Extending Gaussian Probability with Uninformative Priors

Dario Stein

iHub, Radboud University Nijmegen, The Netherlands
dario.stein@ru.nl

Abstract. We introduce *extended Gaussian distributions* as a precise and principled way of combining Gaussian probability with uninformative priors, i.e. idealized uniform distributions which indicate complete absence of information. To give an extended Gaussian distribution on a finite-dimensional vector space X is to give a subspace D together with a Gaussian distribution on the quotient X/D. Our construction can be seen as an extension of linear relations (nondeterminism) with Gaussian noise. Our main result is that the class of extended Gaussians remains closed under taking conditional distributions, making them suitable for Bayesian inference.

Uninformative priors appear naturally when reasoning about probabilistic programs, and we use our construction to solve an outstanding characterization of contextual equivalence in a language for Gaussian probability.

Keywords: bayesian inference · category theory · probabilistic programming

1 Introduction

The *Gaussian* or *normal distribution* is truly ubiquitous throughout statistics, science and engineering. A random variable X is Gaussian with mean μ and variance σ^2 if it has probability density

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

with respect to the Lebesgue measure; we write this as $X \sim \mathcal{N}(\mu,\sigma^2)$. This is generalized to multivariate normal distributions on $\mathbb{R}^n$, which are characterized by a mean $\vec{\mu}$ and covariance matrix Σ.

A crucial feature of Gaussian distributions is that they are *self-conjugate*; conditional distributions of Gaussians are themselves Gaussian. This makes Gaussian probability a powerful setting for a variety of statistical models such as ridge regression, Gaussian processes or Kalman filters. To learn from data, we condition the output of these models on real-world observations, using Bayes' law to obtain an updated posterior. We demonstrate this using a simple noisy measurement example:

Let some unknown quantity X be distributed as $\mathcal{N}(50, 100)$; this represents our *prior knowledge*. We have access to a noisy measurement Y, which is centered around X but itself introduces a variance of 25. After observing a value of $Y = 40$, we obtain an updated *posterior* belief over X, namely[1]

$$X|(Y = 40) \quad \sim \quad \mathcal{N}(42, 20) \qquad (1)$$

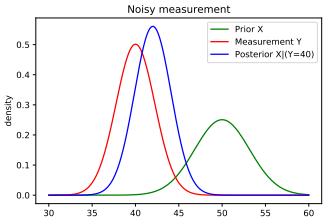


Fig. 1 – Noisy measurement

Notice that the posterior variance is strictly smaller than the prior or measurement variances, because the two sources of knowledge combine to produce a slightly more precise estimate.

In this paper, we address the problem of how to model complete *absence of information* in this framework:

1. Which distribution can we put on X if we have absolutely no prior knowledge about its value? The posterior in this case should precisely reflect the measurement, $\mathcal{N}(40, 25)$, without any modification.
2. Which distribution can we use to model a measurement Y that is completely inconclusive? Observing any particular value here will leave the prior unchanged, because we have gained no information.

We can approximate these properties by putting larger and larger variances $\sigma^2 \to \infty$ on the prior or measurement. However, this process does not converge to a limiting distribution in any meaningful sense. Intuitively, an uninformative prior should be completely flat (uniform) over the entire real line, but such a probability distribution does not exist. In practice, one can sometimes pretend (using the method of *improper priors*, e.g. [13]) that X is sampled from the Lebesgue measure λ (with constant density 1). This measure fails to be normalized, however the resulting density calculations may yield the correct result. We prefer to give a formal account of this situation which avoids unnormalized measures altogether:

Our novel insight is that in the context of Gaussian probability, we can use nondeterminism (relations) to act as uninformative priors. For example, we'll treat the *subset* $\mathbb{R} \subseteq \mathbb{R}$ as the uninformative probability distribution over the real line. Subsets (or relations) have a long tradition in computer science to model nondeterministic functions with multiple possible outputs. In general, it is not possible do combine probability and nondeterminism in a seamless way[2] [15, 31]. It is surprising that we can achieve this in the restricted context of linear algebra and Gaussian probability. In the following discussion, we will use the

[1] see appendix 6.1 for the example calculation

[2] the technical obstruction here is the non-commutativity of the resulting theory

terms nondeterministic, uninformative and uniform interchangeably.

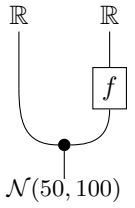
In Section 2, we introduce a class of synthetic distributions on $\mathbb{R}^n$ called *extended Gaussians*, which are a formal combination of two types of noise: nondeterministic noise along a vector subspace $D \subseteq \mathbb{R}^n$, and probabilistic noise distributed according to a multivariate normal distribution $\mathcal{N}(\vec{\mu}, \Sigma)$ with mean $\vec{\mu}$ and covariance matrix Σ. By synthetic distribution, we mean that these are not probability distributions in the usual sense of measure theory. Yet, we can manipulate them like probability distributions in a rigorous way. One way of formalizing this is through categorical probability theory:

1.1 Markov categories and Linear Relations

A *Markov category* [10] is a self-contained universe of probability distributions and stochastic maps. One such universe is *Gaussian probability*, by which we mean the Markov category Gauss whose objects are affine spaces $\mathbb{R}^n$, and whose maps are affine-linear maps with multivariate Gaussian noise, written

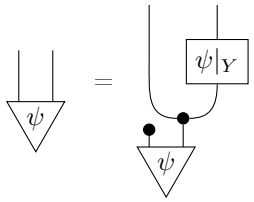
$$f(\vec{x}) = A\vec{x} + \mathcal{N}(\vec{b}, \Sigma) \tag{2}$$

This is a fairly restricted universe (all maps are affine-linear), yet is has good formal properties and we are able to express an array of interesting real-world models in it, as outlined earlier in the introduction. In Markov categories, we can make use of the diagrammatic language of string diagrams to compose morphisms in an intuitive and convenient way [10, 23]. For example, the joint distribution over the variables (X, Y) in the noisy measurement example can be described as the following morphism $\mathbb{R}^0 \to \mathbb{R}^2$ in Gauss,



where $f(x) = x + \mathcal{N}(0, 25)$

Bayesian inference, that is the computation of conditional probabilities, can be tricky to formalize, especially in cases where observations have probability zero [18], such as conditioning on the exact observation $Y = 40$. We can sidestep these issues using the abstract notion of conditional probability in Markov categories. This means we can recover a joint distribution $\psi : I \to X \otimes Y$ from its marginal $\psi_Y : I \to Y$ using a conditional map $\psi|_Y : Y \to X$ as follows

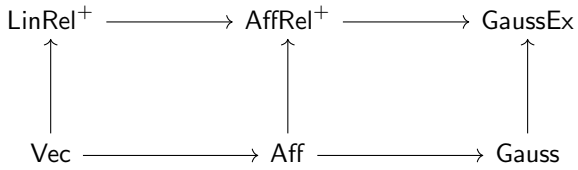


The category Gauss has all conditionals, which lets us reason about statistical inferences such as (1) purely abstractly. This forms the basis for the semantics of a dedicated probabilistic programming language for Gaussian probability with a first-class conditioning construct [28, 26]. We will come back to this language and exact conditioning in Section 4.2.

In Section 3, we extend the category Gauss to incorporate uninformative priors. The resulting category GaussEx consists of affine-linear maps with extended Gaussian noise. Its construction is interesting on its own right and emphasizes the connection with nondeterminism: A *linear relation* between vector spaces X, Y is a relation $R \subseteq X \times Y$ which is also a vector subspace. Linear relations have various applications in computer science and engineering [5, 2, 3]. Left-total linear relations form a Markov category LinRel^+; as we show in Proposition 5, we can think of such a relation as a linear map with *nondeterministic noise*

$$f(\vec{x}) = A\vec{x} + D$$

which is reminiscent of morphisms (2) in Gauss. We define GaussEx as an combination of LinRel^+ with Gaussian probability, using a general formalism of *decorated linear relations*. We obtain the following diagram of faithful inclusions between several Markov categories of interest,



The categories in question are

- Vec: vector spaces and linear functions
- Aff: vector spaces and affine-linear functions
- LinRel^+: vector spaces and left-total linear relations
- AffRel^+: vector spaces and left-total affine relations
- Gauss: vector spaces and Gaussian maps
- GaussEx: vector spaces and extended Gaussian maps

In Theorem 3, we prove that GaussEx has all conditionals, making it a suitable domain for Bayesian inference. The proof combines self-conjugacy of ordinary Gaussians with general structural facts about linear relations.

