $\mathcal{M}$ and $\mathcal{T}$ are shown in Fig. 4 along with that of the bivariate standard normal distribution $\mathcal{Z} = \mathbf{N}(\vec{0}, \mathbf{I})$ for two different combinations of the parameters ρ and γ.

In population masking the marking map g maps the marked distribution $\mathcal{M}$ to the distribution $\mathcal{S}$ in which the marked variables $\vec{X}$ and the free variables $\vec{Y}$ are dependent (unless $\mathcal{M}$ is inherently masked). A masking map $h : \vec{M} \to \vec{T}$ transports $\mathcal{M}$ to the target distribution $\mathcal{T}$ in which the marked variables $\vec{X}$ and the free variables $\vec{Y}$ are independent. In the present bivariate example a linear masking map h from $\vec{M} \sim \mathcal{M}$ to $\vec{T} \sim \mathcal{T}$ can be found from optimal transport theory. This map is of the form $\vec{T} = \mathbf{V}\vec{M}$ where

$$\mathbf{V} = \sqrt{\Sigma_{\mathcal{T}}\Sigma_{\mathcal{M}}^{-1}} = \frac{\sqrt{\zeta}\sqrt{\zeta+2}}{2\sqrt{2}}\mathbf{I} + \frac{\sqrt{\zeta-2}}{2\sqrt{2}\sqrt{\zeta}}\begin{pmatrix} -\mu\sin 2\gamma & 2\rho\cos 2\gamma \\ \frac{2}{\rho}\cos 2\gamma & \mu\sin 2\gamma \end{pmatrix} \tag{7}$$

with $\zeta = \sqrt{4 + \nu^2 \sin^2 2\gamma}$. The masking map $h : \vec{M} \to \vec{T}$ defined by $\vec{T} = \mathbf{V}\vec{M}$ with $\mathbf{V}$ in (7) is the map that "moves" $\mathcal{M}$ to $\mathcal{T}$ in such a way as to minimize the earth-mover (EM) distance between them.

3.4 A population masking problem

A marked population $(\vec{M}, \mathcal{M})$ is readily masked if sufficient appropriate information is available. For example, suppose both $(\vec{M}, \mathcal{M})$ and a marking map g are known. Then, in principle, according to either definition (2) or (3), the target masked population $(\vec{T}, \mathcal{T})$ can be constructed without error. By this we mean that there is a

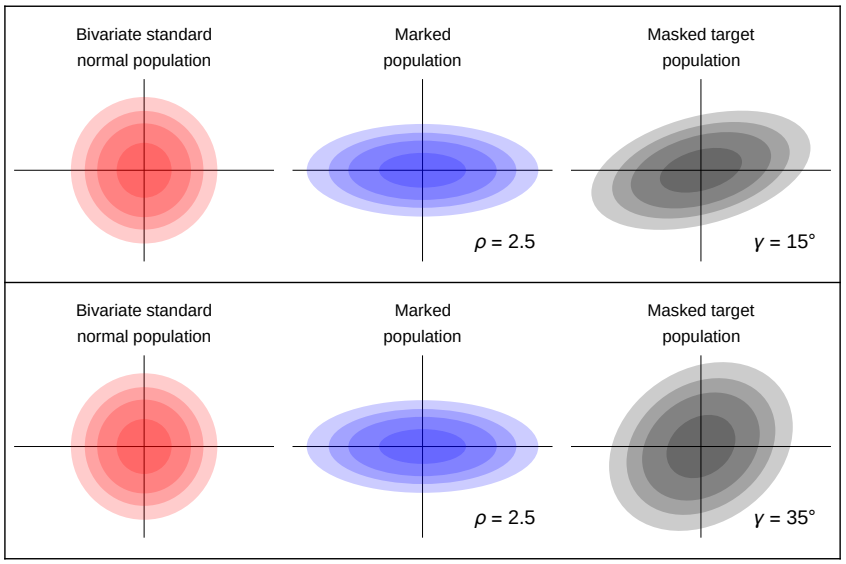

Figure 4: Bivariate normally distributed marked and masked populations $\mathcal{M}$ and $\mathcal{T}$ along with the default bivariate standard normal distribution $\mathcal{Z} = \mathbf{N}(\vec{0}, \mathbf{I})$ for two different combinations of elongation $\rho = 2.5$ and rotation $\gamma = 15°, 35°$.

procedure that will generate members $\vec{T}$ from the masked population distribution $\mathcal{T}$. For example, if both $\mathcal{M}$ and g are known, then according to (3) one would draw two members $\vec{M}$, $\vec{M}'$ independently from $\mathcal{M}$, calculate $\vec{X} = g_x(\vec{M})$ and $\vec{Y} = g_y(\vec{M}')$, and use the inverse of g to obtain $\vec{T} \sim \mathcal{T}$. This procedure can be repeated to obtain an unlimited number of observations from the masked population $(\vec{T}, \mathcal{T})$.

