Functional Conjugacy in Parametric Bayesian Models

Peter Orbanz
University of Cambridge

Abstract

We address a basic question in Bayesian analysis: Can updates of the posterior under observations be represented as a closed-form mapping from the data to the posterior parameters? The question is closely related to the concept of a conjugate prior, but we do not assume that prior and posterior belong to the same model class. We refer to models for which a closed-form mapping exists as *functionally conjugate*, and ask which observation models admit such functionally conjugate priors. For finite-dimensional, dominated models, the answer is almost disappointingly restrictive: Under mild regularity assumptions, such a mapping can only exist if the likelihood is an exponential family model. This is a consequence of a more general result: In dominated models with strictly positive prior densities, existence of a mapping to the posterior parameters of the Bayesian model implies the existence of a sufficient statistic for the sampling model.

1 Introduction

One of the most widely used tools in Bayesian analysis are conjugate priors for exponential family models. The definition of a conjugate prior (Raiffa and Schlaifer, 1961) in general demands that, for a given observation model (likelihood), and a given class of priors, all corresponding posteriors are again in the prior class. This property is known as *closure under sampling*. The particular utility of conjugate priors in the exponential family case, however, is mostly due to the fact that the parameters of the posterior given observations are a simple, known function of the prior parameters and the data, making the posterior easily computable from a set of observations. Due to their coincidence in exponential families, the two concepts are often used interchangeably in the literature, but they are not equivalent: Closure under sampling does not guarantee the existence of a mapping to the posterior parameters. In particular, the set of all probability distributions on parameter space is, when used as the family of priors, conjugate to any sampling model. On the other hand, a closed-form mapping to the posterior parameters may presumably exist even if the models are not closed under sampling, i.e. if the posterior belongs to a different parametric model class than the prior.

In the following, we study the implications of the existence of a mapping to the posterior parameters – that is, of the assumption that there exists a parametric model containing all possible posterior measures, and that the posterior parameters can be expressed as a measurable function of the prior parameters and the data. We do not assume that the posterior and the prior are elements of one and the same model class. A Bayesian model admitting such a mapping will be referred to as *functionally conjugate*. Our main result shows that, in the dominated case and under a mild regularity assumption on the prior family, the existence of such a mapping implies the existence of a sufficient statistic for the observation model. The study of functionally

conjugate models thereby essentially reduces to the well-studied subject of sufficient statistics. In combination with Pitman-Koopman theory, which shows that dominated models admitting a sufficient statistic of dimension bounded w.r.t. sample size are of exponential family type, a particular implication is that if the parameter space of the posterior class has finite dimension, then the likelihood term of a functionally conjugate Bayesian model is an exponential family model.

For a rough overview of the problem, consider a Bayesian estimation problem involving data values $x_1, x_2, \ldots$ generated from a parametric model $P_{\mathrm{X}}(X|\Theta)$. The data is assumed to be conditionally i.i.d. given the true value of Θ. Given a prior distribution P_Θ on the space of parameter values, the objective of Bayesian inference is to compute the corresponding posterior distribution, i.e. the conditional probability $P_\Theta(\Theta|X_1 = x_1, \ldots, X_n = x_n)$. Under suitable conditions on the spaces and models involved, Bayes' theorem guarantees that the density of the posterior with respect to the prior can be expressed in terms of the density of $P_{\mathrm{X}}(X|\Theta)$: If the density of the observation model is denoted $f(x|\theta)$, then

$$\frac{dP_\Theta(\theta|X = x)}{dP_\Theta} = \frac{f(x|\theta)}{\int f(x|\theta) dP_\Theta(\theta)} \,. \tag{1}$$

This density provides the formal means to actually compute the abstract object $P_\Theta(\Theta|X_1 = x_1, \ldots, X_n = x_n)$ from the prior and the data. In order to do so, the integral in the denominator has to be evaluated for the given sample. This innocuous-looking integration problem is not solvable in general – for most choices of observation model and prior, no analytic solution exists, and numerical solutions are feasible in low dimensions at best. For a small (but fortunately important) subset of possible models, the integral admits a closed-form solution. In particular, for the so-called natural conjugate priors in exponential family models, the application of the density (1) to computation of the posterior leads to generic update equations for the posterior parameters: The posterior is a parametric model of known form, and its computation requires only the substitution of the data into a generic and easily evaluated formula. In this case, the prior is also part of a parametric model, denoted $P_\Theta(\Theta|Y)$ in the following. The parameter Y of the prior is generally referred to as a *hyperparameter* in the literature. Denote the density of the prior by $g(\theta|y)$. Assume that $f(x|\theta)$ is an exponential family density, and $g(\theta|y)$ the density of its natural conjugate prior (definitions will be given below). Then for the prior specified by some value y_0 of Y, and n observations, the posterior has density $g(\theta|y = T_n(x_1, \ldots, x_n, y_0))$, where the function T_n is of known, generic form and can be computed from the model's sufficient statistics. The problem studied in the present paper is how the existence of such a mapping to the posterior parameter can be formalized in a suitable manner, and for which model distributions it can be expected to exist.