1.2 Contribution

We show that in the setting of Gaussian probability, one can represent uninformative distributions by nondeterminism. We use this to define a category of extended Gaussian maps which combines linear relations with Gaussian noise. Proving that nondeterminism and probability combine here in a well-defined and commutative way is nontrivial, and the main part of this work is dedicated to presenting this construction in a modular way using the formalism of decorated linear relations. Commutativity, meaning that independent computations can be reordered, is implicit in the use of monoidal categories and string diagrams throughout. In Section 2.2, we show that extended Gaussians admit a more symmetric relationship between covariance and precision than ordinary Gaussians.

We then prove that extended Gaussians remain closed under taking conditionals, which makes them suitable to apply to Bayesian inference. We discuss conditioning in detail in Section 4.1, and show that our relations do indeed satisfy the properties expected of uninformative distributions.

One central application of this work lies in the semantics of probabilistic programs: In Section 4.2, we consider a dedicated programming language for Gaussian distributions with an exact conditioning operator. Using extended Gaussians, we can extend the semantics of this language to include uniform distributions. This allows us to reduce questions about open terms to closed terms and thus helps solve an outstanding characterization of contextual equivalence for this language. In this way, extended Gaussians show up as tools even when reasoning about ordinary Gaussian distributions.

2 Extended Gaussian Distributions

Before we give the formal definition of extended Gaussians in Section 2.1, we begin with a review of Gaussian probability and an informal overview to build intuitions. We have compiled a summary of the required linear algebra terminology in the appendix (Section 6.2).

Recall that a *Gaussian distribution* on $\mathbb{R}^n$ can be written uniquely as $\mathcal{N}(\mu, \Sigma)$ where $\mu \in \mathbb{R}^n$ is its *mean* and $\Sigma \in \mathbb{R}^{n\times n}$ is a symmetric positive-semidefinite matrix called its *covariance matrix*. The *support* of $\mathcal{N}(\mu, \Sigma)$ is the affine subspace $\mu + \mathsf{col}(\Sigma)$ where $\mathsf{col}(\Sigma)$ is the column space (image) of Σ. Gaussian distributions transform as follows under affine-linear maps: If $f(x) = Ax + b$ with $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$ then

$$f_*\mathcal{N}(\mu, \Sigma) = \mathcal{N}(f(\mu), A\Sigma A^T)$$

is its pushforward Gaussian distribution on $\mathbb{R}^m$. We write addition $+$ between distributions to indicate the distribution of the sum of two independent variables; for example, if $X, Y \sim \mathcal{N}(0, 1)$ are independent, then $X + Y \sim \mathcal{N}(0, 2)$ because

variance is additive for independent variables. We therefore have the following equation between distributions

$$\mathcal{N}(0,1) + \mathcal{N}(0,1) = \mathcal{N}(0,2)$$

For an introduction to Gaussian probability see e.g. [29].

We now wish to combine Gaussian distributions with uninformative (nondeterministic) distributions along a vector subspace D. An *extended Gaussian distribution on* $\mathbb{R}^n$ is a *formal sum*

$$\mathcal{N}(\mu, \Sigma) + D \tag{3}$$

of a Gaussian distribution (probabilistic part) and a subspace $D \subseteq \mathbb{R}^n$ (nondeterministic part). We will refer to the space D as the *locus of nondeterminism*. For example, the expression $\mathcal{N}(0,1) + \mathbb{R}$ represents a sum of a standard normal variable and an uninformative variable over the real line.

We would like the outcome of this to be simply uninformative, i.e.

$$(\mathcal{N}(0,1) + \mathbb{R}) = \mathbb{R} \tag{4}$$

This is because the subspace $\mathbb{R}$ is already uninformative, so adding even more noise to it doesn't change anything!

A convenient way to model this non-uniqueness of representation (3) is to use quotient spaces: To give an extended Gaussian, we first specify its locus of nondeterminism D, and then a Gaussian distribution on the quotient space $\mathbb{R}^n/D$. For extended Gaussians on $\mathbb{R}$, we only have two possibilities for the subspace D

1. If $D = \mathbb{R}$, then the quotient space $\mathbb{R}/\mathbb{R}$ consists of a single point; so there is only a single such extended Gaussian, which is the uniform distribution $\mathbb{R}$; this is what happens in (4)
2. If $D = 0$, there is no nondeterministic contribution, so we recover all ordinary Gaussian distributions $\mathcal{N}(\mu, \sigma^2)$ on $\mathbb{R} \cong \mathbb{R}/0$

We give a second example of an extended Gaussian on $\mathbb{R}^2$ with one-dimensional locus of nondeterminism.

Example 1. The following two extended Gaussians on $\mathbb{R}^2$ are equal

$$\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1&0\\0&1\end{pmatrix}\right) + D = \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}0&0\\0&2\end{pmatrix}\right) + D \tag{5}$$

where $D = \{(y, y) : y \in \mathbb{R}\}$ is the diagonal in $\mathbb{R}^2$.

The left hand side represents the distribution of a random vector $(X_1, X_2) + (Y, Y)$ where $X_1, X_2 \sim \mathcal{N}(0,1)$ are independent normal variables and $Y \sim \mathbb{R}$ is uninformative. We can decompose this expression into a sum as follows

$$(X_1, X_2) + (Y, Y) = (0, X_2 - X_1) + \underbrace{(X_1 + Y, X_1 + Y)}_{\in D}$$

and notice that the contribution of the second summand gets absorbed modulo D. The remaining vector $(0, X_2 - X_1)$ has covariance matrix $\begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}$ as claimed.

2.1 Definition of Extended Gaussians

We wish to define an extended Gaussian distribution on $\mathbb{R}^n$ with locus of nondeterminism D as an ordinary Gaussian distribution on the quotient space $\mathbb{R}^n/D$ (Definition 2). So far, we have only considered Gaussian distributions on Euclidean spaces $\mathbb{R}^d$, but not on more general vector spaces like quotients. In practice, we can side-step this issue by identifying the quotient space $\mathbb{R}^n/D$ with any *complement* K of D; that is a subspace K such that $K \oplus D = \mathbb{R}^n$. However, because the choice of such complement is non-canonical, we prefer to work with the quotient directly. This means we have to define Gaussian distributions on general vector spaces X in a coordinate-free way (Definition 1). This development is well-known, and we refer to e.g. [29, Section 1] for an overview:

In what follows, all vector spaces X are assumed to be finite-dimensional. We write X^* for the dual space of X, consisting of all linear maps $X \to \mathbb{R}$. By a *form* on X, we mean a *symmetric bilinear map* $\Omega : X \times X \to \mathbb{R}$. The kernel of Ω is the set

$$\ker(\Omega) \stackrel{\text{def}}{=} \{x \in X : \Omega(x, -) = 0\} = \{y \in X : \Omega(-, y) = 0\}$$

We call Ω *nondegenerate* if $\ker(\Omega) = 0$. Note that Ω can be curried as $\omega : X \to X^*$ with $\omega(x) = \Omega(x, -)$. The notation is consistent in that $\ker(\Omega) = \ker(\omega)$, and Ω is nondegenerate if and only if $\omega : X \to X^*$ is an isomorphism. If Ω has kernel K, then the quotient form $\widetilde{\Omega} : (X/K) \times (X/K) \to \mathbb{R}$ is well-defined and nondegenerate.

A form $\Omega : X \times X \to \mathbb{R}$ is called *positive semidefinite* if $\Omega(x, x) \geq 0$ for all $x \in X$, and *positive definite* if $\Omega(x, x) > 0$ for all $x \neq 0$. A positive semidefinite form is positive definite if and only if it is nondegenerate.

The correct coordinate-free version of the covariance matrix is that of a covariance form $\Sigma : X^* \times X^* \to \mathbb{R}$ on the dual space. Given a random variable U on the space X, and linear functionals $f, g : X \to \mathbb{R}$, we compute the covariance

$$\Sigma(f, g) \stackrel{\text{def}}{=} \mathbb{E}[f(U)g(U)] - \mathbb{E}[f(U)]\mathbb{E}[g(U)]$$

which is symmetric, bilinear and positive semidefinite. A Gaussian distribution is fully determined by its mean and covariance form, which motivates the following definition:

Definition 1. *A Gaussian distribution on a vector space X is a pair written $\mathcal{N}(\mu, \Sigma)$ of a mean $\mu \in X$ and a positive semidefinite form $\Sigma : X^* \times X^* \to \mathbb{R}$. If $f : X \to Y$ is a linear map, the Gaussian distribution pushes forward to $\mathcal{N}(f(\mu), f_*\Sigma)$ where $(f_*\Sigma)(g, h) = \Sigma(gf, hf)$.*

We can now give the following concise definition of extended Gaussians.

