Bayesian Invariant Measurements of Generalisation for Continuous Distributions

Huaiyu Zhu and Richard Rohwer

NCRG/4352

Technical Report NCRG/4352

August 31, 1995

Neural Computing Research Group
Dept. of Computer Science and Applied Mathematics
Aston University, Aston Triangle
Birmingham B4 7ET, UK

Tel: +44 (0)121 359-3611
Fax: +44 (0)121 333-6215

Bayesian Invariant Measurements of Generalisation for Continuous Distributions

Huaiyu Zhu and Richard Rohwer

Department of Computer Science and Applied Mathematics

Aston University, Aston Triangle, Birmingham B4 7ET

August 31, 1995

Abstract

A family of measurements of generalisation is proposed for estimators of continuous distributions. In particular, they apply to neural network learning rules associated with continuous neural networks. The optimal estimators (learning rules) in this sense are Bayesian decision methods with information divergence as loss function. The Bayesian framework guarantees internal coherence of such measurements, while the information geometric loss function guarantees invariance. The theoretical solution for the optimal estimator is derived by a variational method. It is applied to the family of Gaussian distributions and the implications are discussed.

This is one in a series of technical reports on this topic; it generalises the results of [ZR95a] to continuous distributions and serve as a concrete example of a larger picture [ZR95d].

Contents

1   Introduction

The study of optimal statistical estimation and inference has a long history [Fis22, Fis25, Fis34]. It can be cast in (Bayesian) decision theory [Fer67, DeG70]. For discussions of statistical inference in relation to other statistical methods see also [KS79, CH74, BT73, Ber85].

The problem of optimal statistical estimation has gained more importance recently due to the rapidly expanding research on neural networks. Neural networks can be described as statistical models [Whi89], where a learning rule acts as a statistical estimator and a trained network as an estimate. The task of learning, or any statistical estimation, is to find a most representative distribution among all the distributions which could have possibly generated the observed data. [1] The comparison of two learning rules is therefore a particular example of the comparison of two statistical estimators, and therefore a particular instance of decision theory.

We shall confine our attention to Bayesian decision theory for the following two reasons. First, under some regularity conditions, the set of all "good" (admissible) decision rules essentially coincides with the set of all the Bayesian rules. [2] Therefore any decision rule can be considered as a (either perfect or approximate) Bayesian decision rule. Secondly, many of the classical non-Bayesian theories can be regarded as special examples of Bayesian theory with a particular non-informative prior [Aka80].

There is a further, more philosophical, argument in support of a Bayesian theory. The Bayes Theorem is implied in the mathematical theory of probability [Kol56], which is as logically consistent as arithmetic, and is applicable to any interpretation of probability. All consistent mathematical descriptions of probabilistic models must be Bayesian, in the sense that all the variables, known or unknown, deterministic or random, must be described as random variables, hence having probability distributions. Some of them are selected as

---

[1]This is also true of the learning rules which give the posterior as the result, as will be seen later.

[2]There are many exceptions, but they are more technical than the problems of current concern in neural networks.

observables. The (unconditional) distribution of a random variable is called a prior, while its distribution conditional on the observables is called the posterior. On the other hand, it is known that any definition of reasonable belief of events is equivalent to Kolmogorov's axioms [Cox46], but this argument is not as convincing to someone who does not consider the particular value of a parameter as an event [Fis34]. We shall not pursue this argument any further.

Statistical inference problems are special decision problems in that the "loss function" must characterise the difference between the true distribution and the estimated distribution. Many such loss functions have been proposed over the years, such as the mean squared error proposed by Gauss, the cross entropy (Kullback-Leibler divergence) [KL51] and the Hellinger distance. The advent of neural network models calls for the study of measures of "divergence" between two probability distributions which are invariant with respect to the way the model is represented in order for it to be applied universally, since the weights in neural networks usually do not possess distinct external interpretation apart from being parameters to a family of distributions. We also require that they are invariant with respect to one-one input/output transforms, since we are only concerned with the amount of information captured by the learning algorithm which, unlike the content of information, should not depend on a renaming of the samples. Such information divergences have been studied extensively in the theory of information geometry [Che72, Ama82, Ama85, Ama87], See also [Egu83, BN86, Lau87, CMS93]. For more background see [Hou82, Kas84, BNCR86, Kas87, Kas89]. The main result of information geometry particularly relevant to our current enquiry is that there is a unique one-parameter family of "information divergences" satisfying the above invariance conditions and possessing some important properties making them as important to statistics as the Euclidean distance is to functional analysis.

In [ZR95a], Bayesian decision theory was combined with information divergence to define a family of measurements of generalisation error for discrete distributions. It is coherent in the sense that an optimal estimator thus defined gives optimal estimates for almost all the data. It is invariant of the parameterisation, and input-output space transforms. Explicit results on the problem of estimating multinomial distributions were derived which confirm its intuitive interpretation. This gives a more satisfactory framework than the "empirical error" used by many authors [GBD92], and avoids the "bias/variance trade-off" altogether.

In this report we generalise these results to models with continuous distributions. The optimal estimators are derived using a variational argument so that it is applicable to any family of distributions absolutely continuous with respect to a base measure §2.

This theory is then applied to the family of Gaussian distributions, properties of which are reviewed and collated in §3. The results (§4) coincide with the well known results for this particular family; it shows that least mean squares is the only reasonable method for Gaussian models. The optimal estimates can be represented as t distributions and χ^2 distributions, revealing the Bayesian assumptions behind the classical statistical tests.

3

The results in these two reports are applicable to problems in which the computational model coincides with the mathematical model. A more general and mathematically rigorous development will be given in [ZR95d], allowing the computational model to be a proper submanifold of the mathematical model, as is the case with neural networks.

An application of this theory to the problem of sequentially fitting a regression curve is given in [ZR95c].

Discussions and conclusions are given in §5.

2   General Theory

In this section we shall generalise the results obtained in [ZR95a] to an arbitrary family of distributions dominated by a base measure. This is particularly useful for continuous distributions, which are distributions dominated by the Lebesgue measure. Most of the derivations will be similar to the corresponding ones for discrete distributions [ZR95a], except that instead of taking the derivative of the Lagrange function we need to take a variation of it.

2.1   Bayes Rules

Consider a sample space Y. Let y be a random variable taking values in Y. We are interested estimating the distribution p of y. [3] Now consider the set $\mathcal{P}$ of distributions on Y dominated by a carrier measure (also called a base measure) r, ie. $\mathcal{P}$ is the set of all possible distributions p on Y such that p is absolutely continuous with respect to r. This is equivalent to say that every member of $\mathcal{P}$ can be represented as an indefinite integral of a density function with respect to r. The density function is unique up to a set of r-measure zero, and is called the Radon-Nikodym derivative of p with respect to r [HS49]. In the sequel we shall assume that the carrier measure is fixed so that probability distributions can be identified with density functions. The set $\mathcal{P}$ forms an "infinite dimensional manifold" [ZR95d]. Our task is to infer $p = P(\cdot|p)$ from a sample of y.

In the Bayesian framework, we need to define a prior $P(p)$ over $\mathcal{P}$, which is the unconditional distribution of p. For example, we might define $P(p)$ to be the "uniform distribution over all the Gaussians", in the sense of non-informative prior, which will be explained later. [4] The prior $P(p)$ and the sample y, or, more accurately, the likelihood function $P(y|p)$, are combined by the Bayes formula to give a posterior $P(p|y)$, the distribution of

---

[3]How this is represented is not important at the moment. We can either guess a form of the distribution of y, and estimate its parameters, or we can implement a machine which draws samples from Y with the same distribution as y[Nea93], among many other possibilities.