A more interesting problem arises when the marking map g is unavailable. This might happen, for example, if the determinative data authority does not want to so transparently identify the population variables it deems sensitive. Possibly, all that is available are samples of data $\mathcal{S}_\mathcal{M}$ and $\mathcal{S}_\mathcal{T}$ from $\mathcal{M}$ and $\mathcal{T}$—in particular, g is unknown or unavailable. This may be the case with archived data collected for one purpose but now being considered for a new purpose. If the sample $\mathcal{S}_\mathcal{T}$ of target data is large enough, then the target population $\mathcal{T}$ can be modeled by any of various procedures, and the model $\hat{\mathcal{T}}$ can be used to generate masked observations $\vec{T} \sim \hat{\mathcal{T}}$. But suppose that the sample $\mathcal{S}_\mathcal{T}$ is not large and, indeed, that it is so small that an estimate $\hat{\mathcal{T}}$ of the masked population cannot be constructed with satisfactorily small error. Of course, a sample $\mathcal{S}_\mathcal{T}$ of some size must be given because in the absence of g some information has to stand in for the unknown marked population variables. If $\mathcal{S}_\mathcal{T}$ is too small, data must be found elsewhere to aid in constructing $\hat{\mathcal{T}}$. These other data may come in the form of a sample $\mathcal{S}_\mathcal{M}$ from the marked population. These data are just the data that the data authority does not want to reveal, so the statistician tasked with constructing $\hat{\mathcal{T}}$ must have the authority's trust. And the sample $\mathcal{S}_\mathcal{M}$ must be relatively large—large enough that $\mathcal{S}_\mathcal{M}$ and $\mathcal{S}_\mathcal{T}$ together can support an estimate $\hat{\mathcal{T}}$ with satisfactorily small error.

The preceding line of thought suggests the following population masking problem. Suppose we are give a sample $\mathcal{S}_\mathcal{T}$ from the target masked population and a sample $\mathcal{S}_\mathcal{M}$ from the marked population, and we want an estimate $\hat{\mathcal{T}}$ of the target population distribution $\mathcal{T}$; in other words:

$$\text{Problem: given samples } \mathcal{S}_\mathcal{T} \text{ and } \mathcal{S}_\mathcal{M}, \text{ estimate } \mathcal{T}. \tag{8}$$

No other information is available for (8); in particular, the marking map g is not known. We want an estimate $\hat{\mathcal{T}}$ of the target population $\mathcal{T}$ with small estimation error $d(\hat{\mathcal{T}}, \mathcal{T})$. Problem (8) is interesting when 1) the size $N_\mathcal{T}$ of $\mathcal{S}_\mathcal{T}$ is small, too small by itself to meet the goal for the allowed error $d(\hat{\mathcal{T}}, \mathcal{T})$, and 2) the size $N_\mathcal{M}$ of $\mathcal{S}_\mathcal{M}$ is large, large enough to successfully estimate $\mathcal{T}$. The following section explores different approaches to (8).

The population masking problem in (8) posed above is just one of many possible, depending on the scenario under consideration. Any meaningfully formulated problem, though, must address two concerns. First, of course, information must be available in some form to in some way estimate the target population $\mathcal{T}$. This estimation might be accomplished either directly by some means or indirectly by, for example, first estimating one or the other of the masking and marking maps, g or h. Problem (8) addresses this first concern in the simplest way by supposing a random sample $\mathcal{S}_\mathcal{T}$ from $\mathcal{T}$ is given. Second, estimating the multivariate population $\mathcal{T}$ demands some sufficient amount of data, either data $\mathcal{S}_\mathcal{T}$ directly from $\mathcal{T}$ or data from a different source that can augment $\mathcal{S}_\mathcal{T}$. The problem we pose supposes that a large sample $\mathcal{S}_\mathcal{M}$ is available from the marked population $\mathcal{M}$. This large sample addresses the

concern for a large amount of data. Different population masking problems might be posed, and their solutions will entail different solution approaches. The following section lays out three approaches for the masking problem in (8).

4 Three approaches to population masking

We describe three approaches to the population masking problem in (8) posed in the previous section, one based on data augmentation, one based on optimal transport, and one based on transfer learning. Each approach assumes that we have a relatively large sample $\mathcal{S}_\mathcal{M}$ from the marked population and only a small sample $\mathcal{S}_\mathcal{T}$ from the target, masked population.

Two of our three approaches described below involve generative adversarial networks (GANs). A GAN is a generative model introduced by Goodfellow *et al.* [11, 12] involving two artificial neural networks (ANNs), a generator G_θ and a discriminator D_ϕ, where θ and ϕ are the ANNs' weight vectors. These two ANNs are trained simultaneously, with the one pitted against the other, in an iterative fashion according to a combined loss function $\mathcal{L}(\theta, \phi)$. At each iteration the generator presents learned examples to the discriminator, and the discriminator attempts to correctly identify these learned examples from among a pool of training examples. The discriminator reports its successes and failures back to the generator, each network updates its weights based on the discriminator's successes and failures, and a new iteration begins. At each iteration the generator is trying to mislead the discriminator, and the discriminator is trying to avoid being misled. After enough iterations, the generator and discriminator approach a game-theoretic equilibrium where the discriminator cannot distinguish between synthesized examples and training examples any better than guessing. At this equilibrium the GAN generator is trained and ready for use.

4.1 Data augmentation approach

Our population masking problem specifies that we have a sample $\mathcal{S}_\mathcal{T}$ from the target population but that $\mathcal{S}_\mathcal{T}$ is not sufficiently large by itself to construct an estimate $\hat{\mathcal{T}}$ to within the allowed error $d(\hat{\mathcal{T}}, \mathcal{T})$. Insufficient training data is a common problem in machine learning model training, and data augmentation schemes have been proposed for different settings to use data related to the training data to aid model training [13, 14, 15, 16, 17, 18].

