Some notes on the Multivariate Binary distribution in the Grassmann formalism

Cornelius Schöder

July 14, 2022

Abstract

A binary distribution can be seen as the simplest case of a probability distribution, which can only take two values, commonly noted as 0 and 1. However, working with *multivariate* binary distributions poses several difficulties and there is no 'natural' 'normal' distribution for discrete probability distributions. Nowadays, neural networks, for example auto-regressive flows, can be used to approximate multivariate binary distributions, with the caveat that often not all properties, such as marginals or conditional distributions can be derived easily. In these notes we give some comments on binary distributions in the Grassmann formalism [Ara21]. In this framework a lot of properties can be calculated analytically. Additionally, we define a 'mixture of Grassmann' distribution and show how this can be used in a mixture of density network to model flexible conditional distributions.

1 Introduction

The Grassmann formalism for multivariate binary distributions was introduced by Arai [Ara21] and gives an elegant way to define an n-dimensional binary distribution by an $n \times n$ matrix Σ. For a valid distribution $\Sigma^{-1} - I$ must be a P_0 matrix, but has otherwise no further constraints. This definition gives access to the analytical derivations of properties such as the mean, covariance but also marginal and conditional distributions. We summarize some of the analytical formulas in the Appendix, a detailed derivation as well as further details can be found in the aforementioned publication.

In theses notes we will first define a mixture of Grassmann distribution (MoGr) before we will comment on the expressivity of binary distributions in the Grassmann formalism in terms of possible covariances. Next we will give a proof of principle for a conditional MoGr distribution on a low dimensional toy dataset. In the last section we will then comment on the implemenation, which can be found at `https://github.com/mackelab/grassmann_binary_distribution`.

2 Mixture Distribution

We can define a mixture of Grassmann distribution on $\{0,1\}^n$ in the same formalism as the Grassmann distribution $\mathcal{G}$ by defining $p(x) = \sum_i \alpha_i \mathcal{G}_i(x|\Sigma_i)$ for a finite partition $\sum_i \alpha_i = 1$. For each component $\mathcal{G}_i$ we can calculate the mean μ_i and covariance C_i analytically. To calculate these quantities for the mixture distribution we introduce a discrete latent variable Z and reformulate the mixture distribution as

$$\begin{aligned} p(x|Z=i) &= \mathcal{G}_i(x|\Sigma_i), \\ p(Z=i) &= \alpha_i. \end{aligned}$$

We can easily compute the expected value and covariance by using the law of total expectation and variance. Therefore we get

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Z=i]] = \sum_i \alpha_i \mu_i$$

and

$$\begin{aligned} \mathrm{Cov}(X) &= \mathbb{E}[\mathrm{Cov}(X|Z=i)] + \mathrm{Cov}(\mathbb{E}[X|Z=i]) \\ &= \sum_i \alpha_i C_i + \sum_i \alpha_i (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^T, \end{aligned}$$

where $\bar{\mu} = \mathbb{E}[X]$.

3 On the Expressivity

3.1 On the Convariance

Multivariate binary distributions have the interesting property that, depending on their mean, the covariances can only take restricted values. For a two dimensional distribution we can derive the bounds analytically: For any binary random variables X_1 and X_2 with $p(X_1 = 1) = p_1$ and $p(X_2 = 1) = p_2$ their covariance is first restricted by the Cauchy-Schwarz-inequality

$$|\operatorname{Cov}(X_1, X_2)| \leq \sqrt{\operatorname{Var}(X_1)\operatorname{Var}(X_2)} = \sqrt{p_1(1-p_1)p_2(1-p_2)} \leq \frac{1}{4}.$$

Using the Grassmann formalism adds further constraints on the covariance via the P_0 criterion for $\Sigma^{-1} - I$, which we can calculate explicitly. For a two dimensional distribution with $\Sigma = \Lambda^{-1} = \begin{pmatrix} p_1 & b \\ c & p_2 \end{pmatrix}$ and using $(\Lambda_{11} - 1), (\Lambda_{22} - 1) \geq 0$ and $\det(\Lambda - I) \geq 0$ this yields the following criteria:

$$\begin{aligned}
&\operatorname{Cov}(X_1, X_2) \leq p_2(1-p_1), \\
&\operatorname{Cov}(X_1, X_2) \leq p_1(1-p_2), \\
&\operatorname{Cov}(X_1, X_2) \geq -(1-p_1)(1-p_2), \quad \text{and} \\
&\operatorname{Cov}(X_1, X_2) \geq -p_1 p_2.
\end{aligned}$$