2 Background

Conjugacy and the existence of a closed-form mapping to the posterior parameters are two different properties, and will have to be carefully distinguished for the purposes of this article. The general definition of conjugate priors defines a class of priors as conjugate to an observation model $P_{\mathrm{X}}(X|\Theta)$ if the resulting Bayesian model is "closed under sampling" (e.g. Lindley, 1972; Robert, 1994): A prior and a sampling model are called conjugate if the posterior is an element of the same model class as the prior, as exemplified by the exponential family case described

above, where the posterior under prior $g(\theta|y_0)$ is again of the form $g(\theta|y)$. Closure under sampling is only of limited practical importance: In particular, it includes the trivial case in which the prior class is the set of all probability distributions over the given parameter space, which is apparently conjugate to any possible sampling model. Of actual practical importance is the question whether the posterior parameters are a function of the data and the prior, and whether this function is feasible to evaluate, resulting in a closed-form representation of the posterior. Closure under sampling does not imply tractability of the posterior. In order to clearly distinguish the two concepts, we will introduce the following terminology: Conjugacy in the sense of closure under sampling will be referred to as *algebraic* conjugacy, as its defining characteristic is membership of a distribution in a set. A model admitting a closed-form mapping to the posterior parameters will be called *functionally conjugate*.

Definition 1. Let $P_{\mathrm{X}}(X|\Theta)$ a sampling distribution with parameter space Ω_θ, and let $\mathcal{M}$ be a set of prior distributions on Ω_θ. Then the model $P_{\mathrm{X}}(\,.\,|\Theta)$ and the set $\mathcal{M}$ will be called *conjugate* or *algebraically conjugate* if, for every prior measure $P_\Theta \in \mathcal{M}$ and set of observation $X = x$, the corresponding posterior is an element of $\mathcal{M}$.

A formal definition of functional conjugacy will be given in Sec. 3. We will later address the question for which parametric models a functionally conjugate prior can be expected to exist. The remainder of the current section formalizes the mathematical setting and summarizes two concepts relevant for our study of functionally conjugate models: Sufficient statistic for parametric models, and the particular form of conjugate priors in exponential family models. The latter are both algebraically and functionally conjugate.

2.1 Formal Setting

We will generally assume all involved random variables – the observation variable X, the parameter Θ and the hyperparameter Y – to take values in Polish sample spaces, i.e. complete, separable and metrizable topological spaces. These will be equipped with their Borel algebras, and the resulting measurable spaces are denoted $(\Omega_x, \mathcal{B}_x)$, $(\Omega_\theta, \mathcal{B}_\theta)$ and $(\Omega_y, \mathcal{B}_y)$ respectively. All random variables are assumed to be defined on a common abstract probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The models $P_{\mathrm{X}}(X|\Theta)$ and $P_\Theta(\Theta|Y)$ are regular conditional probabilities of the corresponding image measures $P_{\mathrm{X}} = X(\mathbb{P})$ and $P_\Theta = \Theta(\mathbb{P})$. We do not require the spaces involved to be finite-dimensional, but we will assume that the models $P_{\mathrm{X}}(X|\Theta)$ and $P_\Theta(\Theta|Y)$ are dominated[1]. This excludes some infinite-dimensional models which are not dominated, such as the Dirichlet process. As usual in Bayesian analysis, observations will be assumed exchangeable, and in particular conditionally i.i.d. given the value of Θ. As a convenient side-effect, this will allow us to treat only the case of a single observation without any loss of generality.

2.2 Sufficient Statistics

Conjugate models are inextricably linked with sufficient statistics, and we will have to briefly review the concept of sufficiency in both classical and Bayesian models before discussing conjugate models from a technical point of view.

[1] We refer to a conditional probability, or any set of probability measures, as *dominated* if all measures in the set are absolutely continuous w.r.t. a single, σ-finite measure.