Definition 2. *An* extended Gaussian distribution *on a vector space X is a pair (D, ψ) where $D \subseteq X$ is a vector subspace and ψ is a Gaussian distribution on X/D. Every linear map $f : X \to Y$ induces a linear map $f_D : X/D \to Y/f[D]$, and we define the pushforward of (D, ψ) to be $(f[D], (f_D)_*\psi)$.*

Example 2. We verify equation (5) of Example 1 using Definition 2: The two distributions

$$\psi_1 = \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1&0\\0&1\end{pmatrix}\right), \psi_2 = \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}0&0\\0&2\end{pmatrix}\right)$$

have covariance forms $\Sigma_1, \Sigma_2 : (\mathbb{R}^2)^* \times (\mathbb{R}^2)^* \to \mathbb{R}$ given by

$$\Sigma_1(a, b) = a_1b_1 + a_2b_2, \quad \Sigma_2(a, b) = 2a_2b_2$$

when we identify functionals in $(\mathbb{R}^2)^*$ with row vectors. We form the quotient modulo D by pushing forward under the quotient map $\pi : \mathbb{R}^2 \to \mathbb{R}^2/D$. Functionals $a \in (\mathbb{R}^2/D)^*$ are precisely those that vanish on D, i.e. satisfying the constraint $a_1 + a_2 = 0$, and the induced forms $\widetilde{\Sigma_i} = \pi_*\Sigma_i$ are simply the restrictions of Σ_i to those functionals. Hence we have equality

$$a_1b_1 + a_2b_2 = (-a_2)(-b_2) + a_2b_2 = 2a_2b_2$$

as desired. Alternatively, we can verify (5) by identifying $\mathbb{R}^2/D$ with a complement like $K = 0 \times \mathbb{R}$ and projecting ψ_1, ψ_2 onto K. The projection is here given by the matrix $P = \begin{pmatrix}0&0\\-1&1\end{pmatrix}$ and we verify

$$P\begin{pmatrix}1&0\\0&1\end{pmatrix}P^T = P\begin{pmatrix}0&0\\0&2\end{pmatrix}P^T$$

2.2 Precision and Duality

We will show that our definition of extended Gaussians fits into an elegant duality between forms on a space and its dual. This lets us convert between two equivalent representations, using a *covariance form* or a *precision form*, which are convenient for different purposes:

Probability distributions and probability densities are dual to each other. Distributions naturally push forward, and consequently the covariance form must be defined on the *dual space* X^*. On the other hand, *density functions* pull back. If the covariance matrix Σ is nonsingular, the Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ has a Lebesgue density

$$f(x) \propto \exp\left(-\frac{1}{2}(x-\mu)^T\Omega(x-\mu)\right) \tag{6}$$

where $\Omega = \Sigma^{-1}$ is called the *precision matrix*. If Σ is singular, $\mathcal{N}(\mu, \Sigma)$ will only admit a density on its support S, and the precision is only defined on that

subspace. The coordinate-free formulation of the precision-covariance correspondence is given by the following duality statement (where for simplicity we assume centered distributions, that is mean zero):

Theorem 1. *The following pieces of data are equivalent for every vector space X*

1. *a form $\Sigma : X^* \times X^* \to \mathbb{R}$*
2. *a subspace $S \subseteq X$ and a nondegenerate form $\Omega : S \times S \to \mathbb{R}$*

Recall that the evaluation pairing $X^* \times X \to \mathbb{R}$ induces a duality between the subspaces of X and X^* called *annihilators*, which we here denote $(-)^\perp$. For subspaces $D \subseteq X$ and $F \subseteq X^*$, the subspaces $D^\perp \subseteq X^*, F^\perp \subseteq X$ are defined as

$$D^\perp \stackrel{\text{def}}{=} \{f \in X^* : f|_D = 0\}, \qquad F^\perp \stackrel{\text{def}}{=} \{x \in X : \forall f \in F, f(x) = 0\} \tag{7}$$

Taking annihilators is order-reversing and involutive; we list further properties under Proposition 6.

In Theorem 1, the subspace S is taken to be annihilator of the kernel $K = \ker(\Sigma)$, that is $S = \{x : \forall f \in K, f(x) = 0\}$. This recovers the familiar support $S = \mathsf{col}(\Sigma)$ for covariance matrices. We may think of the form Ω as taking the value infinity outside of S (which corresponds to vanishing a density under (6)).

Extended Gaussians admit a generalized and in fact more symmetric version of this correspondence:

Theorem 2. *The following pieces of data are equivalent for every vector space X*

1. *"precision representations", i.e. pairs $\langle S, \Omega \rangle$ of a subspace $S \subseteq X$ and $\Omega : S \times S \to \mathbb{R}$*
2. *"covariance representations", i.e. pairs $\langle F, \Sigma \rangle$ of a subspace $F \subseteq X^*$ and $\Sigma : F \times F \to \mathbb{R}$*

Proof. 1. Given $\langle S, \Omega \rangle$, let $D = \ker(\Omega) \subseteq S$ and define $F = D^\perp$ and $K = S^\perp$. Form the nondegenerate quotient form $\widetilde{\Omega} : (S/D) \times (S/D) \to \mathbb{R}$. Its currying $\widetilde{\omega} : (S/D) \to (S/D)^*$ is an isomorphism. Making use of the canonical isomorphisms described in Proposition 6,

$$\begin{array}{ccc} (S/D)^* & \xrightarrow{\widetilde{\omega}^{-1}} & S/D \\ \sim\uparrow & & \| \\ D^\perp/S^\perp & & K^\perp/F^\perp \\ \| & & \downarrow\sim \\ F/K & \xrightarrow{\widetilde{\sigma}} & (F/K)^* \end{array} \tag{8}$$

we obtain an iso $\widetilde{\sigma} : (F/K) \to (F/K)^*$, which is the same thing as a bilinear form $\Sigma : F \times F \to \mathbb{R}$ with kernel K.

2. Conversely, given $\langle F, \Sigma \rangle$, let $K = \ker(\Sigma)$ and define $D = F^\perp$ and $S = K^\perp$. Turn Σ into an iso $\widetilde{\sigma} : (F/K) \to (F/K)^*$, then reading the diagram (8) backwards uniquely defines the iso $\widetilde{\omega}$ and hence the form Ω with kernel D.

The constructions are clearly inverses to each other. It is furthermore easy to see that the correspondence takes positive semidefinite forms to positive semidefinite forms.

A (centered) extended Gaussian can thus be presented in two interconvertible ways: A *covariance representation*, which is a pair $\langle F, \Sigma \rangle$ with $F \subseteq X^*$ and Σ a positive semidefinite form on F, and a *precision representation* $\langle S, \Omega \rangle$ with $S \subseteq X$ and Ω positive semidefinite on S.

The covariance representation is convenient for computing pushforwards, while the precision representation generalizes density functions and is useful for conditioning. Note that the locus of nondeterminism of an extended Gaussian equals $D = \ker(\Omega)$; for ordinary Gaussians we insisted that Ω be nondegenerate, i.e. $D = 0$! The proof of Theorem 2 is reminiscent of the construction of the Moore-Penrose pseudoinverse, whose relevance to Gaussian probability is well-known (e.g. [20]).

Example 3. The uniform distribution on X, i.e. the relation $X \subseteq X$, has as precision representation the zero form $X \times X \to \mathbb{R}$; its covariance representation is the zero form on the trivial subspace $0 \subseteq X^*$.

3 A Category of Extended Gaussian maps

In this section, we will generalize extended Gaussian distributions to a category of *extended Gaussian maps*, i.e. affine-linear maps with extended Gaussian noise. This is useful because it captures in one precise formalism all the different ways to combine extended Gaussians, such as forming their product distributions, sums and pushforwards. It will also reveal that extended Gaussians are an instance of a general construction of *decorated linear relations*, which makes the relationship with nondeterminism more precise.

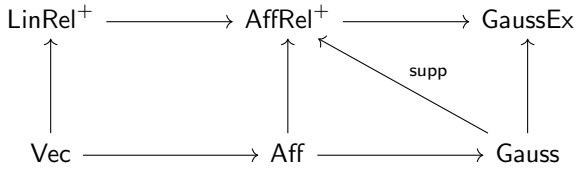
Roughly, a *decorated linear map* $X \to Y$ between vector spaces is a pair (f, s) of a linear map $f : X \to Y$ and a piece of noise or 'decoration' s on Y. Depending on which type of decoration s we allow, we obtain affine-linear maps or Gaussian maps this way.