These are the exact same constraints as for *any* two dimensional binary distribution [MBE+09] and can be summarized as:

$$\begin{aligned}
&\operatorname{Cov}(X_1, X_2) \leq \min(p_1(1-p_2), p_2(1-p_1)), \quad \text{and} \\
&\operatorname{Cov}(X_1, X_2) \geq \max(-p_1 p_2, -(1-p_1)(1-p_2)).
\end{aligned}$$

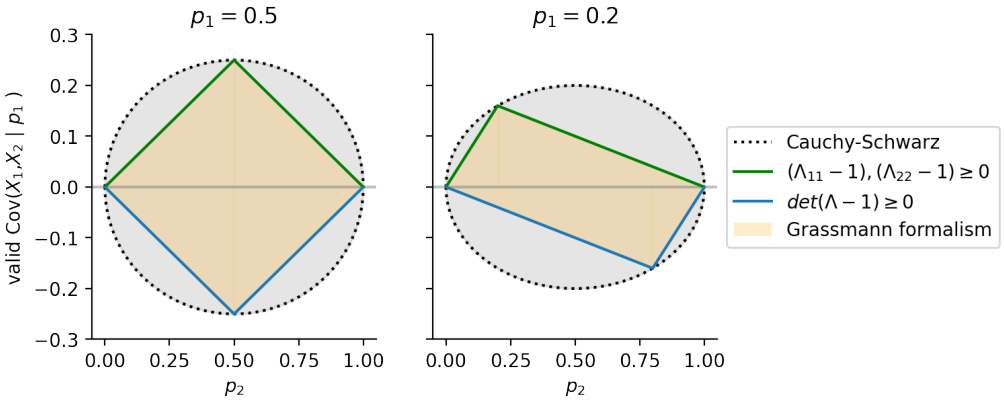

Figure 1: The covariance of binary variables is bounded by the Cauchy-Schwarz inequality. Using the Grassmann formalism adds constraints from the P_0 criterion, here shown for two examples of two dimensional distributions. These constraints coincide in the two dimensional case with the constraints for *any* binary distribution.

We can visualize these constraints in Figure 1. It is obvious that the constraints on the pairwise covariances is not sufficient for the P_0 criterion. We can easily verify this using the same example as in [MBE+09] with $p_1 = p_2 = p_3 = \frac{1}{2}$ and the covariance matrix

$$C = \begin{pmatrix} \frac{1}{4} & -\frac{1}{8} & -\frac{1}{8} \\ -\frac{1}{8} & \frac{1}{4} & -\frac{1}{8} \\ -\frac{1}{8} & -\frac{1}{8} & \frac{1}{4} \end{pmatrix},$$

which gives us a possible parametrization in the Grassmann formalism

$$\Sigma = \begin{pmatrix} \frac{1}{2} & \sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} \\ \sqrt{\frac{1}{8}} & \frac{1}{2} & \sqrt{\frac{1}{8}} \\ \sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & \frac{1}{2} \end{pmatrix}.$$

We can check that $\det(\Sigma^{-1} - I) < 0$ and Σ does therefor not fulfill the P_0 criterion. In fact, as it is shown in [MBE+09], there exists no binary distribution with C as covariance matrix and so the exact choice of Σ doesn't matter in this case.