[4]This is a "uniform distribution over a subspace of $\mathcal{P}$", so the prior is a distribution over distributions. It should not be confused with the unrelated concept of 'the uniform distribution over Y" which is only one particular degenerate Gaussian (with infinite variance).

p conditional on the observation y.

$$(2.1) \qquad\qquad P(p|y) = P(p)P(y|p)/P(y).$$

The posterior is still a distribution of p, not a single-valued estimate of p. What we really need is an estimate $q \in \mathcal{P}$, which is thought to be closest to p, given all the relevant information. A learning method is a mapping $\tau : \; y \rightarrow q = \tau(y) \in \mathcal{P}$, which maps each observed set of data $y \in Y$ to a unique distribution $q \in \mathcal{P}$. The requirement of a single estimator instead of a posterior does not impose any restriction on the general applicability of the theory: On the one hand, all the known applications of the posterior can be transformed into an application of an estimate; On the other, more importantly, the "optimal estimator" together with a few ancillary statistics indicating sample size turns out to be sufficient statistics for the posterior, so that any optimal decisions or inference should be a function of the optimal estimator.

In the terminology of decision theory, we are looking for optimal decision rules τ which will make "good decisions" q, where the objective is that q should approximate the "true parameter" p. Technically, this requires the specification of a "loss function" $D(p, q)$ describing the "divergence" between the two distributions, the posterior mean of which is called the "risk function". An estimator is optimal if it minimises the risk. The loss functions we are interested in are the information divergences developed in the theory of information geometry.

2.2 Information Divergences for Continuous Distributions

We shall use interchangeably the notation of a distribution p with the notation of a conditional distribution $P(\cdot|p)$. This enables us to write the definition of Kullback-Leibler divergence between two distributions p and q as [5]

$$(2.2) \qquad\qquad K(p,q) := \int p \log \frac{p}{q} = \int_{y \in Y} P(y|p) \log \frac{P(y|p)}{P(y|q)}.$$

As in [ZR95a], we use $\delta = (1 - \alpha)/2$ instead of α as in [Ama82, Ama85]. This usage was adopted from [Hou82, Kas84].

Definition 2.1 (δ-divergence) Let $p, q \in \mathcal{P}$. The δ-divergence is defined as,

$$(2.3) \qquad\qquad D_\delta(p,q) := \frac{1}{\delta(1-\delta)} \left(1 - \int p^\delta q^{1-\delta} \right).$$

$$(2.4) \qquad\qquad D_0(p,q) := \lim_{\delta \to 0} D_\delta(p,q), \qquad D_1(p,q) := \lim_{\delta \to 1} D_\delta(p,q),$$

The family of δ-divergences was discovered many times in the history of information theory (See, for example, [AD75, p. 208] and earlier references cited therein) and statistical

---

[5] We adopt the notation := for "defined as" and =: for "denoted by", following [AD75, p. 3].

theories (See [Ama85, p.????] and earlier references cited therein). Special cases of the δ-divergences are well-known [Ama85]

$$(2.5) \qquad\qquad D_0(p,q) = K(q,p),$$

$$(2.6) \qquad\qquad D_{1/2}(p,q) = 2 \int \left(q^{1/2} - p^{1/2} \right)^2,$$

$$(2.7) \qquad\qquad D_1(p,q) = K(p,q).$$

They are the cross entropy, the Hellinger distance, and the reversed cross entropy, respectively.

Information divergences have various important properties ([Ama85, Egu83, Lau87]) which give them the same status in statistics as that enjoyed by the L_p norms in functional analysis. In fact δ-divergence is closely related to the $L_{1/\delta}$ norm [ZR95d].

We shall take the "risk function" defined using any particular "information divergence" as a measurement of "generalisation".

2.3 Generalisation and Optimal Learning Rules

Let Y be a sample space. Let $\mathcal{P}$ be the manifold of a dominated family of probability measures on Y. Let $P(p)$ be a prior over $p \in \mathcal{P}$. The posterior $P(p|y)$, ie. the conditional distribution of p given $y \in Y$, is well defined by the Bayes rule. Let $\tau : Y \to \mathcal{P}$ be a learning rule. Note that both p and q are points on a manifold $\mathcal{P}$ of distributions, while τ itself is a mapping from sample space Y to this manifold of distributions.

Definition 2.2 (Measurement of Generalisation) The estimator error $E_\delta(\tau)$ of τ, and the estimate error $E_\delta(q|y)$ of each estimate $q \in \mathcal{P}$ for a given $y \in Y$, are defined as,

$$(2.8) \qquad\qquad E_\delta(\tau) := \int_{p \in \mathcal{P}} P(p) \int_{y \in Y} P(y|p) D_\delta(p, \tau(y)).$$

$$(2.9) \qquad\qquad E_\delta(q|y) := \int_{p \in \mathcal{P}} P(p|y) D_\delta(p, q).$$

Corollary 2.1 *The error of a learning rule τ is the expected error of estimates it gives, averaged over all possible data.*

$$(2.10) \qquad\qquad E_\delta(\tau) = \int_{y \in Y} P(y) E_\delta(\tau(y)|y).$$

Definition 2.3 (Optimal learning rule) A learning rule τ is called an δ-optimal learning rule (or δ-estimator) if it is the solution of the following optimisation problem

$$(2.11) \qquad\qquad \underset{\tau(y) \in \mathcal{P}}{\text{Min}}\, E_\delta(\tau).$$

Definition 2.4 (Optimal estimate) A probability distribution $q \in \mathcal{P}$ is called an δ-optimal estimate from data y if it is the solution of the following optimisation problem

$$(2.12) \qquad\qquad \underset{q \in \mathcal{P}}{\text{Min}}\, E_\delta(q|y).$$

Theorem 2.2 (Coherence) *A learning rule τ is δ-optimal if and only if for any data y, except a set of probability zero, the result given by the learning rule $\tau(y)$ is a δ-optimal estimate.*

This result in the general decision theory setting is well known in Bayesian decision theory [Fer67, DeG70, Zac71] and is considered to be one of the fundamental advantages of Bayesian methods by advocates of Bayesian statistics [Lor90].

2.4 Optimal learning rules on the whole probability space

Suppose we are allowed to choose a learning rule which gives an arbitrary distribution as an estimate. That is, we consider the whole set of estimators which maps the sample space into the distribution space. It is desirable to find the learning rule which minimises the generalisation error (with respect to a particular prior and a particular information divergence). The solution to this problem can be obtained by the Lagrange multiplier method (See any good textbook on optimisation or variational methods). The only thing which is different from the discrete case discussed in [ZR95a] is that instead of using derivatives, we shall use variations. We first consider $\delta \in (0, 1]$.

Following the coherence theorem, it is only necessary to find optimal estimates for any given data. Consider $y \in Y$. Consider $\mathcal{P}$ as a manifold embedded in the linear space of positive measures. Define the Lagrangian

$$(2.13) \qquad F := E_\delta(q|y) - \lambda \left(\int q - 1 \right)$$

Now we take variations with Δq. From equation (2.3), it can derived that

$$(2.14) \qquad \Delta D_\delta(p, q) = -\frac{1}{\delta} \int \Delta q \left(\frac{p}{q} \right)^\delta .$$

$$(2.15) \qquad \Delta E_\delta(q|y) = -\frac{1}{\delta} \int \Delta q \frac{\left\langle p^\delta \right\rangle_y}{q^\delta} .$$

$$(2.16) \qquad \Delta F = \int \Delta q \left(-\frac{1}{\delta} \frac{\left\langle p^\delta \right\rangle_y}{q^\delta} - \lambda \right) .$$

Therefore $\Delta F = 0$ if and only if

$$q^\delta \sim \left\langle p^\delta \right\rangle_y ,$$

considered as density functions. An analogous derivation for $\delta = 0$ gives similar results, with p^δ replaced by $\log p$. Therefore, we have proved the following theorem.