A *decorated linear relation* $X \to Y$ is a pair (D, f, s) of a subspace $D \subseteq Y$ and decorated linear map $X \to Y/D$ into the quotient. This does indeed generalize linear relations, because of the following lemma (elaborated in Proposition 5)

Lemma 1. *To give a left-total linear relation $R \subseteq X \times Y$ is to give subspace $D \subseteq Y$ and a linear function $f : X \to Y/D$ to the quotient.*

The most elegant way of defining extended Gaussian maps is by decorating linear relations with Gaussian noise. We believe that this level of abstraction is helpful to make the construction more modular, structure the tedious calculations in Section 6, and make it clear which parts of the construction are

completely general, rather than relying on specific properties of Gaussians. In Section 3.4, we use the generality of this construction to establish the key diagram of functors from the introduction, connecting all categories of interest in this paper.



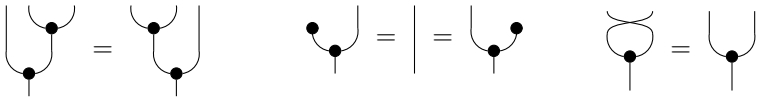
(9)

Notably, the functor supp collapses Gaussian distributions to their supports (which are affine subspaces).

3.1 Markov categories

Markov categories [10] are a recent axiomatization for categorical models of probability theory. They have been used do give abstract characterizations of concepts such as determinism, independence, conditioning, zero-one laws and De Finetti's theorem [12, 11].

A Markov category is a symmetric monoidal category $(\mathbb{C}, \otimes, I)$ in which each object X comes equipped with the structure of a commutative comonoid, formalizing copying ($\mathsf{cpy}_X : X \to X \otimes X$) and deleting ($\mathsf{del}_X : X \to I$), and satisfying certain coherence conditions. Importantly, deleting is natural but copying need not be. We use use string diagram notation to construct morphisms in Markov categories in an intuitive way; for example, the comonoid axioms relating copy and delete can be rendered as



Markov categories are also the semantic counterpart of a class of first-order probabilistic programming languages [27], which makes them a natural structure for denotational semantics. We will return to this perspective in section 5.

Example 4. The prototypical Markov categories, FinStoch, Stoch and Rel^+, capture discrete probability, measure-theoretic probability and nondeterminism respectively [10]:

1. in FinStoch, objects are finite sets X and morphisms $X \to Y$ are stochastic matrices $p \in [0,1]^{Y \times X}$
2. in Stoch, objects are measurable spaces (X, Σ_X) and morphisms $(X, \Sigma_X) \to (Y, \Sigma_Y)$ are Markov kernels $\kappa : X \times \Sigma_Y \to [0,1]$.
3. in Rel^+, objects are sets X and morphisms $X \to Y$ are left-total relations $R \subseteq X \times Y$

In all cases, the tensor product is $X \otimes Y = X \times Y$, and copying and deleting are inherited from the usual copy and deleting operations $x \mapsto ()$ and $x \mapsto (x, x)$ in Set.

Left-total linear relations form a Markov subcategory LinRel^+ of Rel^+. The Markov category Gauss (due to [10, Ch. 6]) is defined as follows: Objects are vector spaces $\mathbb{R}^n$, and morphisms are formal sums of linear maps with Gaussian noise $f + \psi$. On objects, the tensor is cartesian product, $X \otimes Y = X \times Y$, and the categorical structure is given by

$$(f + \psi) \circ (g + \varphi) = fg + \psi + f_*\varphi \qquad (f + \psi) \otimes (f' + \psi') = f \times f' + \psi \otimes \psi'$$

where $\psi \otimes \psi'$ denotes the product distribution of Gaussians, and $f_*\varphi$ is push-forward. One can embed Gauss into Stoch by assigning $\mathbb{R}^n$ its Borel-σ-algebra and interpreting Gaussian maps as actual Markov kernels. This will no longer be possible for extended Gaussian maps, which are not measure theoretic.

We will now generalize the construction of Gauss to decorated linear maps.

3.2 Decorated Linear Maps

Let $S : \mathsf{Vec} \to \mathsf{CMon}$ be a functor from the category of vector spaces into the category of commutative monoids and monoid homomorphisms. We think of elements $s \in S(Y)$ as "decorations" (or noise) for linear maps into Y, and call S a decoration functor.

Definition 3. *We define a category Lin_S of S-decorated linear maps as follows:*

1. *Objects are vector spaces X*
2. *Morphisms are pairs (f, s) where $f : X \to Y$ is a linear map and $s \in S(Y)$*
3. *Composition is defined as follows: for $g : X \to Y$, $f : Y \to Z$, $s \in S(Y)$, $t \in S(Z)$ let*
$$(f, t) \circ (g, s) = (fg, t + S(f)(s))$$
Note that addition takes place in the commutative monoid $S(Z)$.

There is a faithful inclusion $J : \mathsf{Vec} \to \mathsf{Lin}_S$ sending f to $(f, 0)$. The functor $U : \mathsf{Lin}_S \to \mathsf{Vec}$, which forgets the decoration, is an opfibration; a decorated map $(f, s) : X \to Y$ is opcartesian if and only if s is invertible in the monoid $S(Y)$. In fact, Lin_S is precisely the (op-)Grothendieck construction for S seen as a functor $\mathsf{Vec} \to \mathsf{Cat}$.

We argue that Lin_S has the structure of a symmetric monoidal category with the tensor $X \otimes Y = X \times Y$ on objects. For this, we first observe that S is automatically lax monoidal; for this we define natural maps $\oplus : S(X) \times S(Y) \to S(X \times Y)$ given as follows: For $(s, t) \in S(X) \times S(Y)$, let $s \oplus t = S(i_X)(s) + S(i_Y)(t)$ where $i_X : X \to X \times Y, i_Y : Y \to X \times Y$ are the biproduct inclusions. We can now define the tensor of decorated map as $(f, s) \otimes (g, t) = (f \times g, s \oplus t)$. The monoidal category Lin_S is in general not cartesian; it does however inherit copy and delete maps from Vec. The category Lin_S is a Markov category if and only if deleting is natural, i.e. $S(0) \cong 0$, where 0 denotes the terminal vector space/commutative monoid.

Example 5. We reconstruct the bottom row of (9) for the following decoration functors:

1. For $S(X) = 0$, Lin_S is equivalent to Vec.
2. For $S(X) = X$, Lin_S is equivalent to Aff. A map $X \to Y$ consists of a pair (f, y) with $f : X \to Y$ linear and $y \in Y$.
3. Define $\mathsf{Cov}(X) = \{\sigma : X^* \times X^* \to \mathbb{R} \text{ positive semidefinite form}\}$ to be the set of covariance forms. This is a commutative monoid under pointwise addition, and is functorial via $\mathsf{Cov}(f)(\sigma)(g, h) = \sigma(gf, hf)$. Let $S(X) = X \times \mathsf{Cov}(X)$, then Lin_S is equivalent to Gauss.

3.3 Decorated Linear Relations

Given a decoration functor $S : \mathsf{Vec} \to \mathsf{CMon}$, we define S-decorated linear relations by maintaining a locus of nondeterminism D similar to Definition 2.

Definition 4. *We define a category* LinRel_S *as follows:*

1. *objects are vector spaces* X
2. *morphisms in* $\mathsf{LinRel}_S(X, Y)$ *are triples* (D, f, s) *where* $D \subseteq Y$ *is a vector subspace,* $f : X \to Y/D$ *is a linear map and* $s \in S(Y/D)$.

Intuitively, the subspace D represents the direction of complete ignorance, so we only decorate the quotient.

