A question on a model via sufficiency

Luca (pglpm0@gmail.com)

Draft of 4 December 2016 (first drafted 3 December 2016)

Notation

Boldface letters: Latin ones are contravariant quantities: lowercase are vectors and uppercase are rank-2 tensors; Greek ones are covariant quantities: lowercase are covectors and uppercase are rank-2 tensors. Juxtaposition of two quantities indicates tensor product if they are of the same transformation kind and contraction if they are of opposite kind: $\boldsymbol{x}\boldsymbol{x} \equiv \boldsymbol{x} \otimes \boldsymbol{x}$ is rank-2 and $\boldsymbol{\alpha}\boldsymbol{x} \equiv \boldsymbol{\alpha} \cdot \boldsymbol{x}$ is a scalar.

Medium letters denote scalars and combinations of quantities of different type, depending on the context

1 Context

I have a keen interest in the method of *modelling via sufficiency*, which in rough terms works as follows. We have the outcomes $\{X_i\}$ of some observations made with the same set-up. Let's call these outcomes our "data". We want to predict the yet unknown outcome X of one more observation, same set-up, given our data and some background knowledge I; that is, we want to give for each X a numerical value to

$$\mathrm{p}(X|\,\mathrm{data}, I). \tag{1}$$

Now suppose we believe or just entertain the hypothesis that for this prediction we only need to know a function – an "aspect", or "feature" – of the data $f(\mathrm{data})$; the rest of the data is irrelevant. That is, calling our hypothesis S,

$$\forall X, \quad \mathrm{p}[X|\,\mathrm{data}, f(\mathrm{data}), S, I] = \mathrm{p}[X|\,f(\mathrm{data}), S, I]. \tag{2}$$

An example is $f(\{X_i\}) = X_1 + X_2 + \cdots$ if the outcomes are numbers. I call S a hypothesis about sufficiency; f is called *sufficient statistics*.

A consequence of the hypothesis is also that

$$\forall \text{data}, \text{data}', \quad \mathrm{p}(\text{data} \mid S, I) = \mathrm{p}(\text{data}' \mid S, I) \text{ if } f(\text{data}) = f(\text{data}'), \tag{2'}$$

that is, data with the same sufficient statistics must have the same probability.

The method of modelling via sufficiency says that our hypothesis then leads to a *unique* expression for $\mathrm{p}(X \mid \text{data}, S, I)$, given our prior guesses about the limit values in the absence of data. In other words, our hypothesis leads to a specific value for this probability.

This method is based on mathematical theorems reviewed in several works [e.g.: 1; 2; 3, § 4.5; 4–6]. These works also show the unique expression for our probability for particular choices of outcome spaces and sufficient statistics. In many important cases the resulting expression is an integral containing distributions of the exponential family, for example the normal, exponential, binomial, or multinomial ones.

2 Question

2.1 Two hypotheses

I am interested in a particular kind of outcomes: each outcome is a pair of a "quality" q and an "indicator" $\boldsymbol{x}$; that is, $X_i = (q_i, \boldsymbol{x}_i)$. The quality is one from of a discrete set, for example $q \in \{\text{green}, \text{red}, \text{blue}\}$, or $q \in \{\text{healthy}, \text{unhealthy}\}$; the indicator is a real-valued vector. The space of outcomes is therefore homeomorphic to $\mathbf{Z}_n \times \mathbf{R}^m$ for some n, m. Let's say for definiteness that there are two qualities called A and B.

First let's define the number of outcomes of each quality, $n_\mathrm{A}, n_\mathrm{B}$, and the averages of the first and second powers of the indicators for each quality, $\overline{\boldsymbol{x}}_\mathrm{A}, \overline{\boldsymbol{x}}_\mathrm{B}, \overline{\boldsymbol{x}\boldsymbol{x}}_\mathrm{A}, \overline{\boldsymbol{x}\boldsymbol{x}}_\mathrm{B}$:

$$\begin{aligned} n_\mathrm{A} &:= \textstyle\sum_i \delta(q_i = \mathrm{A}), & n_\mathrm{B} &:= \textstyle\sum_i \delta(q_i = \mathrm{B}), \\ \overline{\boldsymbol{x}}_\mathrm{A} &:= \textstyle\sum_i \boldsymbol{x}_i \delta(q_i = \mathrm{A})/n_\mathrm{A}, & \overline{\boldsymbol{x}}_\mathrm{B} &:= \textstyle\sum_i \boldsymbol{x}_i \delta(q_i = \mathrm{B})/n_\mathrm{B}, \\ \overline{\boldsymbol{x}\boldsymbol{x}}_\mathrm{A} &:= \textstyle\sum_i \boldsymbol{x}_i\boldsymbol{x}_i \delta(q_i = \mathrm{A})/n_\mathrm{A}, & \overline{\boldsymbol{x}\boldsymbol{x}}_\mathrm{B} &:= \textstyle\sum_i \boldsymbol{x}_i\boldsymbol{x}_i \delta(q_i = \mathrm{B})/n_\mathrm{B}. \end{aligned} \tag{3}$$

I would like to know the form of the distribution $\mathrm{p}[X \mid f(\text{data}), S, I]$ for two choices of sufficient statistics among the quantities above, that is, two different hypotheses about sufficiency S given data $\{(q_i, \boldsymbol{x}_i)\}$:

S_1**:** The sufficient statistics $f_1(\text{data})$ are: the number of outcomes of each quality and the averages of the indicators for each quality. In formulae, $f_1(\text{data}) := \{n_\mathrm{A}, n_\mathrm{B}, \overline{\boldsymbol{x}}_\mathrm{A}, \overline{\boldsymbol{x}}_\mathrm{B}, \overline{\boldsymbol{x}\boldsymbol{x}}_\mathrm{A}, \overline{\boldsymbol{x}\boldsymbol{x}}_\mathrm{B}\}$.

This sufficiency condition is mathematically expressed by

$$\mathrm{p}[(q,\boldsymbol{x})|\,\mathrm{data},S_1,I] = \mathrm{p}[(q,\boldsymbol{x})|\,n_\mathrm{A},n_\mathrm{B},\overline{\boldsymbol{x}}_\mathrm{A},\overline{\boldsymbol{x}}_\mathrm{B},\overline{\boldsymbol{x}\boldsymbol{x}}_\mathrm{A},\overline{\boldsymbol{x}\boldsymbol{x}}_\mathrm{B},S_1,I]. \quad (4)$$

There are other equivalent ways of specifying this statistics, for example by giving the total number of outcomes, the number of those of quality A, and the totals of the indicators for each quality: $\{n_\mathrm{A}+n_\mathrm{B}, n_\mathrm{A}, n_\mathrm{A}\overline{\boldsymbol{x}}_\mathrm{A}, n_\mathrm{B}\overline{\boldsymbol{x}}_\mathrm{B}\}$.

S_2**:** This hypothesis combines sufficiency on two different levels: the probability for the qualities $\{q_i\}$ is exchangeable, and that of the indicators given the qualities is “unrestrictedly exchangeable” [3, § 4.6.2]. In terms of sufficient statistics: the total number of each quality in the data is a sufficient statistics for q:

$$\mathrm{p}(q|\,\mathrm{data},S_2,I) = \mathrm{p}(q|\,\{q_i\},S_2,I) = \mathrm{p}(q|\,n_\mathrm{A},n_\mathrm{B},S_2,I); \quad (5a)$$

and the total of the first and second powers of the indicator related to a particular quality is a sufficient statistics for that indicator, if the qualities are known:

$$\mathrm{p}(\boldsymbol{x}|\,q,\mathrm{data},S_2,I) = \mathrm{p}(\boldsymbol{x}|\,q,n_q,\overline{\boldsymbol{x}}_q,\overline{\boldsymbol{x}\boldsymbol{x}}_q,S_2,I). \quad (5b)$$

The two conditions (5) together imply the same condition (4) of the first hypothesis, but not vice versa. So hypothesis S_2 seems to be a special case of hypothesis S_1.

2.2 Integral form

We calculate the joint probability for an arbitrary sequence $\{(q_i,\boldsymbol{x}_i)\}$ under the two hypotheses. From this all other probabilities can be calculated.