Theorem 2.3 (δ-optimal estimate) *Let Y be a measurable space. Let $\mathcal{P}$ be the space of all probability measures on Y. Let $P(p)$ be a prior over $\mathcal{P}$. Given a sample $y \in Y$, the*

posterior $P(p|y)$ is given by the Bayes rule. Let $q = \tau_\delta(y) \in \mathcal{P}$ be the δ-optimal estimator of p based on the sample y. Then q is given by

$$(2.17) \qquad q \sim \begin{cases} \left(\int_{p \in \mathcal{P}} P(p|y) p^\delta \right)^{1/\delta}, & \textit{for } \delta > 0, \\ \exp \left(\int_{p \in \mathcal{P}} P(p|y) \log p \right), & \textit{for } \delta = 0. \end{cases}$$

Let us define the δ-coordinate (δ-representation) $\overset{\delta}{l}(p) := p^\delta / \delta$. Denote by $\overset{1/\delta}{l}$ the inverse of $\overset{\delta}{l}$. The above theorem can be expressed more concisely as

$$(2.18) \qquad \tau_\delta(y) \sim \overset{1/\delta}{l} \left(\left\langle \overset{\delta}{l}(p) \right\rangle_y \right).$$

We call the right hand side side the δ-average of p over the posterior $P(p|y)$. This usage is essentially the same as adopted in [AD75, p. 158]. The above two theorems can be summarised as the following: the δ-estimate is the renormalised δ-average over the posterior, and the δ-estimator is an estimator which gives the δ-estimate for almost all the data. If we do not distinguish between estimators which are almost always equal to each other, then the δ-estimator is unique.

The above are derived under the assumption that Δq is free to vary over $\mathcal{P}$. In the case where q is restricted to a smooth manifold $\mathcal{Q}$, we obtain instead that

$$(2.19) \qquad \partial_\theta F = \int \partial_\theta q \left(-\frac{1}{\delta} \frac{\left\langle p^\delta \right\rangle_y}{q^\delta} - \lambda \right) = 0,$$

where θ is a coordinate of the manifold. This implies that $\left\langle p^\delta \right\rangle_y / q^\delta - \lambda$ is normal to $\mathcal{Q}$. The optimisation problem can only be solved by gradient methods.

These results generalise classical results on least squares: The solution in the whole linear space can be obtained analytically by inverting a matrix, while the solution restricted to a submanifold can only be obtained by gradient methods in general. A least squares problem can always be viewed in this way by considering it as estimating the mean of a Gaussian with unit spherical variance. This will be discussed elsewhere [Zhu95].

3   Gaussians With Conjugate Priors

Conjugate priors are of particular importance to Bayesian theories. They were developed in [RS68]. See also [Dem69, DeG70, BT73, KS79, Ber85]. In a Bayesian framework, priors can always be considered as previous information, which may or may not be due to previous statistics. However, "natural conjugate priors" are unique in that they can always be interpreted as information supplied by previous experiments. Non-conjugate priors can

always be approximated by mixture of conjugate priors and explained by hidden variables. For this reason here we shall only consider Gaussians with conjugate priors.

Many of the results collected in this subsection are standard [RS68, Dem69, DeG70]. However, we present them in a unified and concise notion which will be convenient for later development.

3.1 Normal, Gamma t and F distributions

Notation. Denote by $\Gamma(a)$ and $B(a, b)$ the Gamma and Beta functions. Denote $(a)_b := \Gamma(a + b)/\Gamma(a)$.

Definition 3.1 Fix the Lebesgue measure as the carrier measure. The normal distribution is represented by the pdf

$$(3.1) \qquad N(x|b) := (2\pi)^{-\frac{1}{2}} b^{\frac{1}{2}} \exp\left(-\frac{b}{2} x^2\right), \qquad x \in \mathbb{R}.$$

The Gamma distribution is represented by the pdf

$$(3.2) \qquad G(x|m, b) := \frac{1}{\Gamma(m) x} (bx)^m \exp(-bx), \qquad x \in \mathbb{R}_+.$$

The Student's t distribution is represented by the pdf

$$(3.3) \qquad T(x|m, b) := \frac{1}{B(\frac{m}{2}, \frac{1}{2})} \left(\frac{b}{m}\right)^{\frac{1}{2}} \left(1 + \frac{b}{m} x^2\right)^{-\frac{m+1}{2}}, \qquad x \in \mathbb{R}.$$

The F distribution is represented by the pdf

$$(3.4) \qquad F(x|m, n, b) = \frac{1}{B(\frac{m}{2}, \frac{n}{2}) x} \left(\frac{mx}{nb}\right)^{\frac{m}{2}} \left(1 + \frac{mx}{nb}\right)^{-\frac{m+n}{2}}, \qquad x \in \mathbb{R}_+.$$

Notation. Denote

$$(3.5) \qquad G_2(x|m, b) := G\left(x \left| \frac{m}{2}, \frac{m}{2b}\right.\right) = \frac{1}{\Gamma(\frac{m}{2}) x} \left(\frac{mx}{2b}\right)^{\frac{m}{2}} \exp\left(-\frac{mx}{2b}\right).$$

The standard t distribution with m degrees of freedom is $T(x|m, 1)$. The Cauchy distribution is $T(x|1, 1)$. The exponential distribution is $G(x|1, \lambda)$. The χ_m^2 distribution with m degrees of freedom is $G_2(x|m, m)$. The distribution χ_m^2/m is $G_2(x|m, 1)$. The following is also true

$$(3.6) \qquad N(x|b) = \lim_{m \to \infty} T(x|m, b),$$

$$(3.7) \qquad G_2(x|m, b) = \lim_{n \to \infty} F(x|m, n, b),$$

$$(3.8) \qquad N(x|b) = G_2(x^2|1, \frac{1}{b}),$$

$$(3.9) \qquad T(x|m, b) = F(x^2|1, m, \frac{1}{b}).$$

9

Theorem 3.1 *The kth moments of these distributions are given by*

$$(3.10) \qquad \int_x x^k G(x|m, b) = \frac{(m)_k}{b^k}, \qquad \int_x x^k G_2(x|m, b) = \left(\frac{m}{2}\right)_k \left(\frac{2b}{m}\right)^k,$$

$$(3.11) \qquad \int_x x^{2k} T(x|m, b) = \left(\frac{m}{2}\right)_{-k} \left(\frac{1}{2}\right)_k \left(\frac{m}{b}\right)^k,$$

$$(3.12) \qquad \int_x x^{2k} N(x|b) = \left(\frac{1}{2}\right)_k \left(\frac{2}{b}\right)^k.$$

The following two theorems are very useful when Bayes theorem is applied.

Theorem 3.2 *Let* $x, a_1, a_2 \in \mathbb{R}$, $b_1, b_2 \in \mathbb{R}_+$. *Then*

$$(3.13) \qquad N(x - a_1|b_1)N(x - a_2|b_2) = N(x - a_3|b_3)N(a_1 - a_2|b'),$$

where

$$(3.14) \qquad b_3 = b_1 + b_2, \qquad \frac{1}{b'} = \frac{1}{b_1} + \frac{1}{b_2}, \qquad a_3 = \frac{b_1 a_1 + b_2 a_2}{b_1 + b_2}.$$

Theorem 3.3 *Let* $x \in \mathbb{R}$, $h, b, m, n \in \mathbb{R}_+$. *Then*

$$(3.15) \qquad N(x|nh)G_2(h|m, b) = G_2(h|m_1, b_1)T(x|m, nb),$$

where

$$(3.16) \qquad m_1 = m + 1, \qquad \frac{m_1}{b_1} = \frac{m}{b} + nx^2.$$

The following results are useful for deriving δ-optimal estimators.