Composition of $(D, g, s) : X \to Y$ and $(E, f, t) : Y \to Z$ is slightly more involved: We first define the composite subspace F as $E + f[D]$, which is well-defined. The composite $Y \xrightarrow{f} Z/E \to Z/F$ vanishes on D and so descends to $\tilde{f} : Y/D \to Z/F$. We define the composite as

$$(E, f, t) \circ (D, g, s) = (F, \tilde{f}g, S(\tilde{f})(s) + S(Z/E \to Z/F)(t))$$

To understand the name 'decorated linear relation', we notice that by Proposition 5, the category LinRel_S is equivalent to LinRel^+ for $S = 0$. This means we can think of a left-total relation R as a linear function with nondeterministic noise $x \mapsto f(x) + D$ along some subspace D. This is similar to a decorated linear map, however the choice of the linear map f is no longer unique, so further quotienting is required, as we discuss now:

As a quotient: We can demystify the composition of LinRel_S by first constructing an auxiliary category and then quotienting it by a congruence. We first consider the decoration functor $\mathsf{Sub} : \mathsf{Vec} \to \mathsf{CMon}$ defined by $\mathsf{Sub}(X) = \{D \subseteq X \text{ vector subspace }\}$. Each $\mathsf{Sub}(X)$ is a commutative monoid under Minkowski addition $D + E$, and the functorial action for $f : X \to Y$ is direct image $D \mapsto f[D]$, which is a monoid homomorphism. Now we consider the category $\mathsf{Lin}_{S \times \mathsf{Sub}}$ where morphisms $X \to Y$ are by definition triples (f, s, D) with $s \in S(Y)$ and $D \subseteq Y$, and composition is $(f, t, E) \circ (g, s, D) = (fg, t + S(f)(s), E + f[D])$. This

has all the data for LinRel_S and explains why the composite subspace is formed the way it is.

However some pieces of data need to be identified if they agree on the quotient by D: We define an equivalence relation $(f_1, s_1, D_1) \approx (f_2, s_2, D_2)$ if $D_1 = D_2$ are the same subspace D, and $\pi_{Y/D} f_1 = \pi_{Y/D} f_2$ and $S(\pi_{Y/D})(s_1) = S(\pi_{Y/D})(s_2) \in S(Y/V)$ where $\pi_{Y/D} : Y \to Y/D$ is the quotient map.

Proposition 1. *The relation $\approx$ is a congruence relation, and LinRel_S is the quotient of $\mathsf{Lin}_{S \times \mathsf{Sub}}$ under $\approx$. LinRel_S is symmetric monoidal and inherits the copy and delete maps from Vec. It is a Markov category if $S(0) \cong 0$.*

Proof. See Appendix (Section 6.4).

Definition 5. *We define the Markov category $\mathsf{GaussEx}$ to be $\mathsf{LinRel}_{X \times \mathsf{Cov}}$.*

In the style of Proposition 5, we can show that $\mathsf{LinRel}_X \cong \mathsf{AffRel}^+$. This way the LinRel construction applied to Example 5 gives rise to the top row of (9).

3.4 Relationships between the Constructions

The constructions Lin and LinRel are themselves functorial:

Proposition 2. *Let $S, T : \mathsf{Vec} \to \mathsf{CMon}$ be decoration functors and $\alpha : S \to T$ a natural transformation. Then we have induced monoidal identity-on-objects functors*

$$F_\alpha : \mathsf{Lin}_S \to \mathsf{Lin}_T, (f, s) \mapsto (f, \alpha(s))$$
$$G_\alpha : \mathsf{LinRel}_S \to \mathsf{LinRel}_T, (D, f, s) \mapsto (D, f, \alpha(s))$$

which preserve copy and delete structure.

This proposition accounts for all functors in the diagram (9) except for supp. For every decoration functor S, we obtain an inclusion functor $\mathsf{Lin}_S \to \mathsf{LinRel}_S$ by choosing locus of nondeterminism $D = 0$, that is forming the composite

$$\mathsf{Lin}_S \xrightarrow{F_{\mathsf{id} \times 0}} \mathsf{Lin}_{S \times \mathsf{Sub}} \twoheadrightarrow \mathsf{LinRel}_S$$

From a decorated linear relation, we can extract its locus of nondeterminism by forgetting the decoration via the following composite:

$$\mathsf{LinRel}_S \xrightarrow{G_0} \mathsf{LinRel}_0 \cong \mathsf{LinRel}^+$$

Recall that the support of a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ on $\mathbb{R}^n$ is the affine space $\mu + \mathsf{col}(\Sigma)$. This construction can be extended functorially to Gauss as follows:

Proposition 3. *We have a functor $\mathsf{supp} : \mathsf{Gauss} \to \mathsf{AffRel}^+$ which takes the Gaussian noise to its support. Concretely on $\mathbb{R}^n$, $\mathsf{supp}(x \mapsto Ax + \mathcal{N}(\mu, \Sigma)) = (x \mapsto Ax + \mu + \mathsf{col}(\Sigma))$.*

Proof. We construct supp as the composite

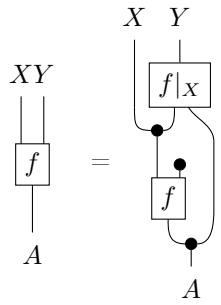
$$\mathsf{Lin}_{X\times\mathsf{Cov}} \xrightarrow{F_{\mathrm{id}_X\times\alpha}} \mathsf{Lin}_{X\times\mathsf{Sub}} \twoheadrightarrow \mathsf{LinRel}_X \cong \mathsf{AffRel}^+$$

where $\alpha_X : \mathsf{Cov}(X) \to \mathsf{Sub}(X)$ is takes the covariance form σ to the annihilator of its kernel. Naturality is nontrivial and can be shown using the Cholesky decomposition, which makes use of the positive semidefiniteness of σ.

4 Conditioning and Probabilistic Programs

The self-conjugacy of Gaussians can be expressed through the fact that the category Gauss has conditionals. This lets us not only express statistical models in Gauss, but also answer inference queries.

Recall that a *conditional* [10, Definition 11.5] for a morphism $f : A \to X \otimes Y$ in a Markov category is a morphism $f_{|X} : X \otimes A \to Y$ which lets us reconstruct f from one of its marginals as $f(x, y|a) = f_{|X}(y|x, a)f_X(y|a)$, or in string diagrams



Conditionals in Gauss exist and are given by the usual conditional distributions [10, 11.8]. LinRel^+ also has conditionals, which are essentially given by re-ordering a relation $R \subseteq A \times (X \times Y)$ to a relation $R_{|X} \subseteq (X \times A) \times Y$. This $R_{|X}$ may not be left-total yet, but any linear relation can be extended to a left-total one outside of its domain.

We will now prove that extended Gaussians have conditionals too. By picking a convenient complement to the locus of nondeterminism D, we can show the following fact about decorated linear relations:

Theorem 3. *If* Lin_S *has conditionals, so does* LinRel_S.

Proof. Let $\varphi \in \mathsf{LinRel}_S(A, X \times Y)$ be represented modulo $\approx$ by (f, μ, D) with $\mu \in S(X \times Y)$. By Lemma 2, there exists a complement $K \subseteq X \times Y$ of D such that K_X is a complement of D_X in X. We then take $P_K, P_D : X \times Y \to X \times Y$ to be the projection endomorphisms, and replace the representative μ by $S(P_K)(\mu)$ and f by $P_K f$ without affecting φ.

Now we consider $(f, \mu) \in \mathsf{Lin}_S(A, X \times Y)$ and find a conditional $(g, \psi) \in \mathsf{Lin}_S(X \times A, Y)$. This means we can obtain $(X_1, Y_1) \sim f(a) + \mu$ as

$$X_1 \sim f_X(a) + \mu_X \qquad Y_1 \sim g(x, a) + \psi$$

Similarly we can use conditionals in LinRel^+ to find a linear function $h : X \to Y$ and a subspace $H \subseteq Y$ such that $(X_2, Y_2) \sim D$ can be obtained as

$$X_2 \sim D_X \qquad Y_2 \sim h(X_2) + H$$

Thus a joint sample $(X, Y) \sim \varphi(a)$ can be obtained as follows

$$\begin{aligned} X_1 &\sim f_X(a) + \mu_X & Y_1 &\sim g(X_1, a) + \psi \\ X_2 &\sim D_X & Y_2 &\sim h(X_2) + H \\ X &= X_1 + X_2 & Y &= Y_1 + Y_2 \end{aligned}$$

Because we have chosen K such that $K_X \oplus D_X = X$, we can extract the values of X_1, X_2 from X via the projections $P_{K_X}, P_{D_X} : X \to X$ as $X_1 = P_{K_X}(X)$ and $X_2 = P_{D_X}(X)$. We can thus read off a conditional for φ by combining the two individual conditionals, namely

$$\varphi|_X(x, a) = g(P_{K_X}(x), a) + h(P_{D_X}(x)) + \psi + H$$

Corollary 1. $\mathsf{GaussEx}$ *has all conditionals.*

Proof. By combining Theorem 3 with the fact that Gauss has all conditionals.