3.2 Comparison to Dichotomized Gaussian

To get a first intuition on the expressivity of a multivariate binary distribution in the Grassmann formalism we compared it to the dichotomized Gaussian distribution [MBE+09]. For this we fixed the mean of all and the covariance of the first two dimensions and sampled the remaining parameters to get three dimensional binary distributions. For the Grassmann distribution we rejected the sampled parameters if they did not fulfill the P_0 criteria. For the DG non valid samples were projected via the Higham projection to a positive semi-definite matrix with minimal Frobenius distance to the initial sample.
Figure 2 clearly shows the projection boundaries for the DG distribution, whereas the boundaries for the Grassmann distribution are more fuzzy. However we observe that there are regions in the second quadrant for which it is able to parameterize a DG but not a Grassmann distribution. Vice versa is true for regions in the forth quadrant. Fitting a mixture of Grassmann distribution allows for more extreme covariances and to cover regions outside the boundaries imposed by the P_0 criteria, shown by one example point in Figure 2. This hints to a higher expressivity of the mixture of Grassmann distribution compared to a single Grassmann distribution and also DG distributions. However a more systematic and theoretical investigation is an open research project.

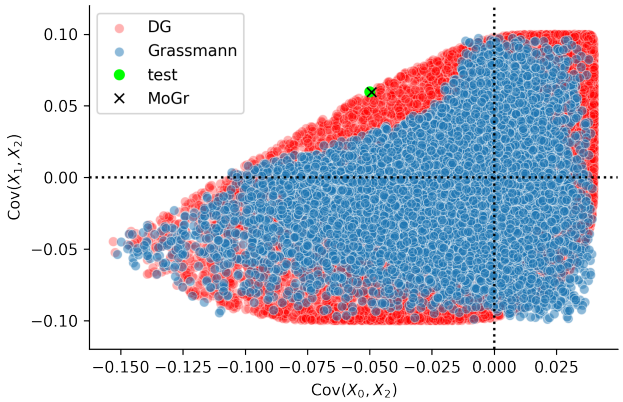

Figure 2: Example covariances for randomly sampled parameters for a dichotomized Gaussian (DG) and a Grassmann distribution with fixed mean, $\mu = (0.8, 0.5, 0.2)$, and $\mathrm{Cov}_{01} = 0.05$. The test sample shows $(\mathrm{Cov}_{02}, \mathrm{Cov}_{12}) = (-0.05, 0.06)$ together with a fitted mixture of Grassmann (MoGr) distribution with 3 components.

While this comparison focuses on the first and second moments, we want to highlight that even if we fix these moments, the Grassmann formalism still has degrees of freedom as $\mathrm{Cov}[X_i, X_j] = -\Sigma_{ij}\Sigma_{ji}$ while the DG is fully specified by the first and second moment.

4 Conditional Distributions

In the same spirit as a conditional mixture of Gaussian distribution we can implement a conditional mixture of Grassmann distribution. We use a hidden network h to map some input vector y via the parametrization described in section 5.1 to $\mathcal{G}(\cdot|y, \Sigma)$:

$$y \xrightarrow{h} \text{latent} \xrightarrow{linear} B, C|y \xrightarrow{\text{see 5.1}} \Sigma|y \longrightarrow \mathcal{G}(\cdot|y, \Sigma).$$

We show the a proof of concept by fitting a mixture of Grassmann distribution to samples from a three dimensional dichotomized Gaussian (DG) distribution [MBE+09]. For this we sample randomly the parameters $\mu^{\mathcal{N}}$ and $\Sigma^{\mathcal{N}}$ for the DG distribution, draw one sample x from the specified DG and define the input y as the flattened and concatenated vector $\mu^{\mathcal{N}}$ and the upper diagonal half of $\Sigma^{\mathcal{N}}$. We then train the network to maximize $\mathcal{G}(x|y, \Sigma)$ end to end. See Appendix 6.2 for details.
Given the simple task of predicting a corresponding Σ from the DG mean and covariance it is not surprising that we can capture the first and second moments quite well (Fig. 3). For more difficult tasks the expressivity and training properties need still to be tested.