Theorem 3.4 *Let* $x \in \mathbb{R}$, $m, b \in \mathbb{R}_+$, $\delta \in (0, 1)$. *Then*

$$(3.17) \qquad N(x|b)^\delta \sim b^{\frac{\delta-1}{2}} N(x|b\delta),$$

$$(3.18) \qquad G(x|m, b)^\delta \sim G(x|m, b\delta),$$

$$(3.19) \qquad x^\delta G(x|m, b) = (m)_\delta b^{-\delta} G(x|m + \delta, b),$$

$$(3.20) \qquad x^\delta G_2(x|m, b) = \left(\frac{m}{2}\right)_\delta \left(\frac{2b}{m}\right)^\delta G_2(x|m + 2\delta, (1 + 2\delta/m)b),$$

$$(3.21) \qquad T(x|m, b)^\delta \sim T(x|\delta_1 m, \delta_1 b),$$

where $\delta_1 := \delta + (\delta - 1)/m$,

$$(3.22) \qquad cT(cx|m, b) = T(x|m, bc^2), \qquad cN(cx|b) = N(x|bc^2).$$

10

3.2 Natural conjugate prior of Gaussians

The following notation makes it easy to refer to joint, marginal, and conditional distributions at the same time.

Notation. Suppose $a_1, \ldots, a_n$ are n easily distinguished factors. Suppose the same holds for $b_1, \ldots, b_n$. We use the notation $a_1 \cdots a_n \overset{n}{=} b_1 \cdots b_n$ to denote that $a_i = b_i$ for all i, and the whole expression is also used to denote the product of these factors.

Let f denote probability density function of random variables x, y, z. Then,

$$(3.23) \qquad f(x; y; z) :\overset{3}{=} f(x|y, z)f(y|z)f(z).$$

The notation $f(x; y; z)$ simultaneously supplies expressions for $f(x, y, z)$, $f(x, y|z)$, $f(x|y, z)$, $f(y, z)$, $f(y|z)$ and $f(z)$. Analogous notation can also defined for any number of variables.

Theorem 3.5 *If any one of the expressions (3.24–3.29) for joint distribution of x, μ, h is true, then all the identities implied (total 30) in the whole formula are true.*

$$(3.24) \qquad f(x; \mu; h) \overset{3}{=} N(x - \mu|h)\, N(\mu - a|nh)\, G_2(h|m, b)$$

$$(3.25) \qquad = f(\mu; x; h) \overset{3}{=} N(\mu - a_1|n_1 h)\, N(x - a|n'h)\, G_2(h|m, b)$$

$$(3.26) \qquad = f(\mu; h; x) \overset{3}{=} N(\mu - a_1|n_1 h)\, G_2(h|m_1, b_1)\, T(x - a|m, n'b)$$

$$(3.27) \qquad = f(h; \mu; x) \overset{3}{=} G_2(h|m_2, b_2)\, T(\mu - a_1|m_1, n_1 b_1)\, T(x - a|m, n'b)$$

$$(3.28) \qquad = f(h; x; \mu) \overset{3}{=} G_2(h|m_2, b_2)\, T(x - \mu|m_1, b_1')\, T(\mu - a|m, nb)$$

$$(3.29) \qquad = f(x; h; \mu) \overset{3}{=} N(x - \mu|h)\, G_2(h|m_1, b_1')\, T(\mu - a|m, nb),$$

where

$$(3.30) \qquad m_1 = m + 1, \qquad m_2 = m_1 + 1,$$

$$(3.31) \qquad n_1 = n + 1, \qquad n' = n/n_1,$$

$$(3.32) \qquad a_1 = \frac{na + x}{n + 1} = a + \frac{x - a}{n_1}.$$

$$(3.33) \qquad \frac{1}{b_1'} = \frac{1}{m_1}\left(\frac{m}{b} + n(\mu - a)^2\right),$$

$$(3.34) \qquad \frac{1}{b_1} = \frac{1}{m_1}\left(\frac{m}{b} + n'(x - a)^2\right),$$

$$(3.35) \qquad \frac{1}{b_2} = \frac{1}{m_2}\left(\frac{m_1}{b_1} + n_1(\mu - a_1)^2\right) = \frac{1}{m_2}\left(\frac{m_1}{b_1'} + (x - \mu)^2\right).$$

Proof: The equivalence of every two consecutive lines can be proved by repeated application of Theorem 3.2 and Theorem 3.3. The identities implied are six of the form

$f(x, y, z)$, six of the form $f(x, y|z)$, three of the form $f(x|y, z)$, six of the form $f(y, z)$, six of the form $f(y|z)$, and three of the form $f(z)$. $\qquad\qquad\qquad\qquad\square$

This theorem summarises all that is known about functional relationships between density functions for univariate Gaussians with natural conjugate priors, where the likelihood is

$$(3.36) \qquad\qquad f(x|\mu, h) = N(x - \mu|h),$$

and the prior is

$$(3.37) \qquad\qquad f(\mu, h) = N(\mu - a|nh) \, G_2(h|m, b).$$

All the prior, posterior, joint, conditional and marginal distributions can be directly read off these identities. Many of these identities appeared in the literature, but it appears that not all of them have previously been collected together.

The equations (3.34), (3.33) and (3.35) can be alternatively represented in the following useful form.

$$(3.38) \qquad\qquad \frac{m_1}{b_1} = \frac{m}{b} + n'(x - a)^2,$$

$$(3.39) \qquad\qquad \frac{m_1}{b_1'} = \frac{m}{b} + n(\mu - a)^2,$$

$$(3.40) \qquad\qquad \frac{m_2}{b_2} = \frac{m_1}{b_1} + n_1(\mu - a_1)^2 = \frac{m_1}{b_1'} + (x - \mu)^2.$$

The case of Gaussians with fixed variance is of special interest since it is the basis of least mean square methods. This can be obtained by setting $m = \infty$ in Theorem 3.5, which implies that

$$(3.41) \qquad\qquad b_1, b_2, b_1' \rightarrow b,$$

$$(3.42) \qquad\qquad G_2(h|m, b) \rightarrow \delta(h - b),$$

$$(3.43) \qquad\qquad T(x|m, h) \rightarrow N(x|h),$$

and Theorem 3.5 reduces a corresponding theorem for fixed h.

Theorem 3.6 *If either of the expressions (3.44–3.45) for joint distribution of x, μ is true, then all the identities (total 6) implied in the whole formula is true.*

$$(3.44) \qquad\qquad f(x; \mu) = N(x - \mu|h) \, N(\mu - a|nh)$$

$$(3.45) \qquad\qquad f(\mu; x) = N(\mu - a_1|n_1 h) \, N(x - a|n'h).$$

3.3 Sufficient statistics for Gaussians

The family of Gaussian distributions with unknown mean μ and precision h has a two dimensional sufficient statistic $[a, b]$, with corresponding ancillary statistic $[n, m]$.

A sufficient statistic summarises all the information in the sample. The remaining variations within a sample can be considered as pure noise. An ancillary statistic does not contain any information about the sample. It instead contains information about the experimental settings. In this context, the ancillary statistic describes the sample size.

The concept of sufficient statistics was proposed and studied by Fisher [Fis22, Fis25, Fis34]. It was shown that under suitable regularity conditions the existence of sufficient statistics of fixed dimension is the defining property of exponential family [Dar35, Koo36, Pit36]. Rigorous measure theoretical definition of sufficient statistics was given by [HS49]. A whole chapter of [Zac71] is devoted to sufficient statistics. Ancillary statistics was first studied by [Fis34] and clarified in [Ama85, Ama87]. See also [KL51, Fra56, Fra63, CH74, KS79, Kas89].