The sample $\mathcal{S}_\mathcal{T}$ in our masking problem is insufficient by itself to estimate $\mathcal{T}$. The sample $\mathcal{S}_\mathcal{M}$ is large, though, and $\mathcal{S}_\mathcal{M}$ is similar to $\mathcal{S}_\mathcal{T}$ in important respects; $\mathcal{S}_\mathcal{M}$ and $\mathcal{S}_\mathcal{T}$ come from populations with the same marginal distribution of free variables and the same marginal distribution of marked variables. We propose as illustrated in Fig. 5 to use the data in $\mathcal{S}_\mathcal{M}$ to augment $\mathcal{S}_\mathcal{T}$ for training a GAN. In Fig. 5 the sample $\mathcal{S}_\mathcal{T}$ provides the training data for the GAN. Usually, a GAN is trained with and its generator is driven by white noise $\mathcal{Z}$, $\mathcal{Z}$ being typically the distribution of a vector of independent standard uniform or standard Gaussian variables. We propose, instead, to estimate the marked distribution $\mathcal{M}$ from $\mathcal{S}_\mathcal{M}$ and use the estimate $\hat{\mathcal{M}}$ as the GAN generator's source.

The intuition behind Fig. 5's data augmentation approach is illustrated in Fig.

6. If the distance $d(\mathcal{M}, \mathcal{T})$ in Fig. 6 is smaller than the distance $d(\mathcal{Z}, \mathcal{T})$, and if the metrical distance d well-reflects the distance reduction that might otherwise be gained by a larger training sample $\mathcal{S}_\mathcal{T}$, then $\mathcal{S}_\mathcal{T}$ could be augmented by $\mathcal{S}_\mathcal{M}$ to good effect. If the distance $d(\mathcal{M}, \mathcal{T})$ is not smaller than $d(\mathcal{Z}, \mathcal{T})$, using $\mathcal{M}$ to solve problem (8) may actually be counterproductive.

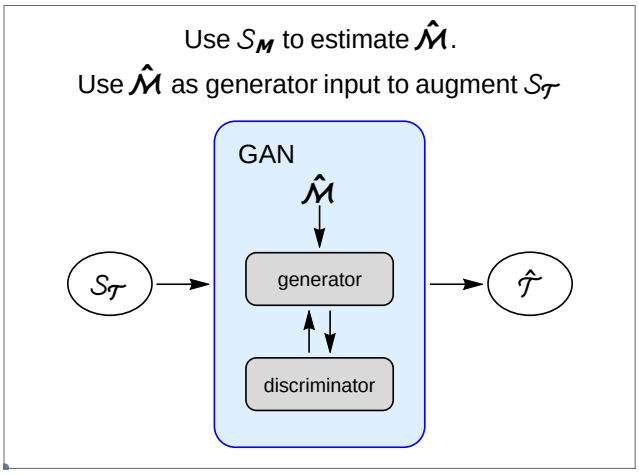

Figure 5: A data augmentation approach to the population masking problem. The small sample $\mathcal{S}_\mathcal{T}$ available for training the GAN is augmented by driving the GAN generator with data from the marked population estimate $\hat{\mathcal{M}}$.

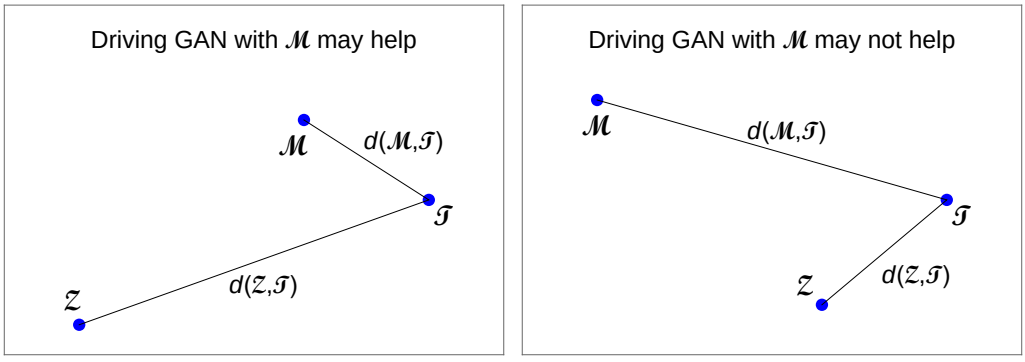

Figure 6: Relative distances of a marked population distribution $\mathcal{M}$ and a multivariate standard normal distribution $\mathcal{Z}$ to a target masked distribution $\mathcal{T}$. The left panel shows $\mathcal{M}$ closer than $\mathcal{Z}$ to $\mathcal{T}$, in which case driving a GAN generator with $\mathcal{M}$ may be advantageous. The right panel shows $\mathcal{Z}$ closer to $\mathcal{T}$, in which case $\mathcal{M}$ may offer no advantage.