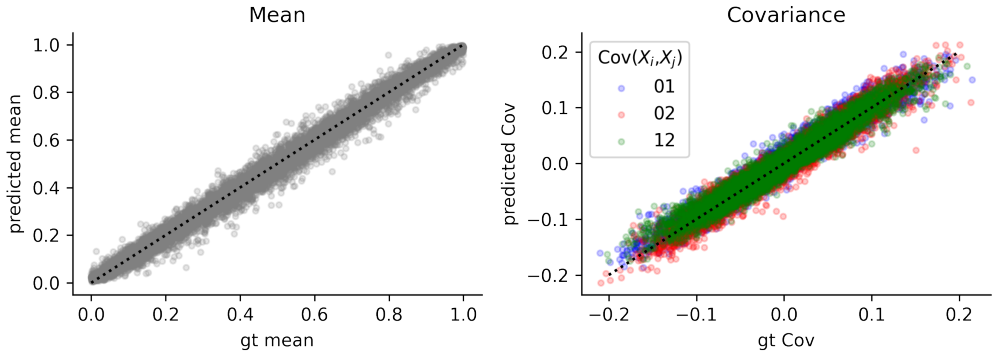

Figure 3: Results for the fitted conditional distribution. We plot the flattened mean and covariance of the ground truth (gt) vs. the predicted mean. For the covariance the colorcode indicates covariance between which dimensions.

5 Implementation

5.1 Parametrization

In [Ara21] Arai proposes the following parametrization for Σ that ensures the P_0 criterion:

$$\Sigma^{-1} = BC^{-1} + I,$$

where B and C are strictly row diagonal dominant matrices:

$$b_{ii} > \sum_{j \neq i} |b_{ij}|,$$
$$c_{ii} > \sum_{j \neq i} |c_{ij}|.$$

We use this to parametrization to optimize unconstraint matrices $\tilde{B}$ and $\tilde{C}$ and define B by replacing the diagonal elements of $\tilde{B}$ by

$$b_{ii} = \exp(\tilde{b}_{ii}) + \sum_{j \neq i} |\tilde{b}_{ij}|$$

and analogously for C. Instead of exp any other positive function could be chosen and even the non-negative Relu function showed good training properties in initial experiments.
For a mixture of Grassmann distribution we used a similar paramterization for each component and in the fitting procedure a softmax layer to learn the partition $\sum_i \alpha_i = 1$.

5.2 Moment Matching

As the P_0 criterion gives us a sufficient criterion for valid Σ, we can easily check for a given mean μ and covariance matrix C if the 'almost-symmetric' matrix Σ with $\Sigma_{ii} = \mu_i$ and $\Sigma_{ij} = -\sqrt{|C_{ij}|}$, $\Sigma_{ji} = \text{sign}(C_{ij})\sqrt{|C_{ij}|}$ yields a valid multivariate binary distribution in the Grassmann formalism. This is implemented in the class `EstimateGrassmannMomentMatching`.
However, the Grassmann formalism is more flexible and in general distributions with fixed first and second moments but with $|\Sigma_{ij}| \neq |\Sigma_{ij}|$ result in different distributions.
Nevertheless initial experiments showed that an initialization with the moment matching parametrization can be beneficial to get the maximum likelihood estimate for Σ.

References

[Ara21] Takashi Arai. Multivariate binary probability distribution in the grassmann formalism. *Physical Review E*, 103(6):062104, 2021.

[MBE+09] Jakob H Macke, Philipp Berens, Alexander S Ecker, Andreas S Tolias, and Matthias Bethge. Generating spike trains with specified correlation coefficients. *Neural computation*, 21(2):397–423, 2009.

6 Appendix

6.1 Formulas

We summarize some of the most important formulas here, but refere to [Ara21] for further details. Let X be a multivariate binary random variable, $X \sim \mathcal{G}(\Sigma)$, $\Sigma \in \mathbb{R}^{n \times n}$, and $\Lambda = \Sigma^{-1}$. Let $x \in \{0,1\}^n$.

- $\mathcal{G}(x|\Sigma) = \det(\begin{pmatrix} \Sigma_{11}^{x_1}(1-\Sigma_{11})^{1-x_1} & \Sigma_{12}(-1)^{1-x_2} & \cdots \\ \Sigma_{21}(-1)^{1-x_1} & \Sigma_{22}^{x_2}(1-\Sigma_{22})^{1-x_2} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix})$
- $\mathbb{E}[X_i] = \Sigma_{ii}$
- $\mathrm{Cov}[X_i, X_j] = -\Sigma_{ij}\Sigma_{ji}$

6.2 Details for the Conditional Distribution

We use a fully connected network with three layers and ReLU activaton functions as hidden network h. For the training we use 100k DG distributions with one sample per DG and a batchsize of 10k/20k to minimize the negative log likelihood.