In the case of Gaussians, there are many different formulas to update the statistics. They have been comprehensively documented since [RS68, Dem69, DeG70]. Here we shall adopt a particular set of notations which is capable to summarise these results in a more systematic way, which is also useful for our later developments. Note that for conjugate priors it is also meaningful to talk about the sufficient statistics for the prior, the likeliwhichhood, and the posterior. The classical meaning of sufficient statistics is that of likelihood, whs meaningful for both Bayesian and non-Bayesian theories. Our usage of the term will always be clear from the context.

Notation. With reference to Theorem 3.5. Denote

$$(3.46) \qquad f(\mu; h | n, m, a, b) :\stackrel{2}{=} N(\mu - a | nh) \, G_2(h | m, b),$$

interpreted as specifying the prior. Let $x^k := [x_1, \ldots, x_k]$ be a sample of size k. The posterior is denoted as

$$(3.47) \qquad f(\mu; h | n, m, a, b, x^k).$$

Theorem 3.7 *Consider the likelihood $f(x | \mu, h)$ with prior $f(\mu; h | n, m, a, b)$. Then for a sample x^k, the posterior is given by*

$$(3.48) \qquad f(\mu; h | n, m, a, b, x^k) = f(\mu; h | n_k, m_k, a_k, b_k).$$

where

$$(3.49) \qquad n_k = n + k,$$
$$(3.50) \qquad m_k = m + k,$$
$$(3.51) \qquad n_k a_k = na + \sum x,$$
$$(3.52) \qquad n_k a_k^2 + m_k / b_k = na^2 + m/b + \sum x^2.$$

Proof: It is easy to see from the expressions for $f(\mu; h)$ and $f(\mu; h | x)$ in Theorem 3.5, that after sampling one point, the posterior is expressed in exactly the same form as the

prior, with the parameters updated as follows

$$(3.53) \qquad n_1 = n + 1,$$

$$(3.54) \qquad m_1 = m + 1,$$

$$(3.55) \qquad n_1 a_1 = na + x,$$

$$(3.56) \qquad n_1 a_1^2 + m_1/b_1 = na^2 + m/b + x^2.$$

This can obviously be generalised to a sample of size k by applying mathematical induction on k. $\qquad \square$

Lemma 3.8 *With the notations defined as in Theorem 3.5, the following holds,*

$$\frac{x - a_1}{1} = \frac{a_1 - a}{n} = \frac{x - a}{n + 1}.$$

Lemma 3.9 *With the notations defined as in Theorem 3.5, the following holds,*

$$na^2 + x^2 - n_1 a_1^2 = \frac{n}{n_1}(x - a)^2 = (x - a)(a_1 - a)$$

$$= n(x - a_1)(x - a) = n_1(x - a_1)(a_1 - a)$$

$$= \frac{n_1}{n}(a_1 - a)^2 = nn_1(x - a_1)^2.$$

Corollary 3.10 *The updating rule for b_k can also be written in the following forms,*

$$m_k/b_k - m/b = na^2 + \sum x^2 - n_k a_k^2$$

$$= \frac{n}{n + k} \sum (x - a)^2 + \frac{k}{n + k} \sum (x - \overline{x})^2$$

$$= \frac{n}{n + k} \left(a^2 - a_k^2 \right) + \frac{k}{n + k} \left(\overline{x^2} - a_k^2 \right)$$

$$= \frac{n}{n + k} \left(a - a_k \right)^2 + \sum (x - a_k)^2$$

$$= \frac{nk}{n + k} \left(\overline{x} - a \right)^2 + k \left(\overline{x^2} - \overline{x}^2 \right).$$

Many of these updating rules are used in different text books, but they do not appear to have previously been collected in a single formula.

The updating rule clearly shows that given n_k and m_k, the statistics a_k and b_k are sufficient, with the following intuitive interpretation. The ancillary statistics n_k and m_k are the amount of information with regard to μ and h, respectively, measured in the unit of sample size. The sufficient statistics a_k and b_k are the contents of the information. The posterior information about μ is the algebraic sum of the prior and sample information, weighted by the ancillary. The posterior information about h is not expressible as a simple sum of the prior and sample information, since the information about μ is also involved. It is easy to see that the posterior information about $\langle x^2 \rangle = \mu^2 + 1/h$ is an algebraic sum of prior and sample information.

Theorem 3.11 *For any given parameter μ and h, the statistics a_k and b_k are consistent estimators of μ and h. That is, $a_k \to \mu$ and $b_k \to h$, as $k \to \infty$.*

Theorem 3.12 *For any given sample x^k ,*

$$(3.57) \qquad \langle \mu \rangle_k = a_k, \qquad \left\langle \mu^2 \right\rangle_k = \frac{1}{1 - \frac{2}{m_k}} \frac{1}{n_k b_k} = \frac{\langle \sigma^2 \rangle}{n_k}.$$

$$(3.58) \qquad \langle h \rangle_k = b_k, \qquad \left\langle h^2 \right\rangle_k = \left(1 + \frac{2}{m_k} \right) b_k^2,$$

$$(3.59) \qquad \langle h, h \rangle_k = \frac{2 b_k^2}{m_k}, \qquad \left\langle \sigma^2 \right\rangle_k = \langle 1/h \rangle_k = \frac{1}{b_k} \frac{1}{1 - \frac{2}{m_k}}.$$

The posterior variances of μ and h are of orders $1/n_k$ and $1/m_k$, respectively.

The non-informative prior [Jef61, BT73, DeG70] is given by $m = 0$, $n = 0$. Correspondingly, the ancillary and sufficient statistics are

$$(3.60) \qquad n_k = k,$$
$$(3.61) \qquad m_k = k,$$
$$(3.62) \qquad a_k = \overline{x},$$
$$(3.63) \qquad 1/b_k = \overline{x^2} - \overline{x}^2 = \overline{(x - \overline{x})^2}.$$

The statistics a_k and b_k are exactly the maximum likelihood estimates of μ and h.

3.4 Information Divergence between Two Gaussians

This subsection is intended to provide some concrete examples illustrating the meaning of information divergence. Let $p_i \sim N(\mu_i | h_i)$, $i \in \{1, 2\}$ be two Gaussian distributions. Denote

$$(3.64) \qquad d_0(h_1, h_2) := \left(\frac{h_1^\delta h_2^{1-\delta}}{\delta h_1 + (1 - \delta) h_2} \right)^{1/2},$$

$$(3.65) \qquad d_1(\mu | h) := \exp \left(-\frac{h}{2} \mu^2 \right).$$

It follows that

$$(3.66) \qquad 0 \le d_0 \le 1, \qquad 0 \le d_1 \le 1,$$

and

$$(3.67) \qquad d_0(h_1, h_2) = 1 \iff h_1 = h_2.$$
$$(3.68) \qquad d_1(\mu | h) = 1 \iff \mu = 0.$$
$$(3.69)$$

15

It is also easy to verify that

$$(3.70) \qquad \int p_1^\delta p_2^{1-\delta} = d_0(h_1, h_2)d_1(\mu_1 - \mu_2|H),$$

where

$$(3.71) \qquad \frac{1}{H} = \frac{1}{\delta h_1} + \frac{1}{(1-\delta)h_2}.$$

Therefore the δ-divergence is given by

$$(3.72) \qquad D_\delta(p_1, p_2) = \frac{1}{\delta(1-\delta)}\Big(1 - d_0(h_1, h_2)d_1(\mu_1 - \mu_2|H)\Big).$$

It vanishes if and only if $h_1 = h_2$, $\mu_1 = \mu_2$. The special forms of d_0 and d_1 have intimate connections with the tests of similarity between two Gaussian samples.