To see how the distances $d(\mathcal{M}, \mathcal{T})$ and $(d(\mathcal{Z}, \mathcal{T})$ in Fig. 6 might actually compare, consider the class of examples in Subsect. 3.3. In these examples both the marked distribution $\mathcal{M}$ in (4) and the target, masked distribution $\mathcal{T}$ in (6) are bivariate normal. The default latent distribution $\mathcal{Z} \sim N(\vec{0}, \mathbf{I})$ is also bivariate normal. The distances separating $\mathcal{M}$, $\mathcal{Z}$, and $\mathcal{T}$ can be measured in different ways. The EM distance separating two zero-mean multivariate normal distributions $\mathcal{D}_1, \mathcal{D}_2$ with common dimension k and respective covariances Σ_1, Σ_2 is

$$d(\mathcal{D}_1, \mathcal{D}_2) = \mathrm{tr}\Sigma_1 + \mathrm{tr}\Sigma_2 - 2\mathrm{tr}\left[(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right]. \tag{9}$$

In terms of Kullback-Leibler divergence the same separation is

$$d(\mathcal{D}_1, \mathcal{D}_2) = \frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_2|} - \frac{k}{2} + \mathrm{tr}\left[\Sigma_1^{-1}\Sigma_2\right], \tag{10}$$

and in terms of Hellinger distance it is

$$d(\mathcal{D}_1, \mathcal{D}_2) = \sqrt{1 - \sqrt{\frac{|\Sigma_1|^{1/4}|\Sigma_2|^{1/4}}{|\frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2|^{1/2}}}}. \tag{11}$$

Figure 7 shows by the measures (9), (10), and (11) the distances $d(\mathcal{M}, \mathcal{T})$ and $d(\mathcal{Z}, \mathcal{T})$ for the bivariate normal examples in Subsect. 3.3 with elongation $\rho = 2.5$ and rotations in the interval $\gamma \in [0°, 90°]$. In this class of examples with any of

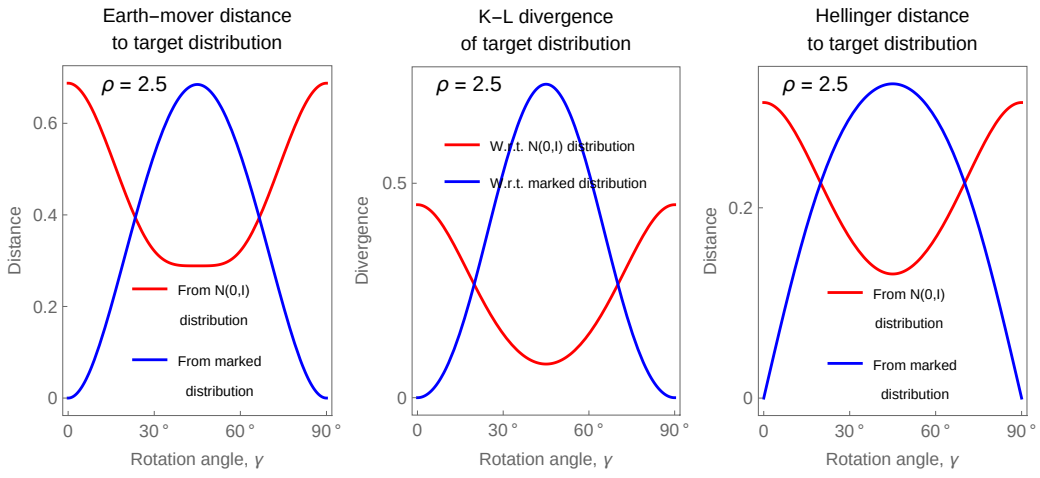

Figure 7: Distances $d(\mathcal{M}, \mathcal{T})$ (blue curves) and $d(\mathcal{Z}, \mathcal{T})$ (red curves) from the marked distribution $\mathcal{M}$ and a bivariate standard normal distribution $\mathcal{Z}$ to the target masked distribution $\mathcal{T}$ for the class of bivariate normal examples with elongation $\rho = 2.5$ in Subsect. 3.2. The same generic qualitative differences in the separations are seen with any of EM distance, K-L divergence, or Hellinger distance.

these measures of separation, we see generically for small rotations γ that $\mathcal{M}$ is closer to $\mathcal{T}$, while for bigger rotations approaching 45°, $\mathcal{Z}$ is closer. There is an intuition for this. When γ is small, $\mathcal{M}$ is closer to $\mathcal{T}$ because $\mathcal{M}$ and $\mathcal{T}$ have the same marginal distributions; on the other hand, $\mathcal{Z}$ and $\mathcal{T}$ both have independent marginal distributions, and that shared feature comes to matter more when the large rotation γ distorts the marginal distributions. Of course, in the end what matters for solving problem (8) is the estimation error $d(\hat{\mathcal{T}}, \mathcal{T})$ and how that depends on the starting points $\mathcal{M}$ and $\mathcal{Z}$ as γ varies. Experiments are underway to see whether the estimation error associated with $\hat{\mathcal{T}}$ behaves in accordance with the intuition offered here.

4.2 Optimal transport approach

The theory of optimal transport began with Monge [19] in 1781 and was given a probabilistic reformulation by Kantarovich [20] in 1942. Villani and Peyre survey modern developments in the theory in [21, 22, 23, 24]. This theory's key point is that while there are many ways to transform one probability distribution into another, there is a unique map that does so optimally in the sense of earth-mover distance. Significantly, powerful numerical algorithms are available [23, 25, 26] to estimate the optimal transport map $\mathbf{T} : \mathcal{D}_1 \to \mathcal{D}_2$ between two distributions $\mathcal{D}_1$, $\mathcal{D}_2$ from samples $\mathcal{S}_{\mathcal{D}_1}$, $\mathcal{S}_{\mathcal{D}_2}$ from those distributions.