We exemplify the procedure of Theorem 3 for Example 1: In order to find a conditional distribution for the joint distribution $(Z_1, Z_2) \sim \mathcal{N}(0, I_2) + D$ on $\mathbb{R} \times \mathbb{R}$, we choose the particular complement $K = 0 \times \mathbb{R}$ of the diagonal and obtain desired decomposition, $Z_1 \sim \mathbb{R}$ and $Z_2 \sim Z_1 + \mathcal{N}(0, 2)$.

4.1 Conditioning on Equality

In the context of Gaussian probability, we can condition two random variables U, V to be *exactly equal* [28] by introducing an auxiliary variable $Z = U - V$ for their difference, and computing the conditional distribution $(U, V)|Z = 0$. We can now formally show that the uniform extended Gaussians $X \subseteq X$ is really uninformative in the sense that conditioning on equality with a uniform variable does not change the prior:

Proposition 4. *For every extended Gaussian prior ψ on X, if $U \sim \psi$ and $V \sim X$, then $U|(U = V)$ still has distribution ψ.*

Proof. We introduce an auxiliary random variable $Z = U - V$ and show that the following two joint distributions over (U, V, Z) are equal:

$$\left\{ \begin{array}{c} U \sim \psi \\ V \sim X \\ Z = U - V \end{array} \right\} = \left\{ \begin{array}{c} Z \sim X \\ U \sim \psi \\ V = U - Z \end{array} \right\}$$

The right-hand side lets us read off the conditional on Z immediately. Conditioning on equality now means setting $Z = 0$, after which we obtain $U \sim \psi, V = U$.

4.2 Probabilistic Programming with Exact Conditions

Probabilistic programming is a powerful and flexible paradigm for statistical inference which has gained traction in recent years (e.g. [30, 25, 14]). Its goal is to express probabilistic models and observed data compositionally within the same language; executing a probabilistic program then means employing various inference algorithms to answer statistical questions, such as sampling from a posterior distribution.

In [28], we argued that the exact conditioning operation (conditioning on equality) described in Section 4.1 is a fundamental primitive in such programs, and enjoys good logical properties. We presented a programming language for Gaussian probability featuring a first-class exact conditioning operator ($=:=$), with Python/F# implementations available under [26].

The behavior of ($=:=$) can be quite intricate. While the denotational semantics defined in [28] on the basis of the category Gauss is fully abstract, it is still lacking a concrete description of when two exact conditioning programs are contextually equivalent. We demonstrate now how, by generalizing the language to extended Gaussians, we can obtain such a concrete description:

As an example, consider the following function of type $\mathbb{R} \to \mathbb{R}$

$$f(x) \stackrel{\text{def}}{=} (\mathsf{let}\, y \,=\, \mathcal{N}(0,1)\, \mathsf{in}\, x =:= y; \mathsf{return}(y)) \tag{10}$$

which takes a variable x, conditions it on equality with a local Gaussian variable y and then returns y. For every constant $\underline{c} \in \mathbb{R}$, it holds that $f(\underline{c}) = \underline{c}$ because y assumes the value $\underline{c}$ exactly after conditioning. However f does not behave like the identity function on distributional inputs; for example $f(\mathcal{N}(0,1)) = \mathcal{N}(0, 0.5)$ because conditioning two independent variables on equality reduces their variance [28, Ex. II.1].

Extending the Gaussian language with improper priors lets us reduce the problem of characterizing functions $f : X \to Y$ to the easier problem of characterizing distributions on $X \otimes Y$. Write $\mathsf{uniform}_X$ for a new language construct which gets interpreted as the uniform distribution over X. Then we define the *name of* f [16, Def. 3.3] as

$$\ulcorner f \urcorner = (\mathsf{let}\, x \,=\, \mathsf{uniform}_X\, \mathsf{in}\, (x, f(x)))$$

We can bijectively recover the function from its name using exact conditioning

$$f(x) = (\mathsf{let}\, (x', y) \,=\, \ulcorner f \urcorner\, \mathsf{in}\, (x =:= x'); \mathsf{return}(y))$$

This is reminiscent of how functions get encoded as predicates in Prolog. This duality lets us identify the contextual equivalence classes of programs $\mathbb{R} \to \mathbb{R}$ with extended Gaussian distributions on $\mathbb{R}^2$. In our example, the name of (10) turns out to be an ordinary Gaussian distribution supported on the diagonal; in contrast, the name of the identity function $\mathsf{id}(x) = x$ is uniform on the diagonal

$$\ulcorner f \urcorner = (\mathsf{let}\, x \,=\, \mathcal{N}(0,1)\, \mathsf{in}\, \mathsf{return}(x, x))$$
$$\ulcorner \mathsf{id} \urcorner = (\mathsf{let}\, x \,=\, \mathsf{uniform}_X\, \mathsf{in}\, (x, x))$$

Asking for an uninformative distribution is very natural from this formal standpoint: it is simply a *unit* for $(=:=)$. Logically, we can think of the uniform distribution as analogous to the existential quantifer $\exists$, or as a fresh variable in logic programming (e.g. [24]).

5 Discussion and Future Work

The principal motivation for this work comes from programming language theory and the semantics of probabilistic programs. In this paper we have defined the extended Gaussian model and established desired properties. The Cond-construction of [28] provides a way to turn it into a model of an exact conditioning language, with Theorem 3 being the crucial ingredient for the construction. The category $\mathsf{Cond}(\mathsf{GaussEx})$ can express both logical (solving linear equations) and probabilistic inference (conditioning Gaussians). We believe that it is a hypergraph category [9], with conditioning as multiplication and uniform distributions as units. Hypergraph categories enjoy a rich duality theory and were suggested by [8] as the natural setting for inference; celebrated algorithms such as message-passing inference can be formulated in them fully abstractly [22].

Further connections between probability and logic arise both through the analogy with unification in logic programming [24] and our new and generalized perspective on linear relations (Section 3.4). It will also be interesting to analyze $\mathsf{Cond}(\mathsf{GaussEx})$ from the perspective of categorical logic [17, 7].

We hope that the concept of extended Gaussians will also be useful outside of theoretical computer science. In this paper, we have introduced them primarily from an algebraic and category-theoretic viewpoint and shown a duality theorem (Theorem 2) linking them to certain types of quadratic forms. We would like to explore further connections to statistics [21] and functional analysis: An important step is to topologize the homsets of $\mathsf{GaussEx}$ and characterize which in which ways Gaussian distributions can converge to extended Gaussians. Ideally this would exhibit the construction of extended Gaussians as a form of topological completion. The aspect of considering improper priors as limits of normalized ones is treated in [1, 4].

It also seems interesting to consider extended Gaussians under the 'principle of transformation groups' (e.g. [19]). In [28, VI], we remark that Gauss is essentially presented as a PROP by the invariance of the standard normal distribution under the orthogonal group $O(n)$. We expect the uniform distributions to be presented by invariance under all of $\mathsf{GL}(n)$.

Acknowledgements: It has been useful to discuss this work with many people. Particular thanks go to Tobias Fritz, Bart Jacobs, Dusko Pavlovic, Sam Staton and Alexander Terenin.