When $\mu_1 = \mu_2 = \mu$,

$$(3.73) \qquad D_\delta(p_1, p_2) = \frac{1}{\delta(1-\delta)}\left(1 - d_0(h_1, h_2)\right),$$

independent of μ. When $h_1 = h_2 = h$,

$$(3.74) \qquad D_\delta(p_1, p_2) = \frac{1}{\delta(1-\delta)}\left(1 - d_1(\mu_1 - \mu_2|H)\right),$$

where $H = \delta(1-\delta)h$. This is a monotonic function of the squared difference between μ_1 and μ_2, and was considered in [LS72]. This is also true in general multidimensional case. Therefore mean square theory of approximation is a special case of the present theory by identifying the inner product with the information matrix. The details are omitted.

The two extreme cases, $\delta = 0$ and $\delta = 1$, correspond to the cross entropy (KL distance) and the reversed cross entropy, which play important roles in information theory. Let $\delta \to 0$. Then

$$(3.75) \qquad d_0(h_1, h_2) = \left(\frac{\left(\frac{h_1}{h_2}\right)^\delta}{1 + \delta\left(\frac{h_1}{h_2} - 1\right)}\right) \to \left(\frac{h_1}{h_2}\right)^{\delta/2} \exp\left(\frac{\delta}{2}\left(\frac{h_1}{h_2} - 1\right)\right).$$

Therefore, as $\delta \to 0$,

$$(3.76) \qquad -\frac{1}{\delta}\log d_0(h_1, h_2) \to \frac{1}{2}\left(-\log\frac{h_1}{h_2} + \frac{h_1}{h_2} - 1\right),$$

$$(3.77) \qquad -\frac{1}{\delta}\log d_1(\mu_1 - \mu_2|H) \to \frac{h_1}{2}(\mu_1 - \mu_2)^2,$$

$$(3.78) \qquad D_0(p_1, p_2) \to -\frac{1}{\delta}\left(\log d_0(h_1, h_2) + \log d_1(\mu_1 - \mu_2|H)\right)$$

$$\to \frac{1}{2}\left(\log\frac{h_2}{h_1} + \frac{h_1}{h_2} - 1 + h_1(\mu_1 - \mu_2)^2\right),$$

16

This is of course the reverse cross entropy $K(p_2, p_1)$ between p_1 and p_2. Similarly, $D_1(p_1, p_2)$ is the cross entropy $K(p_1, p_2)$ between p_1 and p_2.

When p_1 and p_2 are close to each other

$$(3.79) \qquad\qquad \mu_1, \mu_2 \approx \mu, \qquad h_1, h_2 \approx h,$$

all the δ-divergences approach the quadratic form defined by Fisher information matrix,

$$(3.80) \qquad\qquad D_\delta(p_1, p_2) \approx \frac{1}{4}\left(\frac{\Delta h}{h}\right)^2 + \frac{h}{2}(\Delta \mu)^2 .$$

This formula can be given directly by simple differential geometry calculations (See [Ama85, Ex.2.3]).

4 Optimal Estimators for Gaussians

We derive the δ-optimal estimators for the families of Gaussian distributions, by applying the general theory of δ-optimal estimators to the explicit formulas for sufficient statistics for Gaussians. We consider separately the cases of fixed or variable variances, although the former is a limiting case of the latter.

4.1 When both the mean and the variance are unknown

The δ-optimal estimates of Gaussian distributions with natural conjugate priors are summarised in the following theorem. It connects the concepts of maximum likelihood estimator, the t test and χ^2 tests, the sufficient statistics, and the conjugate priors in a straight-forward manner.

Theorem 4.1 *Consider the class of Gaussians with conjugate priors $f(\mu; h | n, m, a, b)$. Let $x^k = [x_1, \ldots, x_k]$ be a sample of size k. Then the δ-optimal estimate $q = \tau_\delta(n, m, a, b, x^k)$ is given by the pdf*

$$(4.1) \qquad\qquad f(y|q) = f_\delta(y|x^k) = T\left(y - a_k \left| \frac{m_k}{\delta}, \frac{b_k}{1 + \delta/n_k} \right.\right) .$$

Proof: It follows from Theorem 2.3, Theorem 3.5 and Theorem 3.4 that

$$
\begin{aligned}
f_\delta(y|x^k)^\delta &\sim \int_{\mu,h} f(y|\mu,h)^\delta \, f(\mu;h|x^k) \\
&= \int_{\mu,h} N(y-\mu|h)^\delta \, N(\mu-a_k|n_k h) \, G_2(h|m_k,b_k) \\
&\sim \int_{\mu,h} N(y-\mu|h\delta)h^{(\delta-1)/2} \, N(\mu-a_k|n_k h) \, G_2(h|m_k,b_k) \\
&\sim \int_{\mu,h} N(\mu-*|*) \, N(y-a_k|n_\delta h) \, G_2\left(h\middle|m_{k\delta},\frac{m_{k\delta}}{m_k}b_k\right) \\
&= \int_h N(y-a_k|n_{k\delta}h) \, G_2\left(h\middle|m_{k\delta},\frac{m_{k\delta}}{m_k}b_k\right) \\
&= \int_h G_2(h|*,*) \, T\left(y-a_k\middle|m_{k\delta},\frac{m_{k\delta}n_{k\delta}}{m_k}b_k\right) \\
&= T\left(y-a_k\middle|m_{k\delta},\frac{m_{k\delta}n_{k\delta}}{m_k}b_k\right) \\
&\sim T\left(y-a_k\middle|\frac{m_k}{\delta},\frac{b_k}{1+\delta/n_k}\right)^\delta,
\end{aligned}
$$

where

$$
\frac{1}{n_{k\delta}} := \frac{1}{n_k} + \frac{1}{\delta}, \qquad m_{k\delta} := m_k + \delta - 1.
$$

This completes the proof, by noting that the t distribution is normalised. $\qquad\square$

Corollary 4.2 *The following limits hold for the δ-optimal estimates*

$$(4.2) \qquad \lim_{\delta\to 0} f_\delta(y|x^k) = N(y-a_k|b_k).$$

$$(4.3) \qquad \lim_{\delta\to 1} f_\delta(y|x^k) = T\left(y-a_k\middle|m_k,\frac{b_k}{1+1/n_k}\right).$$

$$(4.4) \qquad \lim_{n\to\infty} f_\delta(y|x^k) = T\left(y-a_0\middle|\frac{m_k}{\delta},b_k\right). \qquad \lim_{n\to\infty} a_k = a_0.$$

$$(4.5) \qquad \lim_{m\to\infty} f_\delta(y|x^k) = N\left(y-a_k\middle|\frac{b_0}{1+\delta/n_k}\right). \qquad \lim_{m\to\infty} b_k = b_0.$$

Several observations can be made from these results:

- The δ-optimal estimate incorporates all the sufficient statistics a_k and b_k, given the ancillary statistics n_k and m_k. They are "minimum sufficient", or "necessary and sufficient".

- For $\delta = 0$, the 0-optimal estimate $N(y - a_k | b_k)$ is the posterior optimal estimate in the sense of classical statistics. The δ-estimators for $\delta > 0$ do not map to the original manifold (the support of the prior).