Our optimal transport approach to problem (8) depends on estimating the optimal transport map $\mathbf{T}$ from $\mathcal{M}$ to $\mathcal{T}$. The specific steps in this approach are shown in Fig. 8. The blue panel in Fig. 8 shows the optimal map $\mathbf{T}$ that exists from $\mathcal{M}$ to $\mathcal{T}$.

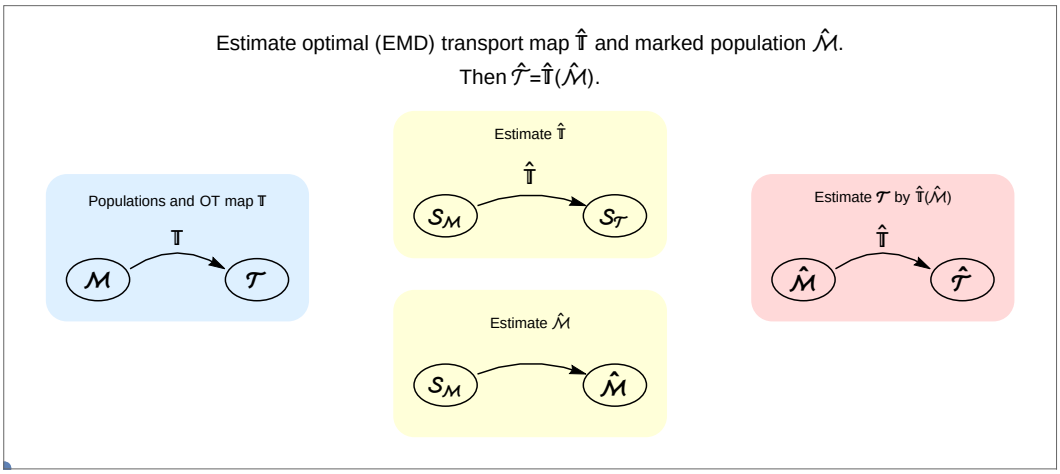

Figure 8: An optimal transport approach to the population masking problem. According to this approach an optimal transport map $\mathbf{T}$ exists (blue panel) that maps the marked population $\mathcal{M}$ to the target population $\mathcal{T}$. Both $\mathbf{T}$ and $\mathcal{M}$ are estimated (yellow panel) from the two available data samples $\mathcal{S}_{\mathcal{M}}$ and $\mathcal{S}_{\mathcal{T}}$. These estimates $\hat{\mathbf{T}}$ and $\hat{\mathcal{M}}$ are used to estimate (red panel) the target population distribution $\mathcal{T}$.

Both $\mathbf{T}$ and $\mathcal{M}$ are estimated (yellow panel) from the two available data samples $\mathcal{S}_{\mathcal{M}}$ and $\mathcal{S}_{\mathcal{T}}$. These estimates $\hat{\mathbf{T}}$ and $\hat{\mathcal{M}}$ are used to form an estimate $\hat{\mathcal{T}} = \hat{\mathbf{T}}(\hat{\mathcal{M}})$ (red panel) of the target population distribution $\mathcal{T}$, solving the population masking problem in (8).

4.3 Transfer learning approach

Transfer learning is a procedure in which a model developed for one task is used as the starting point for training a model with the same architecture on a second task [27, 28]. More specifically, a base model is trained on a base training sample and task, and the learned features (base model parameter estimates) are repurposed, or transferred, to a second target model. These transferred features are used by the second model as starting values for training on a different training sample and target task. This process will tend to aid the second model on its target task to the degree that features learned by the base model are general to both the base and the target tasks, rather than specific to just the base task.

In our application of transfer learning to the population masking problem in (8), the base task is set to be a generative model for $\mathcal{M}$ trained with $\mathcal{S}_{\mathcal{M}}$ and the target task is set to be a generative model for $\mathcal{T}$ trained with $\mathcal{S}_{\mathcal{T}}$. In Fig. 9 the base and target tasks for this learning transfer are shown (blue and yellow, resp.) with GANs as the generative models. In this application of transfer learning, we have only the limited data in $\mathcal{S}_{\mathcal{T}}$ to train GAN_2. The limited size of $\mathcal{S}_{\mathcal{T}}$ may require a nuanced approach to training GAN_2 in which not all the weights in GAN_2 are updated [29]. For example, it may be more effective to fix the weights in the earlier layers of GAN_2 to be those from GAN_1 and only use $\mathcal{S}_{\mathcal{T}}$ to update the weights in GAN_2's later layers.