S_1. Following Bernardo & Smith’s discussion on sufficiency and the exponential family [3, § 4.5.3] I assume that hypothesis S_1 leads to the following form for the joint probability:

$$
\begin{aligned}
&\mathrm{p}[\{(q_i, \boldsymbol{x}_i)\} \mid S_1, I] = \\
&\qquad \int \prod_i \Big\{ \tfrac{1}{Z(\theta')} \exp[\alpha_\mathrm{A}\delta(q_i = \mathrm{A}) + \alpha_\mathrm{B}\delta(q_i = \mathrm{B}) + \\
&\qquad\qquad\qquad \boldsymbol{\beta}_\mathrm{A}\boldsymbol{x}_i\delta(q_i = \mathrm{A}) + \boldsymbol{\beta}_\mathrm{B}\boldsymbol{x}_i\delta(q_i = \mathrm{B}) + \\
&\qquad\qquad\qquad \boldsymbol{x}_i\boldsymbol{\Gamma}_\mathrm{A}\boldsymbol{x}_i\delta(q_i = \mathrm{A}) + \boldsymbol{x}_i\boldsymbol{\Gamma}_\mathrm{B}\boldsymbol{x}_i\delta(q_i = \mathrm{B})]\Big\} \times \\
&\qquad\quad \mathrm{p}(\theta' \mid S_1, I)\,\mathrm{d}\theta'
\end{aligned}
\tag{6}
$$

with $\theta' = (\alpha_\mathrm{A}, \alpha_\mathrm{B}, \boldsymbol{\beta}_\mathrm{A}, \boldsymbol{\beta}_\mathrm{B}, \boldsymbol{\Gamma}_\mathrm{A}, \boldsymbol{\Gamma}_\mathrm{B})$ and $Z(\theta')$ a normalization factor for the exponential. Either of the terms α_A or α_B is actually redundant and is eliminated via a Dirac delta in the parameter prior $\mathrm{p}(\theta' \mid S_1, I)$. Compare Definition 4.11 and Proposition 4.13 (p. 201) [3, § 4.5.3]. The mathematical type of the integration variables should be clear from the context.

The reason for the prime in θ' is that we now change to a more convenient set of parameters $\theta = (\nu_\mathrm{A}, \nu_\mathrm{B}, \boldsymbol{m}_\mathrm{A}, \boldsymbol{V}_\mathrm{A}, \boldsymbol{m}_\mathrm{B}, \boldsymbol{V}_\mathrm{B})$:

$$
\begin{aligned}
&\alpha_\mathrm{A} = \ln \frac{\nu_\mathrm{A}}{\nu_\mathrm{A} + \nu_\mathrm{B}} - \tfrac{1}{2}\boldsymbol{m}_\mathrm{A}{\boldsymbol{V}_\mathrm{A}}^{-1}\boldsymbol{m}_\mathrm{A} - \ln\det(2\pi\boldsymbol{V}_\mathrm{A}), \\
&\boldsymbol{\beta}_\mathrm{A} = \boldsymbol{m}_\mathrm{A}{\boldsymbol{V}_\mathrm{A}}^{-1}, \\
&\boldsymbol{\Gamma}_\mathrm{A} = -\tfrac{1}{2}{\boldsymbol{V}_\mathrm{A}}^{-1} \quad \text{(inversion exchanges co- and contravariance)},
\end{aligned}
\tag{7}
$$

and similarly for $(\alpha_\mathrm{B}, \boldsymbol{\beta}_\mathrm{B}, \boldsymbol{\Gamma}_\mathrm{B})$. Combining the JAKOBIAN determinant of the transformation into the parameter prior, and denoting the normal distribution with N, the joint probability now takes the form