- Assuming the non-informative prior $n_0 = 0$, $m_0 = 0$, the 0-optimal estimator is exactly the maximum likelihood estimator, $a_k = \overline{x}$, $1/b_k = \overline{x^2} - (\overline{x})^2$. [6]

- For $\delta > 0$, the ancillary statistic m_k is also given by the estimator. However, the sufficient statistic b_k is merged with n_k. Therefore, the estimator is a sufficient statistic only when n_k is known. This is a simple example showing the necessity of ancillary statistics for certain kinds of estimators.

- All the δ-optimal estimators are obtained by averaging the δ-coordinates and renormalisation. For $\delta = 1$ the renormalisation is redundant since the 1-optimal estimator is simply the posterior marginal distribution. [7]

The limit $m_k \to \infty$ directly gives results for Gaussians with known variance. Since this is a very common special case, we shall give explicit formulas in the next subsection.

4.2 Optimal estimator when the variance is known

As a special case, which is intimately related to the practice of least mean squares, let us consider estimating Gaussians with fixed variance. This situation is equivalent to setting $m = \infty$ in Theorem 3.5. We shall give direct proofs since this case is of particular interest, and the proof is much simpler than that of the case with unknown variance.

Theorem 4.3 *Consider the class of Gaussian distributions of fixed variance, with natural conjugate priors as specified by Theorem 3.6. Let $x^k := [x_1, \ldots, x_k]$ be a sample of size k. Then the δ-optimal estimate is given by the density*

$$(4.6) \qquad\qquad f_\delta(y | x^k) = N\left(y - a_k \,\middle|\, \frac{h}{1 + \delta/n_k}\right).$$

---

[6] It is an interesting conjecture that most of the nice properties attributed to ML estimators, most of them discovered by Fisher in the early half of this century, are also present for any δ-optimal estimators with ϵ-uniform priors.

[7] It is effectively the distribution used by the Monte Carlo methods proposed by [Nea95].

Proof: Use Theorem 2.3, Theorem 3.6 and Theorem 3.4,

$$
\begin{aligned}
f_\delta(y|x^k)^\delta &\sim \int_\mu f(y|\mu)^\delta f(\mu|x^k) \\
&= \int_\mu N(y-\mu|h)^\delta N(\mu-a_k|n_k h) \\
&\sim \int_\mu N(y-\mu|\delta h) N(\mu-a_k|n_k h) \\
&= N(y-a_k|n_{k\delta} h), \\
&\sim N\left(y-a_k\left|\frac{h}{1+\delta/n_k}\right.\right)^\delta,
\end{aligned}
$$

where $n_{k\delta}$ is given by

$$
\frac{1}{n_{k\delta}} := \frac{1}{n_k} + \frac{1}{\delta}.
$$

This completes the proof. □

Corollary 4.4 *Assuming the non-informative prior $m = 0$, $n = 0$, the δ-optimal estimate is*

(4.7)
$$
f_\delta(y|x^k) = N\left(y-\overline{x}\left|h/\left(1+\delta/k\right)\right.\right).
$$

Two special cases are of particular interest.

- When $\delta = 0$, the optimal estimate $N(y-\overline{x}|h)$ is the conditional distribution with maximum likelihood estimate.

- When $\delta = 1$, the optimal estimate $N(y-\overline{x}|h/(1+1/k))$ is the posterior marginal distribution.

Therefore, the δ-optimal estimators provide a smooth interpolation between "optimising" and "integrating" over the posterior. All the δ-optimal estimates are internally coherent, and invariant with respect to parameter and sample space transforms.

It is interesting to note that the δ-optimal estimates are not members of the model described by the prior unless $\delta = 0$. The situation for $\delta = 0$ is different since the th Gaussian family, being an exponential family, is 0-flat.

On the other hand, the projection of these optimal estimates onto the allowed family coincide with the 0-optimal estimate. This means that the δ-optimal estimates restricted to the allowed manifold are identical to each other for any δ, and are identical to the maximum likelihood estimator. This provides a justification for using least mean squares to estimate Gaussian means. This also suggests that least mean square estimates for other

distributions, for which squared distances between the parameters are not equivalent to the divergences between distributions, are not expected to have the similarly nice properties.

It can be shown, using the formulas derived so far, that asymptotically, ie. as $k \to \infty$,

$$(4.8) \qquad\qquad E_\delta(\tau_\delta) = \left\langle D_\delta(p, \tau_\delta(x^k)) \right\rangle_k \approx \frac{1}{k}.$$

We conjecture that a similar result also holds for the δ-estimate of a Gaussian with both unknown mean and variance. We have not verified this conjecture since it appears to require an explicit formula for the δ-divergence between the normal and t distributions, the derivation of which is likely to be quite tedious.

5 Conclusions and discussions

We have shown that it is possible to define a generalisation measure which enables selection from the Bayes posterior a unique representative which is optimal in the sense of information geometry. It is shown that the δ-optimal estimates are characterised by the fact that their δ-coordinates are proportional to the posterior expectation of the δ-coordinates of the true distributions. This is shown to be true for an arbitrary dominated family of distributions, either continuous or discrete, or mixed.

The explicit formulas for δ-optimal estimators are given for normal distributions as examples. The δ-optimal estimators are sufficient statistics. The 0-optimal estimate coincides with the optimal Bayesian estimate. With a non-informative prior, this reduces to the distribution with the maximum likelihood estimate of its parameters. The 1-optimal estimate is the posterior marginal distribution.

The maximum likelihood estimators in general are invariant with respect to the parameterisation. This is quite clear in this special case, since the δ-optimal estimates are invariant. The least mean square estimates, or the unbiasedness of an estimator, are not invariant. In this special case, it can be seen that the least mean square estimate is equivalent to the δ-optimal estimate only for the mean of a Gaussian.

These results show that it is possible to use invariant information theoretic criteria in conjunction with a Bayesian theory. Learning rules can therefore be viewed as approximation to the δ-optimal estimators. This unifies classical results based on various different points of view, showing their intrinsic relations.

It would be of particular interest to continue this research into the optimal estimation of multivariate Gaussians, the Bayesian updating formulas are readily available [Dem69, DeG70]. This will be important in its own right, since it will provide a basis for least mean squares and the regressions. It will also be of theoretical importance since it will provide a means of unification with the theory of function approximation in Hilbert spaces. Finally, it will offer some insight to theoretical statistics in general in light of the asymptotic normality of most estimators.

Even more interesting would be direct applications to non-Gaussian continuous distributions. For exponential families, which are exactly the distribution families admitting a sufficient statistic of fixed dimension [Dar35, Koo36, Pit36], the procedure would be quite similar. For curved exponential families [Efr75], one of the possible routes of exploration is to assume an exponentially-uniform prior on the whole exponential family, and consider the δ-estimates restricted to the curved submanifold [Ama85]. It is expected that this will reduce to the results of information geometry for curved exponential families [Ama85, Ama87], but anomalies might appear because of the improper prior [DSZ73]. At the moment, we are unclear as to what kind of results can be derived for families not admitting fixed finite dimensional sufficient statistics, although one way of approximating the optimal estimator in the Gaussian (generalised) linear regression context is developed in [ZR95b].

In some applications, specific error functions might be used in place of the δ-divergence. They are, however, dependent on the given problem which assigns specific meaning to the parameters or the sample points. Since the δ-estimates are sufficient statistics, an optimal estimator in such special cases must be functions of τ_δ. It is itself a sufficient statistic if and only if the function is invertible.

References

[ABNK+87] S. Amari, O. E. Barndoff-Nieldon, R. E. Kass, S. L. Lauritzen, and C. R Rao, editors. *Differential Geometry in Statistical Inference*, volume 10 of *IMS Lecture Notes Monograph*. Inst. Math. Stat., Hayward, CA, 1987.

[AD75] J. Aczél and Z. Daróczy. *On measures of information and their characteristics*. Academic Press, New York, 1975.