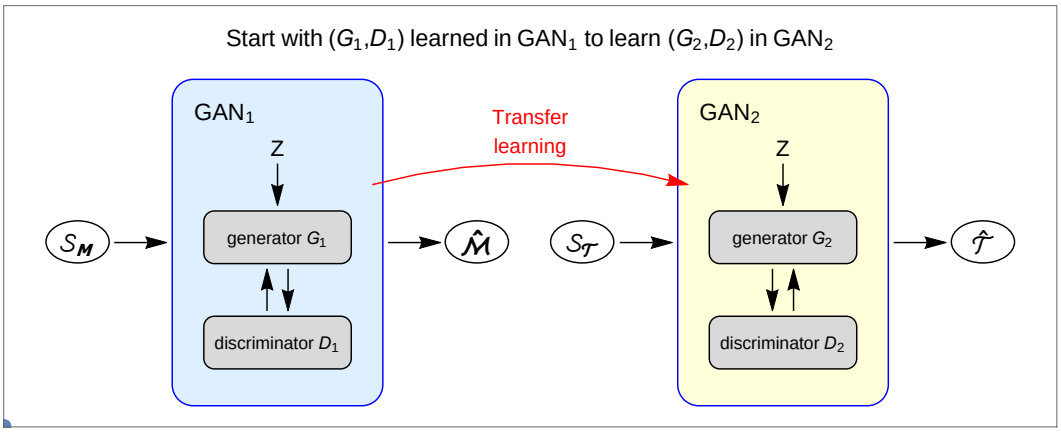

Figure 9: A transfer learning approach to the population masking problem. This approach involves two GANs, GAN_1 and GAN_2. The sample $\mathcal{S}_{\mathcal{M}}$ from the marked population $\mathcal{M}$ is used to train GAN_1 (blue) to approximate $\mathcal{M}$. The discriminator and generator parameters of GAN_1 learned in this process are transferred to GAN_2 (yellow) and used there as initial values for training GAN_2 with the target data sample $\mathcal{S}_{\mathcal{T}}$.

4.4 Hybrid approaches

The three just-described approaches to the population masking problem in (8) can be combined in different ways to create hybrid approaches. For example, the masked population distribution $\mathcal{T}$ estimated by $\hat{\mathcal{T}} = \hat{\mathbf{T}}(\hat{\mathcal{M}})$ in the optimal transport approach could be used instead of $\hat{\mathcal{M}}$ as a data generator in the data augmentation approach to drive the GAN. This hybrid approach is diagrammed in Fig. 10. This approach is interesting because it uses the data in $\mathcal{S}_{\mathcal{T}}$ twice: first, to estimate the optimal transport map $\mathbf{T}$ and then, second, as training data for the data augmentation approach's GAN. Experiments will show whether this hybrid can out-perform the data augmentation and optimal transport approaches from which it is constituted.

5 Inherent masking

A marked population $(\vec{M}, \mathcal{M})$ with marking map g is inherently masked if its sets of marked variables $\vec{X} = g_x(\vec{M})$ and free variables $\vec{Y} = g_y(\vec{M})$ happen to be *ab initio* stochastically independent. We give a simple example of inherent masking in a bivariate population and then generalize the example to the interesting class of univariate frequency-marked moving-average time series.

Bivariate example: Suppose $\mathcal{M}$ is defined by the random vector $\vec{M} = \mathbf{U}\vec{N}$ where $\vec{N} \sim \mathrm{N}(\vec{0}, \mathbf{I})$ is a two-component standard Gaussian noise and

$$\mathbf{U} = \frac{1}{1-x^2}\begin{pmatrix} 1 & x \\ x & 1 \end{pmatrix}, \tag{12}$$

where $x \in [-1, 1]$. The components of $\vec{M}$ can be considered a two-observation moving average time series in which each observation M_k is the corresponding latent

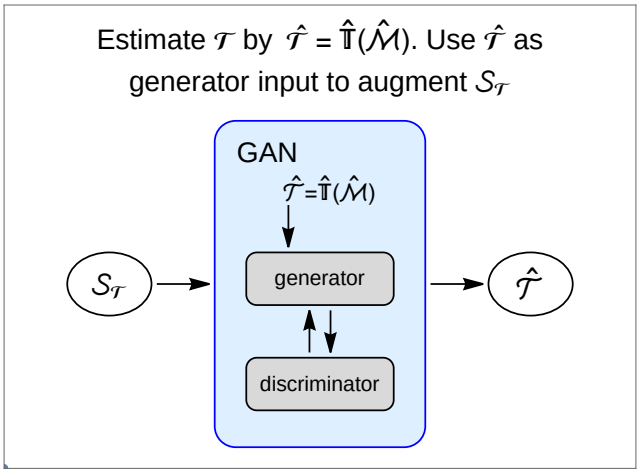

Figure 10: A hybrid approach to problem (8) that finds $\hat{\mathcal{T}} = \hat{\mathbf{T}}(\hat{\mathcal{M}})$ by our optimal transport approach, while training a GAN with $\mathcal{S}_{\mathcal{T}}$ and using $\hat{\mathcal{T}}$ to drive the GAN generator.

variable N_k plus/minus some fraction x of the "preceding" latent variable [30]. This time series $\vec{M}$ is distributed $\vec{M} \sim \mathrm{N}(\vec{0}, \Sigma_{\mathcal{M}})$ with unit-determinant covariance

$$\Sigma_{\mathcal{M}} = \mathbf{U}\mathbf{I}\mathbf{U}^{\mathsf{T}} = \frac{1}{(1-x^2)^2}\begin{pmatrix} 1+x^2 & 2x \\ 2x & 1+x^2 \end{pmatrix}. \tag{13}$$