$$
\begin{aligned}
\mathrm{p}[\{(q_i, \boldsymbol{x}_i)\} \mid S_1, I] &= \int \prod_i \Big[\nu_\mathrm{A}^{\delta(q_i=\mathrm{A})} N(\boldsymbol{x}_i \mid \boldsymbol{m}_\mathrm{A}, \boldsymbol{V}_\mathrm{A})^{\delta(q_i=\mathrm{A})} \times \\
&\qquad\qquad \nu_\mathrm{B}^{\delta(q_i=\mathrm{B})} N(\boldsymbol{x}_i \mid \boldsymbol{m}_\mathrm{B}, \boldsymbol{V}_\mathrm{B})^{\delta(q_i=\mathrm{B})} \Big] \times \\
&\qquad \mathrm{p}(\nu_\mathrm{A}, \nu_\mathrm{B}, \boldsymbol{m}_\mathrm{A}, \boldsymbol{V}_\mathrm{A}, \boldsymbol{m}_\mathrm{B}, \boldsymbol{V}_\mathrm{B} \mid S_1, I)\,\mathrm{d}\theta \\
&= \int \nu_\mathrm{A}^{n_\mathrm{A}} \prod_i^{q_i=\mathrm{A}} N(\boldsymbol{x}_i \mid \boldsymbol{m}_\mathrm{A}, \boldsymbol{V}_\mathrm{A}) \times \\
&\qquad \nu_\mathrm{B}^{n_\mathrm{B}} \prod_i^{q_i=\mathrm{B}} N(\boldsymbol{x}_i \mid \boldsymbol{m}_\mathrm{B}, \boldsymbol{V}_\mathrm{B}) \times \\
&\qquad \mathrm{p}(\nu_\mathrm{A}, \nu_\mathrm{B}, \boldsymbol{m}_\mathrm{A}, \boldsymbol{V}_\mathrm{A}, \boldsymbol{m}_\mathrm{B}, \boldsymbol{V}_\mathrm{B} \mid S_1, I)\,\mathrm{d}\theta,
\end{aligned}
\tag{8}
$$

with a Dirac delta $\delta(\nu_\mathrm{A} + \nu_\mathrm{B} - 1)$ in the parameter prior. The new normalization factor turns out to be unity. This expression is a mixture of products of Bernoulli distributions for the qualities and normal distributions for the indicators. The normals are of two kinds depending on either quality corresponding to each indicator.

S_2. From Bernardo & Smith's discussion on exchangeability [3, § 4.3], exchangeability for the probability of the qualities $\{q_i\}$, eq. (5a) leads to

$$\begin{aligned} \mathrm{p}(\{q_i\} \mid S_2, I) &= \int \prod_i \Big[\nu_A^{\delta(q_i=A)} \nu_B^{\delta(q_i=B)} \Big] \, \mathrm{p}(\nu_A, \nu_B \mid S_1, I) \, \mathrm{d}\nu_A \, \mathrm{d}\nu_B \\ &= \int \nu_A^{n_A} \nu_B^{n_B} \, \mathrm{p}(\nu_A, \nu_B \mid S_1, I) \, \mathrm{d}\nu_A \, \mathrm{d}\nu_B, \end{aligned} \tag{9}$$

with the usual implicit Dirac delta.

From the discussion on unrestricted exchangeability [3, § 4.6.2], the probability of the indicators $\{\boldsymbol{x}_i\}$ given the corresponding qualities $\{q_i\}$ is

$$\begin{aligned} \mathrm{p}(\{\boldsymbol{x}_i\} \mid \{q_i\}, S_2, I) = \int & \prod_i^{q_i=A} N(\boldsymbol{x}_i \mid \boldsymbol{m}_A, \boldsymbol{V}_A) \times \\ & \mathrm{p}(\boldsymbol{m}_A, \boldsymbol{V}_A \mid S_2, I) \, \mathrm{d}\boldsymbol{m}_A \, \mathrm{d}\boldsymbol{V}_A \times \\ \int & \prod_i^{q_i=B} N(\boldsymbol{x}_i \mid \boldsymbol{m}_B, \boldsymbol{V}_B) \times \\ & \mathrm{p}(\boldsymbol{m}_B, \boldsymbol{V}_B \mid S_2, I) \, \mathrm{d}\boldsymbol{m}_B \, \mathrm{d}\boldsymbol{V}_B. \end{aligned} \tag{10}$$