[Aka80] H. Akaike. The interpretation of improper prior distributions as limits of data dependent proper prior distributions. *J. Roy. Stat. Soc., B*, 42(1):46–52, 1980.

[Ama82] S. Amari. Differential geometry of curved exponential families—curvature and information loss. *Ann. Stat.*, 10(2):357–385, 1982.

[Ama85] S. Amari. *Differential-Geometrical Methods in Statistics*, volume 28 of *Springer Lecture Notes in Statistics*. Springer-Verlag, New York, 1985.

[Ama87] S. Amari. Differential geometrical theory of statistics. In Amari et al. [ABNK+87], chapter 2, pages 19–94.

[Ber85]   J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 1985.

[BN86]   O. E. Barndorff-Nielsen. Likelihood and observed geometries. *Ann. Statist.*, 14(3):856–873, 1986.

[BNCR86]   O. E. Barndorff-Nielsen, D. R. Cox, and N. Reid. The role of differential geometry in statistical theory. *Intern. Stat. Rev.*, 54(1):83–96, 1986.

[BT73]   G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. J. Wiley & Sons, New York, 1973.

[CH74]   D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman and Hall, New York, 1974.

[Che72]   N. N. Chentsov. *Optimal Decision Rules and Optimal Inference*. Nauka, Moscow, 1972. In Russian. English translation AMS: Rhode Island, 1982.

[CMS93]   F. Critchley, P. Marriott, and M. Salmon. Preferred point geometry and statistical manifolds. *Ann. Statist.*, 21(3):1197–1224, 1993.

[Cox46]   R. T. Cox. Probability, frequency and reasonable expectations. *Amer. J. Phys.*, 14:1–26, 1946.

[Dar35]   G. Darmois. Sur les lois de probabilités à estimation exhaustive. *C. R. Acad. Sci. Paris*, 200:1265–1266, 1935.

[DeG70]   M. H. DeGroot. *Optimal Statistical Decisiosns*. McGraw-Hill, New York, 1970.

[Dem69]   A. P. Dempster. *Elements of Continuous Multivariate Analysis*. Addison-Wesley, Reading, MA, 1969.

[DSZ73]   A. P. Dawid, M. Stone, and J. V. Zidek. Marginalization paradoxes in Bayesian and structural inference (with discussion). *J. Roy. Stat. Soc., B*, 35:189–233, 1973.

[Efr75]   B. Efron. Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion). *Ann. Statist.*, 3:1189–1242, 1975.

[Egu83]   S. Eguchi. Second order efficiency of minimum contrast estimators in a curved exponential family. *Ann. Statist.*, 11:793–803, 1983.

[Fer67]   T. S. Fersuson. *Mathematical Statistics : A Decision Theoretic Approach*. Academic Press, New York, 1967.

[Fis22]   R. A. Fisher. On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc., A*, 222:309–368, 1922. Reprinted in [Fis50].

[Fis25]   R. A. Fisher. Theory of statistical estimation. *Proc. Camb. Phi. Soc.*, 122:700–725, 1925. Reprinted in [Fis50].

[Fis34]   R. A. Fisher. Two new properties of mathematical likelihood. *Proc. Roy. Soc., A*, 144:285–307, 1934. Reprinted in [Fis50].

[Fis50]   R. A. Fisher. *Contributions to Mathematical Statistics*. J. Wiley & Sons, New York, 1950.

[Fra56]   D. A. Fraser. Sufficient statistics with nuisance parameters. *Ann. Math. Statist.*, 27:838–842, 1956.

[Fra63]   D. A. Fraser. On sufficiency and the exponential family. *J. Roy. Stat. Soc., B*, 25:115–123, 1963.

[GBD92]   S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[Hou82]   P. Hougaard. Parameterization of non-linear models. *J. Roy. Stat. Soc., B*, 44:244–252, 1982.

[HS49]   P. R. Halmos and L. J. Savage. Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Ann. Math. Statist.*, pages 225–241, 1949.

[Jef61]   H. Jeffreys. *Theory of Probability*. Clarendon Press, Oxford, 1961. First edition in 1939.

[Kas84]   R. E. Kass. Canonical parameterization and zero parameter effects curvature. *J. Roy. Stat. Soc., B*, 46:86–92, 1984.

[Kas87]   R. E. Kass. Introduction. In Amari et al. [ABNK+87], chapter 1, pages 1–17.

[Kas89]   R. E. Kass. The geometry of asymptotic inference (with discussion). *Statistical Science*, 4(3):188–234, 1989.

[KL51]   S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22:79–86, 1951.

[Kol56]   A. N. Kolmogorov. *Foundations of the Theory of Probability*. Chelsea Publishing Co., New York, 1956. Translation of *Grundbegriffe der Wahrscheinlichkeitsrechnung, 1933*, with added bibliography, edited by N. Morrison.

[Koo36]   B. O. Koopman. On distributions admitting a sufficient statistic. *Trans. Amer. Math. Soc.*, 39:399–409, 1936.

[KS79]   M. Kendall and A. Stuart. *The Advanced Theory of Statistics: Inference and Relationship*, volume 2. Griffin, London, 4 edition, 1979.

[Lau87]   S. L. Lauritzen. Statistical manifolds. In Amari et al. [ABNK$^+$87], chapter 4, pages 163–216.

[Lor90]   T. J. Loredo. From Laplace to supernova SN 1987A: Bayesian inference in astrophysics. In P. F. Fougère, editor, *Maximum Entropy and Bayesian Methods*, pages 81–142. Kluwer Academic Publishers, 1990.

[LS72]   D. V. Lindley and A. F. M. Smith. Bayes estimation for the linear model. *J. Roy. Stat. Soc., B*, 34:1–41, 1972.

[Nea93]   R. M. Neal. Bayesian learning via stochastic dynamics. In S. J. Hanson, J. D. Cowan, and C. Lee Giles, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 475–482. Morgan Kaufmann, San Mateo, CA, 1993.

[Nea95]   R. M. Neal. *Bayesian Learning for Neural Networks*. PhD thesis, Dept. of Computer Science, University of Toronto, 1995.

[Pit36]   E. J. G. Pitman. Sufficient statistics and intrinsic accuracy. *Proc. Camb. Phi. Soc.*, 32:567–579, 1936.

[RS68]   H. Raiffa and R. Schlaifer. *Applied Statistical Decision Theory*. MIT Press, Cambridge, Mass., 1968.

[Whi89]   H. White. Learning in artificial neural networks: A statistical perspective. *Neural Computation*, 1(4):425–464, 1989.

[Zac71]   S. Zacks. *The Theory of Statistical Inference*. Wiley Series in Probability and Mathematical Statistics. J. Wiley & Sons, New York, 1971.

[Zhu95]   H. Zhu. On the relationship between minimising information divergence and least mean squares. In preparation, 1995.

[ZR95a]   H. Zhu and R. Rohwer. Bayesian invariant measurements of generalisation for discrete distributions. Technical Report NCRG/4351, Dept. Comp. Sci. & Appl. Math., Aston University, August 1995. `ftp://cs.aston.ac.uk/neural/zhuh/discrete.ps.Z`.

[ZR95b]   H. Zhu and R. Rohwer. Bayesian regression filters. Manuscript, 1995.

[ZR95c]   H. Zhu and R. Rohwer. Bayesian regression filters and the issue of priors. Submitted, 1995.

[ZR95d]   H. Zhu and R. Rohwer. Information geometric measurements of generalisation. Technical Report NCRG/4350, Dept. Comp. Sci. & Appl. Math., Aston University, August 1995. `ftp://cs.aston.ac.uk/neural/zhuh/generalisation.ps.Z`.