Suppose that the marked information in $\mathcal{M}$ is contained in one of its Fourier frequencies. Then the marking map g is linear, defined by $g(\vec{M}) = \mathbf{F}\vec{M}$, where $\mathbf{F}$ is the discrete Fourier transform (DFT) matrix. The DFT matrix $\mathbf{F}$ for an N-component time series is $\mathbf{F} = [\omega^{(j-1)(k-1)}]$ for $j, k = 1, ..., N$ with $\omega = \exp(-2\pi i/N)$. In the present two-dimensional example,

$$\mathbf{F} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \tag{14}$$

and $\vec{S} = \mathbf{F}\vec{M}$ is normally distributed $\vec{S} \sim \mathrm{N}(\vec{0}, \Sigma_{\mathcal{S}})$ with covariance

$$\Sigma_{\mathcal{S}} = \mathbf{F}\Sigma_{\mathcal{M}}\mathbf{F}^{\mathsf{H}} = \frac{1}{2}\begin{pmatrix} (1-x)^2 & 0 \\ 0 & (1+x)^2 \end{pmatrix}. \tag{15}$$

The two components of $\vec{S}$ are the Fourier frequencies of the time series $\vec{M}$. The off-diagonal entries of $\Sigma_{\mathcal{S}}$ are zero, so the Fourier frequencies are independent, and the frequency-marked information in $\mathcal{M}$ is inherently obfuscated; no action is required to mask whichever frequency is identified as the marked population variable.

Moving average time series: The preceding example generalizes to Gaussian moving-average time series of any length and parametric structure. We now show that, for any univariate time series in this class, marked information composed of *any* subset of Fourier frequencies is inherently obfuscated. Let $\vec{M}$ be a univariate N-observation Gaussian moving-average time series [30]. Then $\vec{M} = \mathbf{R}\vec{N}$, where $\vec{N} \sim \mathrm{N}(\vec{0}, \mathbf{I})$ is an N-component standard Gaussian noise, and $\mathbf{R}$ is the matrix

$$\mathbf{R} = a_0\mathbf{I} + \sum_{k=1}^{N-1} a_k\mathbf{Q}^k, \tag{16}$$

where $\mathbf{Q}$ is the cyclic shift matrix

$$\mathbf{Q} = \begin{pmatrix} \vec{0}^{\mathsf{T}} & 1 \\ \mathbf{I} & \vec{0} \end{pmatrix}. \tag{17}$$

The zero vector $\vec{0}$ and the identity matrix $\mathbf{I}$ in (17) are $(N-1)$-dimensional and $(N-1)\times(N-1)$-dimensional, respectively. Let the marked population $\mathcal{M}$ be the time series represented by $\vec{M} = \mathbf{R}\vec{N}$, and suppose $\mathcal{M}$ is frequency-marked with marking map defined by $\vec{S} = \mathbf{F}\vec{M}$ where $\mathbf{F}$ is the N-point DFT matrix. Then $\vec{S} = \mathbf{F}\vec{M}$ is normally distributed $\vec{S} \sim \mathrm{N}(\vec{0}, \Sigma_{\mathcal{S}})$ with covariance $\Sigma_{\mathcal{S}} = \mathbf{F}\Sigma_{\mathcal{M}}\mathbf{F}^{\mathsf{H}}$. Denote the rows of $\mathbf{F}$ by F_k. These rows are orthogonal, $F_iF_j^{\mathsf{H}} = 0$ for $i \neq j$. We have as well $(F_i\mathbf{Q})(F_j\mathbf{Q})^{\mathsf{H}} = 0$ for $i \neq j$ since $\mathbf{Q}$ cyclically shifts the entries of both F_i and F_j in tandem. We have, too, the identity $(F_i\mathbf{Q}^k)(F_j)^{\mathsf{H}} = 0$ for $i \neq j$ and any

$k = 1, ..., N-1$ [31, 32]. Therefore,

$$\begin{aligned}(F_i\mathbf{R})(F_j\mathbf{R})^{\mathsf{H}} &= F_i\mathbf{R}\mathbf{R}^{\mathsf{T}}F_j^{\mathsf{H}} \\ &= F_i\left(\sum_{k=0}^{N-1} a_k\mathbf{Q}^k\right)\left(\sum_{m=0}^{N-1} a_m(\mathbf{Q}^m)^{\mathsf{T}}\right)F_j^{\mathsf{H}} \\ &= \sum_{k=0}^{N-1}\sum_{m=0}^{N-1} a_k a_m F_i\mathbf{Q}^k(\mathbf{Q}^m)^{\mathsf{T}}F_j^{\mathsf{H}} \\ &= \sum_{k=0}^{N-1}\sum_{m=0}^{N-1} a_k a_m (F_i\mathbf{Q}^k)(F_j\mathbf{Q}^m)^{\mathsf{H}}.\end{aligned}$$

This shows that $(F_i\mathbf{R})(F_j\mathbf{R})^{\mathsf{H}} = 0$ for $i \neq j$, from which it follows that

$$\Sigma_{\mathcal{S}} = \mathbf{F}\Sigma_{\mathcal{M}}\mathbf{F}^{\mathsf{H}} = \mathbf{F}\mathbf{R}(\mathbf{F}\mathbf{R})^{\mathsf{H}} \tag{18}$$

is diagonal. All of the off-diagonal entries in $\Sigma_{\mathcal{S}}$ are zero, meaning that the Fourier frequencies of $\vec{M}$ are mutually independent; in particular, the marked variables in $\mathcal{M}$ composed of any subset of frequencies is independent of the remaining shareable frequencies, and $\mathcal{M}$ marked in this way is inherently obfuscated.