Multiplying the above two probabilities we find

$$\begin{aligned} \mathrm{p}[\{(q_i, \boldsymbol{x}_i)\} \mid S_2, I] = \int & \nu_A^{n_A} \prod_i^{q_i=A} N(\boldsymbol{x}_i \mid \boldsymbol{m}_A, \boldsymbol{V}_A) \times \\ & \nu_B^{n_B} \prod_i^{q_i=B} N(\boldsymbol{x}_i \mid \boldsymbol{m}_B, \boldsymbol{V}_B) \times \\ & \mathrm{p}(\nu_A, \nu_B \mid S_2, I) \times \\ & \mathrm{p}(\boldsymbol{m}_A, \boldsymbol{V}_A \mid S_2, I) \, \mathrm{p}(\boldsymbol{m}_B, \boldsymbol{V}_B \mid S_2, I) \, \mathrm{d}\theta. \end{aligned} \tag{11}$$

2.3 Comparison

The final integral forms, eqs (8) and (11), of the joint probabilities under the two hypotheses, $\mathrm{p}[\{(q_i, \boldsymbol{x}_i)\} \mid S_1, I]$ and $\mathrm{p}[\{(q_i, \boldsymbol{x}_i)\} \mid S_2, I]$, shows that they only differ in the parameter prior distribution, which factorizes in the second hypothesis S_2. This confirms our previous remark that S_2 seems to be a special case of S_1.

The factorized form of the parameter prior in the second hypothesis is connected to the fact that data about indicators for quality B are irrelevant for prediction of the indicator for quality A, and vice versa. This is clear by constructing that conditional probability from eq. (11). This irrelevance is generally not true under the first hypothesis.

The use of a flat parameter prior (and factorizing boundaries) under the first hypothesis erases its difference from the second one.

2.4 Questions

My questions are these:

1. It is easy to show that the integral forms of eqs (8), (11) for the joint probabilities imply the assumptions (4), (5) of the corresponding hypotheses. Is the converse true?
2. I assumed that the exponential form in the integral also holds for variables of mixed, discrete-continuous character. Is this true? Possibly Barndorff-Nielsen [7] has the answer?
3. A flat parameter prior makes hypothesis S_1 and S_2 equivalent. Does this make sense? Why does a flat prior introduce independence between indicators of different qualities for example?
4. Are the calculations in this note correct?

Bibliography

[1] A. P. Dawid: *Exchangeability and its ramifications*. In: Damien et al. [8] (2013), Ch. 2, 19–29.

[2] P. Diaconis: *Sufficiency as statistical symmetry*. In: Browder [9] (1992), 15–26. First publ. 1991 as technical report https://statistics.stanford.edu/research/sufficiency-statistical-symmetry.

[3] J.-M. Bernardo, A. F. Smith: *Bayesian Theory*, reprint. Wiley, New York (2000). First publ. 1994.

[4] O. Kallenberg: *Probabilistic Symmetries and Invariance Principles*. Springer, New York (2005).

[5] S. L. Lauritzen: *Extremal Families and Systems of Sufficient Statistics*. Springer, Berlin (1988). First publ. 1982.

[6] P. Diaconis, D. Freedman: *Partial exchangeability and sufficiency*. In: Ghosh, Roy [10] (1981), 205–236. Also publ. 1982 as technical report https://www.stat.berkeley.edu/~aldous/206-Exch/Papers/diaconis_freedman_PES.pdf, http://statweb.stanford.edu/~cgates/PERSI/year.html.

[7] O. E. Barndorff-Nielsen: *Information and Exponential Families: In Statistical Theory*, reprint. Wiley, New York (2014). First publ. 1978.

[8] P. Damien et al., eds.: *Bayesian Theory and Applications*. Oxford University Press, Oxford (2013).

[9] F. E. Browder, ed.: *Mathematics into the Twenty-first Century: 1988 Centennial Symposium August 8–12*. American Mathematical Society, Providence, USA (1992).

[10] J. K. Ghosh, J. Roy, eds.: *Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference. Calcutta, 16–19 Dec. 1981*. Indian Statistical Institute, Calcutta (1981).