References

1. Akaike, H. The interpretation of improper prior distributions as limits of data dependent proper prior distributions. *Journal of the Royal Statistical Society. Series B (Methodological) 42*, 1 (1980), 46–52.
2. Baez, J. C., Coya, B., and Rebro, F. Props in network theory, 2018.
3. Baez, J. C., and Erbele, J. Categories in control. *Theory Appl. Categ. 30* (2015), 836–881.
4. Bioche, C., and Druilhet, P. Approximation of improper priors. *Bernoulli 22*, 3 (2016), 1709–1728.
5. Bonchi, F., Sobocinski, P., and Zanasi, F. The calculus of signal flow diagrams I: linear relations on streams. *Inform. Comput. 252* (2017).
6. Cho, K., and Jacobs, B. Disintegration and Bayesian inversion via string diagrams. *Mathematical Structures in Computer Science 29* (2019), 938 – 971.
7. Cho, K., Jacobs, B., Westerbaan, B., and Westerbaan, A. An introduction to effectus theory.
8. Coecke, B., and Spekkens, R. W. Picturing classical and quantum Bayesian inference. *Synthese 186* (2011), 651–696.
9. Fong, B., and Spivak, D. I. Hypergraph categories. *ArXiv abs/1806.08304* (2019).
10. Fritz, T. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Adv. Math. 370* (2020).
11. Fritz, T., Gonda, T., and Perrone, P. De Finetti's theorem in categorical probability, 2021.
12. Fritz, T., and Rischel, E. F. Infinite products and zero-one laws in categorical probability. *Compositionality 2* (Aug. 2020).
13. Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian Data Analysis*, 2nd ed. ed. Chapman and Hall/CRC, 2004.
14. Goodman, N. D., Tenenbaum, J. B., and Contributors, T. P. Probabilistic Models of Cognition. http://probmods.org, 2016. Accessed: 2021-3-26.
15. Goy, A., and Petrişan, D. Combining probabilistic and non-deterministic choice via weak distributive laws. In *Proceedings of the 35th Annual ACM/IEEE Symposium on Logic in Computer Science* (New York, NY, USA, 2020), LICS '20, Association for Computing Machinery.
16. Heunen, C., and Vicary, J. *Categories for quantum theory: an introduction.* Oxford University Press, United Kingdom, Nov. 2019.
17. Jacobs, B. *Categorical Logic and Type Theory.* No. 141 in Studies in Logic and the Foundations of Mathematics. North Holland, Amsterdam, 1999.
18. Jacobs, J. Paradoxes of probabilistic programming. In *Proc. POPL 2021* (2021).
19. Jaynes, E. T. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics 4*, 3 (1968), 227–241.
20. Lauritzen, S., and Jensen, F. Stable local computation with conditional Gaussian distributions. *Statistics and Computing 11* (11 1999).
21. McCullagh, P. Quotient spaces and statistical models. http://www.stat.uchicago.edu/ pmcc/reports/quotient.pdf.
22. Morton, J. Belief propagation in monoidal categories. *Electronic Proceedings in Theoretical Computer Science 172* (12 2014), 262–269.
23. Selinger, P. *A Survey of Graphical Languages for Monoidal Categories.* Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 289–355.

24. Staton, S. An algebraic presentation of predicate logic. In *Foundations of Software Science and Computation Structures* (Berlin, Heidelberg, 2013), F. Pfenning, Ed., Springer Berlin Heidelberg, pp. 401–417.
25. Staton, S. Commutative semantics for probabilistic programming. In *Programming Languages and Systems* (Berlin, Heidelberg, 2017), H. Yang, Ed., Springer Berlin Heidelberg, pp. 855–879.
26. Stein, D. GaussianInfer. https://github.com/damast93/GaussianInfer, 2021.
27. Stein, D. *Structural Foundations for Probabilistic Programming Languages*. PhD thesis, University of Oxford, 2021.
28. Stein, D., and Staton, S. Compositional semantics for probabilistic programs with exact conditioning (long version), 2021.
29. Terenin, A. *Gaussian Processes and Statistical Decision-making in Non-Euclidean spaces*. PhD thesis, Imperial College London, 2022.
30. van de Meent, J.-W., Paige, B., Yang, H., and Wood, F. An introduction to probabilistic programming, 2018.
31. Zwart, M., and Marsden, D. No-go theorems for distributive laws. In *Proceedings of the 34th Annual ACM/IEEE Symposium on Logic in Computer Science* (2019), LICS '19, IEEE Press.

6 Appendix

6.1 Noisy measurement example

Example 6. We elaborate the noisy measurement example from the introduction. Formally, we introduce random variables

$$X \sim \mathcal{N}(50, 100)$$
$$Y \sim \mathcal{N}(X, 25)$$

The vector (X, Y) is multivariate Gaussian with mean $(50, 50)$ and covariance matrix

$$\Sigma = \begin{pmatrix} 100 & 100 \\ 100 & 125 \end{pmatrix}$$

The conditional distribution $X|(Y = 40)$ is $\mathcal{N}(42, 20)$.

Proof. The random vector (X, Y) has joint density function

$$f(x, y) = \frac{1}{2\pi \cdot 10 \cdot 5} \exp\left(-\frac{(x-50)^2}{2 \cdot 100}\right) \cdot \exp\left(-\frac{(y-x)^2}{2 \cdot 25}\right)$$

The conditional density of x given y has the form

$$f(x|y) = \frac{f(x, y)}{\int f(x, y) \mathrm{d}x}$$

By expanding and 'completing the square', it is easy to check that

$$f(x|y = 40) \propto \exp\left(-\frac{(x-42)^2}{2 \cdot 20}\right)$$

is again a Gaussian density, from which we read off $\mu = 42$ and $\sigma^2 = 20$.

6.2 Glossary: Linear Algebra

All vector spaces are assumed finite dimensional. For vector subspaces $U, V \subseteq X$, their *Minkowski sum* is the subspace $U + V = \{u + v : u \in U, v \in V\}$. If furthermore $U \cap V = 0$, we call their sum a *direct sum* and write $U \oplus V$. A *complement* of U is a subspace V such that $U \oplus V = X$. An *affine subspace* $W \subseteq X$ is a subset of the form $x + U$ for some $x \in X$ and a (unique) vector subspace $U \subseteq X$. The space W is called a *coset* of U and the cosets of U organize into the quotient vector space $X/U = \{x + U : x \in X\}$.

An affine-linear map $f : X \to Y$ between vector spaces is a map of the form $f(x) = g(x) + y$ for some linear function $g : X \to Y$ and $y \in Y$. Vector spaces and affine-linear maps form a category Aff.

A *linear relation* $R \subseteq X \times Y$ is a relation which is also a vector subspace of $X \times Y$. An *affine relation* $R \subseteq X \times Y$ is a relation which is also an affine subspace of $X \times Y$. We write $R(x) \overset{\text{def}}{=} \{y \in Y : (x, y) \in R\}$. A relation $R \subseteq X \times Y$ is *left-total* if $R(x) \neq \emptyset$ for all $x \in X$.

Linear relations, affine relations and left-total relations are closed under the usual composition of relations. We denote by LinRel^+ and AffRel^+ the categories whose objects are vector spaces, and morphisms are left-total linear and affine relations respectively. Those categories are Markov categories (left-totality must be assumed for discarding to be natural).

The following characterization underlies Definition 4: Every left-total linear relation can be written as a 'linear map with nondeterministic noise' $x \mapsto f(x) + D$.

Proposition 5. *Let $R \subseteq X \times Y$ be a left-total linear relation. Then*

1. *$R(0)$ is a vector subspace of Y*
2. *$R(x)$ is a coset of $R(0)$ for every $x \in X$*
3. *the assignment $x \mapsto R(x)$ is a well-defined linear map $X \to Y/R(0)$*
4. *every linear map $X \to Y/D$ is of that form for a unique left-total linear relation*

Proof. For 1, consider $y, y' \in R(0)$ (by assumption nonempty), then by linearity of R

$$(0, y) \in R, (0, y') \in R \Rightarrow (0, \alpha y + \beta y') \in R$$

so $R(0)$ is a vector subspace. For 2, we can find some $w \in R(x)$ and wish to show that $R(x) = w + R(0)$. Indeed if $y \in R(x)$ then $(x, y) - (x, w) = (0, y - w) \in R$ so $y - w \in R(0)$, hence $y \in w + R(0)$. Conversely for all $z \in R(0)$ we have $(x, w + z) = (x, w) + (0, z) \in R$ so $w + z \in R(x)$. This completes the proof that $R(x)$ is a coset. For 3, the previous point shows that the map $\rho : x \mapsto R(x)$ is a well-defined map $X \to Y/R(0)$. It remains to show it is linear. That is, if

$w \in R(x)$ and $z \in R(y)$ then $\alpha w + \beta z \in R(\alpha x + \beta y)$. This follows immediately from the linearity of R. For the last point 4, given a linear map $f : X \to Y/V$ we construct the relation

$$(x, y) \in R \Leftrightarrow y \in f(x)$$

which is left-total because $f(x) \neq \emptyset$. To see that R is linear, let $(x, y) \in R, (x', y') \in R$ meaning $y - z \in V$ and $y' - z \in V$ for representatives z, z' of $f(x), f(x')$. Linearity of f means that $\alpha z + \beta z'$ is a representative of $f(\alpha x + \beta x')$. Thus

$$\alpha y + \beta y' - (\alpha z + \beta z') = \alpha(y - z) + \beta(y' - z') \in V$$

Annihilators

Proposition 6. *1. Taking annihilators is order-reversing and involutive*

2. If $D \subseteq S \subseteq X$, then $S^\perp \subseteq D^\perp \subseteq X^*$ and we have a canonical isomorphism

$$(S/D)^* \cong D^\perp / S^\perp \tag{11}$$

and similarly for $K \subseteq F \subseteq X^*$, we have

$$(F/K)^* \cong K^\perp / F^\perp \tag{12}$$

3. We have

$$(V + W)^\perp = V^\perp \cap W^\perp$$
$$(F \cap W)^\perp = F^\perp + G^\perp$$

If $D \subseteq X$ and $f : X \to Y$, then

$$(f[D])^\perp = \{g \in Y^* : gf \in D^\perp\}$$

If $U \subseteq X, V \subseteq Y$, we have a canonical isomorphism

$$(U \times V)^\perp \cong U^\perp \times V^\perp$$

Proof. Standard. An explicit description of the canonical iso (11) is given as follows.