6 Summary and final remarks

This work extends the notion of sample obfuscation to population obfuscation. In the infinite-size populations studied here, population units and entries are naturally secured by the population's size, so the security of a population's variables is of most interest. Masking is a prominent form of obfuscation for sample variables, and we made population masking this work's primary focus.

Our extension of sample masking to population masking led us specifically to consider population masking as a task, with (at least) three different possible approaches to its accomplishment. We have experiments underway to discover each of these approaches' potential in terms of data sample sizes $N_{\mathcal{T}}$ and $N_{\mathcal{M}}$ and the estimation error $d(\hat{\mathcal{T}}, \mathcal{T})$ that results. An important metrological question is the appropriate metric $d(\cdot, \cdot)$ for representing error in the setting of population masking. Potential metrical candidates are the Hellinger, p-Wasserstein, total variation, and Bhattacharya distances [33] and the Kullback-Leibler divergence. Experiments designed to compare the suitabilities of these measures are currently underway.

We noted at the end of Subsect. 3.2 that population masking does not asymmetrically privilege marked variables over free variables; the labeling is arbitrary and "marked" and "free" are interchangeable with no effect. Therefore, the ideas of population masking extend without modification beyond separating two groups of population variables to separating any number of groups. For example, in the last section's example of an N-observation moving average time series, we might be tasked to jointly mask all N of the Fourier frequencies from one another. We saw in that example that the Fourier frequencies are jointly independent, so we would find that that time series population is inherently masked even then.

References

[1] De Capitani Di Vimercati, S., Foresti, S., Livraga, G., & Samarati, P. (2012). Data privacy: definitions and techniques. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **20**(6), 793–817.

[2] Raghunathan, B. (2013). The complete book of data anonymization: from planning to implementation. CRC Press.

[3] Wasserman, L., & Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, **105**(489), 375–389.

[4] Dwork, C. (2008, April). Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation* 1–19. Springer, Berlin, Heidelberg.

[5] Duncan, G., & Stokes, L. (2009). Data masking for disclosure limitation. *Wiley Interdisciplinary Reviews: Computational Statistics*, bf 1(1), 83–92.

[6] Gerban, M. (2019). Why Distinctions Within Mobile Wallets and Tokenization Matter. *The PayTech Book: The Payment Technology Handbook for Investors, Entrepreneurs and FinTech Visionaries*, 59–60.

[7] Adam, N. R., & Worthmann, J. C. (1989). Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys (CSUR)*, **21**(4), 515–556.

[8] Moore, R. A. (1996). Controlled data swapping for masking public use microdata sets. US Census Bureau Research Report, **96**(04).

[9] Goyal, C. (2015). Data Masking: Need, Techniques Solutions. *International Research Journal of Management Science & Technology (IRJMST)*, **6**(5), 221–229.

[10] Sarada, G., Abitha, N., Manikandan, G., & Sairam, N. (2015, March). A few new approaches for data masking. In *2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015]*, 1–4. IEEE.

[11] Goodfellow I., Pouget-Abadi J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., & Bengio Y. (2018). Generative adversarial networks. *Advances in Neural Information Processing Systems*. Eds. Jordan M.I., LeCun Y., Solla S.A. (MIT Press, Cambridge, MA), 2672-–2680.

[12] Creswell A., White T., Dumoulin V., Arulkumaran K., Sengupta B., & Bharath A. (2018). Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine* **35**(1), 53-–65.

[13] Taylor, L., Nitschke, G. (2018, November). Improving deep learning with generic data augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)* (1542–1547). IEEE.

[14] Ioffe, S. & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). PMLR.

[15] Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, **25**.

[16] Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2020, April). Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence* **34**(7), 13001–13008.

[17] Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[18] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, **6**(1), 1–48.

[19] Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*. 666–704.

[20] Kantorovich, L.V. (1942). On the translocation of masses. In *Dokl. Akad. Nauk.* USSR (NS) (Vol. 37, pp. 199–201).

[21] Villani, C. (2003). *Topics in optimal transportation*, **58**. American Mathematical Society.

[22] Villani, C. (2009). *Optimal transport: old and new* (Vol. 338, p. 23). Berlin: Springer.

[23] Peyré, G. & Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, **11**(5-6), 355–607.

[24] Thorpe, M. (2019). *Introduction to Optimal Transport*. Lecture Notes.

[25] Lévy, B. & Schwindt, E.L. (2018). Notions of optimal transport theory and how to implement them on a computer. *Computers & Graphics*, **72**, 135–148.

[26] Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., ... & Vayer, T. (2021). POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78), 1–8.

[27] Pan, S.J., & Yang, Q. (2009). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, **22**(10), 1345–1359.

[28] Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, **3**(1), 1–40.

[29] https://cs231n.github.io/transfer-learning/

[30] Percival, D. B., & Walden, A. T. (2020). *Spectral analysis for univariate time series (Vol. 51)*. Cambridge University Press.

[31] Kra, I., & Simanca, S.R. (2012). On circulant matrices. *Notices of the AMS*, **59**(3), 368–377.

[32] Gray, R.M. (2006). Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, **2**(3), 155—239.

[33] Panaretos, V.M., & Zemel, Y. (2018). Statistical aspects of Wasserstein distances. arXiv preprint arXiv:1806.05500.