# Standard Notation

Amos Storkey

January 20, 2014

**Abstract**

This document is used to formulate standard notation used by Amos Storkey in *most* publications and on courses. Things can vary to match others notation in places.

## 1 Notation

Following common use, the term dataset will be used, but will actually mean a family of data points. Any presumption of unorderdering, exchangeability or independence is given by the context and not immediately presumed (for example data from a time series could also be called a dataset).

All random variables will be denoted in sans-serif fonts (or oversize fonts for lowercase Greek symbols) or in serif fonts if the maths is in sans-serif (such as in presentation slides), and the values of such variables in standard font. In a discussion using only univariate quantities, random variables may be upper case, but in general can be lower case. $P(a)$ will be used for the probability of an event $a$. However $P$ will be significantly overloaded. In $P(\mathsf{x})$, $P$ returns the object that is the whole distribution of the random variable $\mathsf{x}$. $P(\mathsf{x} = x)$ denotes the probability value that $\mathsf{x} = x$. Alternatively, and slightly sloppily, for real random variables, $P(\mathsf{x} = x)$ denotes the probability density of variable $\mathsf{x}$ at point $x$. In most cases the underlying random variable for a given value will be implicit. Hence $P(x)$ will be used as a shorthand for $P(\mathsf{x} = x)$ and the existence of a random variable $\mathsf{x}$ distributed according to $\mathsf{x} \sim P(\mathsf{x} = x)$ will be presumed. In general $P$ is an operator on a random variable rather than a simple function, so $P(x)$ and $P(y)$ can be different even if $x = y$ as the $P$ can be operating on different random variables.

Though some might find this notation conflates too many things, it has the advantage of not creating a plethora of different letter notations for each density that is considered, which can make keeping track of which density matches which random variable somewhat cumbersome. It significantly reduces the number of labels needed in most circumstances.

No notational distinction between probabilities and probability densities is made. Which is the case can be ascertained from the nature of the underlying variable. In some cases this distinction could be hard to infer. In these cases

the nature of the variable will be stated. In the case that an equation is valid for either probabilities or probability densities, this will be stated and it will be presumed that any integrals should be exchanged for sums as appropriate. Variables not explicitly or implicitly given distributions in a given context are presumed known in that particular context.

Scalar quantities will be given in normal font $x$, vector quantities in bold lowercase $\mathbf{x}$, and all vectors will be presumed to be column vectors (except where tensorial properties dictate otherwise). Matrix quantities will be in bold uppercase $\mathbf{X}$, and families in uppercase $X$ (which will not be confused with univariate random variables, which will be in lowercase in any situation involving families). By using a vector or matrix format, there is a presumption that linear algebraic methods might be appropriately used on such a quantity, but no particular presumption of any tensorial properties, though in some cases this may hold, and we will presume the context will make that clear. As a result it could be that a particular family of values could be represented either as a family or as a vector, or even a matrix, and we may (explicitly) swap between notations if a vectorial form is more appropriate in a particular context. The family form is the default.

A superscript reference generally selects a family member, and is typically used in the context of a dataset. If the superscript reference is a set or family itself, then it refers to the family obtained by restricting the elements to the superscripted set. When the elements of this restricted family is referred to directly, rather than implicitly via references to the parent family, it is presumed index set will be mapped in order to the standard indices (e.g. $Y = X^D$ implies $Y^1 = X^{D_1}$, $Y^2 = X^{D_2}$ *etc.*, where $D$ is a family or ordered set of integers, or a set of integers whose elements are indexed in ascending order). There is potential for superscript references to be confused with powers. Where such confusion is possible, brackets will be used to indicate raising to a power. E.g. $(x)^2$.

$n = 1, 2, \ldots N$ are used to label training points, and $n^* = 1, 2, \ldots N^*$ to label test points Sometimes $n^+ = 1, 2, \ldots N^+$ may be used. Here $n^+$ counts over the concatenation of training and test points, and $N^+$ is the total number of such points. We will use $\mathcal{D}$ to denote the (training) dataset, $\mathcal{T}$ to denote the test dataset and $\mathcal{M}$ to denote a model. The number of training and test points is given by $N$,$N^*$ respectively. The number of dimensions the data or variable has (i.e. the number of attributes) is denoted by $D$. The dimension of a feature space is given by $M$, whereas the number of classes is given by $m$.

We will use $\Sigma$ to denote a finite covariance matrix, whereas a covariance matrix derived from a Kernel is typically denoted by $K$.