1. We define $\alpha : D^\perp / S^\perp \to (S/D)^*$ as follows. If $f \in D^\perp$, then f is a function $X \to \mathbb{R}$ such that $f|_D = 0$. The restriction $f|_S : S \to \mathbb{R}$ thus descends to the quotient $S/D \to \mathbb{R}$, and we let $\widetilde{\alpha}(f) = f|_S$. To check this is well-defined, notice that the kernel of $\widetilde{\alpha}$ consists of those $f \in X^*$ such that $f|_S = 0$, that is $S^\perp$.
2. We define $\alpha^{-1} : (S/D)^* \to D^\perp / S^\perp$ as follows. An element $f \in (S/D)^*$ is a function $f : S \to \mathbb{R}$ with $S|_D = 0$. Find any extension of f to a linear function $\bar{f} : X \to \mathbb{R}$ (such an extension exists because S is a retract of X). Then still $\bar{f}|_D = 0$, so $\bar{f} \in D^\perp$. It remains to show that the choice of extension does not matter in the quotient $D^\perp / S^\perp$. Indeed if $\bar{f}_2$ is another extension, then $(\bar{f} - \bar{f}_2)|_S = f - f = 0$, hence $(\bar{f} - \bar{f}_2) \in S^\perp$.

6.3 Conditioning

The proof of the existence of conditionals in LinRel_S proceeds by picking a good complement to the locus of nondeterminism $D \subseteq X \times Y$ as follows:

Lemma 2. *Let $V \subseteq X \times Y$ be a vector subspace, and let $V_X \subseteq X$ be its projection. Then there exists a complement $K \subseteq X \times Y$ of V such that K_X is a complement of V_X.*

Proof. We give an explicit construction, where in fact we can choose K to be a cartesian product of subspaces $U \times W$. Define

$$V_X = \{x : (x, y) \in V\} \qquad H = \{y : (0, y) \in V\}$$

We argue that if $U \oplus V_X = X$ and $W \oplus H = Y$, then $(U \times W) \oplus V = X \times Y$. First we prove that $(U \times W) \cap V = 0$: Indeed, if $(u, w) \in V$ for $u \in U, w \in W$, then $u \in V_X$, but that implies $u = 0$. So we know $(0, w) \in V$, i.e. $w \in H$. Thus $w = 0$.

It remains to show that we can write every (x, y) as $(u + v_1, w + v_2)$ with $u \in U, w \in W$ and $(v_1, v_2) \in V$.

1. We can write $x = u + v_1$ with $u \in U$ and $v_1 \in V_X$.
2. We claim that there exists a $b \in W$ such that $(v_1, b) \in V$. Because $v_1 \in V_X$, there exists some $b' \in Y$ such that $(v_1, b') \in V$. We now decompose $b' = b + h$ for $b \in W, h \in H$. By definition of H, we have $(0, h) \in V$, so $(v_1, b) = (v_1, b') - (0, h) \in V$.
3. Write $y = w' + h$ with $w' \in W, h \in H$ and define $w = w' - b$ and $v_2 = h + b$. Then we have $w \in W$ and $(v_1, v_2) = (v_1, b) + (0, h) \in V$, and as desired

$$(u, w) + (v_1, v_2) = (x, w' - b + h + b) = (x, w' + h) = (x, y).$$

6.4 Composition and Congruence

For the construction of LinRel_S, it remains to check that the relation $\approx$ is a monoidal congruence on $\mathsf{Lin}_{S \times \mathsf{Sub}}$. Recall that $(f, s, U) \approx (g, t, U)$ if and only if $\pi f = \pi g$ and $S(\pi)(s) = S(\pi)(t)$ where $\pi = \pi_{Y/U} : Y \to Y/U$ is the quotient map.

Transitivity: Let $(f, r, V) \approx (g, s, V) \approx (h, t, V)$ meaning $\pi f = \pi g = \pi h$ and $S(\pi)(r) = S(\pi)(s) = S(\pi)(t)$. Then clearly also $(f, r, V) \approx (h, t, V)$.

Congruence: Let $f_i : Y \to Z$, $g_i : X \to Y$ and $U \subseteq Y, V \subseteq Z$ be given for $i = 1, 2$, and assume that $(f_1, r_1, V) \approx (f_2, r_2, V)$ and $(g_1, s_1, U) \approx (g_2, s_2, U)$. We need to show that $(f_1 g_1, r_1 + S(f_1)(s_1), V + f_1[U]) \approx (f_2 g_2, r_2 + S(f_2)(s_2), V + f_2[U])$.

Firstly, we need that $W = V + f_1[U] = V + f_2[U]$ is well-defined. Let $\pi = \pi_{Y/V}$ then by assumption $\pi f_1 = \pi f_2$, so $f_1(y) - f_2(y) \in V$ for all $y \in Y$. Hence

$f_1[U] + V = f_2[U] + V$.

Now, we need to show that $f_1 g_1(x) - f_2 g_2(x) \in W$. We know that $f_1(x) - f_2(x) \in V$ and $g_1(y) - g_2(y) \in U$, hence

$$f_2(g_2(x)) - f_1(g_1(x)) = \underbrace{f_1(g_2(x) - g_1(x))}_{\in f_1[U]} + \underbrace{(f_2(g_2(x)) - f_1(g_2(x)))}_{\in V} \in W$$

For the decorations, we know by assumption that $S(\pi_{Y/U})(s_1) = S(\pi_{Y/U})(s_2)$ and $S(\pi_{Z/V})(r_1) = S(\pi_{Z/V})(r_2)$. We need to show that $S(\pi_{Z/W})(r_1 + S(f_1)(s_1)) = S(\pi_{Z/W})(r_2 + S(f_2)(s_2))$.

The composites $Y \xrightarrow{f_i} Z \xrightarrow{\pi_{Z/W}} Z/W$ are equal for $i = 1, 2$ and vanish on U, thus descending to a map $\tilde{f} : Y/U \to Z/W$, and we obtain

$$\begin{aligned} S(\pi_{Z/W})(S(f_1)(s_1)) &= S(\pi_{Z/W} f_1)(s_1) \\ &= S(\tilde{f}\pi_{Y/U})(s_1) \\ &= S(\tilde{f})(S(\pi_{Y/U})(s_1)) \\ &= S(\tilde{f})(S(\pi_{Y/U})(s_2)) \\ &= S(\tilde{f}\pi_{Y/U})(s_2) \\ &= S(\pi_{Z/W} f_2)(s_2) \\ &= S(\pi_{Z/W})(S(f_2)(s_2)) \end{aligned}$$

The desired proposition now follows easily from the additivity of $S(\pi_{Z/W})(-)$.

Tensor: Let $f_i : X' \to X$, $g_i : Y' \to Y$ and $U \subseteq X, V \subseteq Y$ be given for $i = 1, 2$ and assume $(f_1, s_1, U) \approx (f_2, s_2, U)$ and $(g_1, t_1, V) \approx (g_2, t_2, V)$. We need to show that $(f_1 \times g_1, s_1 \oplus t_1, U \times V) \approx (f_2 \times g_2, s_2 \oplus t_2, U \times V)$.

It is immediate that $(f_1(x), g_1(y)) - (f_2(x), g_2(x)) \in U \times V$ for all $x \in X', y \in Y'$. For the sum of decorations, we chase them around the diagram

$$\begin{array}{ccccccc} X & \xrightarrow{i_X} & X \times Y & \xrightarrow{\pi} & (X \times Y)/(U \times V) & \xleftarrow{\pi} & X \times Y & \xleftarrow{i_Y} & Y \\ \downarrow{\scriptstyle \pi_{X/U}} & & & & \uparrow\wr & & & & \downarrow{\scriptstyle \pi_{Y/V}} \\ X/U & & \xrightarrow[i_1]{} & & (X/U) \times (Y/V) & & \xleftarrow[i_2]{} & & Y/V \end{array}$$

to notice that $S(\pi)(S(i_X)(s_i))$ depends only on $S(\pi_{X/U})(s_i)$.