A sufficient statistic for a parametric model $P_{\mathrm{X}}(X|\Theta)$ is, intuitively speaking, a function of the data summarizing all relevant information the data provides for estimation of the parameter. Since the notion of providing information about the data has different meanings in the classical and Bayesian approach, there are two different definitions of sufficiency.

Definition 2 (Classical Sufficiency)**.** Let $P_{\mathrm{X}}(X|\Theta)$ be a parametric model on $(\Omega_x, \mathcal{B}_x)$ with parameter space $(\Omega_\theta, \mathcal{B}_\theta)$. Let $(\Omega_s, \mathcal{B}_s)$ be a measurable Polish space, and $S : \Omega_x \to \Omega_s$ a measurable map. Then S is called a *classically sufficient statistic* for $P_{\mathrm{X}}(X|\Theta)$ if there exists a Markov kernel $k : \mathcal{B}_x \times \Omega_s \to [0, 1]$ such that

$$P_{\mathrm{X}}(A|\Theta = \theta, S = s) =_{\text{a.e.}} k(A, s) \tag{2}$$

for all $A \in \mathcal{B}_x$, $\theta \in \Omega_\theta$ and $s \in S(\Omega_x)$. The a.e. equality is with respect to the image measure $S[P_{\mathrm{X}}(\,.\,|\Theta = \theta)] = P_{\mathrm{X}}(S^{-1}\,.\,|\Theta = \theta)$ of the sampling distribution under the map S.

In the Bayesian case, the notion of S providing all information contained in the data about the parameter is formalized as complete determination of the posterior by the value of S.

Definition 3 (Bayesian Sufficiency)**.** Let $P_{\mathrm{X}}(X|\Theta)$ be a parametric model and $S : \Omega_x \to \Omega_s$ a measurable mapping as above. Then S is called a *Bayesian sufficient statistic* for the model if the posterior $P_\Theta(\Theta|X)$ of the model under any prior P_Θ on $(\Omega_\theta, \mathcal{B}_\theta)$ satisfies

$$P_\Theta(\Theta|X = x) = P_\Theta(\Theta|S = S(x)) \qquad P_{\mathrm{X}}\text{-a.e.} \tag{3}$$

On Polish or Borel spaces, the two notions of sufficiency are equivalent if the model $P_{\mathrm{X}}(X|\Theta)$ is dominated. In the undominated case, classical sufficiency always implies Bayesian sufficiency, but the converse need not be true. The implication of a statistic being sufficient in the Bayesian, but not classical sense is that the Bayesian model is unable to resolve at least some cases which are distinguished as different by the classical model. Blackwell and Ramamoorthi (1982) give a hypothesis testing example for which the two notions of sufficiency differ – with the consequence that there is no classical test achieving zero error probability, whereas the Bayesian version of the test does so with probability 1. The Bayesian is always certain to be right, the classical test is always uncertain. An excellent, in-depth exposition of classical and Bayesian sufficiency is given by Schervish (1995). Historically, the classical notion of sufficiency is due to Fisher (1922). The Bayesian notion is attributed to Kolmogorov (1942).

The following well-known theorem characterizes dominated models which admit a sufficient statistic and will be used below to construct the generic conjugate prior of an exponential family model. In the form generally reproduced in the literature, it is due to Halmos and Savage (1949), who attribute the result to Neyman (1935).

Theorem 1 (Neyman Factorization Theorem)**.** *Let $P_X(X|\Theta)$ be a dominated parametric family with conditional density $p(x|\theta)$. Then the model admits a classically sufficient statistic S_n (for each sample size n) if and only if the density factorizes in the form*

$$p(x|\theta) = f_n(x_1, \ldots, x_n) g_n(S_n(x_1, \ldots, x_n), \theta) \tag{4}$$

for two suitable functions f_n and g_n.

2.3 Canonical Conjugate Priors in Exponential Families

We will now consider the specific case in which $P_{\mathrm{X}}(\,.\,|\Theta)$ is an exponential family model. Let $P_{\mathrm{X}}(X|\Theta)$ be a parametric model on a Polish measurable space $(\Omega_x, \mathcal{B}_x)$, for which the parameter variable Θ takes values in $\mathbb{R}^d$. Assume further that the model is dominated. The model is called an *exponential family model* if its density w.r.t. to some dominating measure ν is of the form

$$p(x|\theta) = \exp\Big(\langle S(x)|\theta\rangle - \phi(\theta) - \psi(x)\Big)\ , \tag{5}$$

where $S : \Omega_x \to \mathbb{R}^d$ is a measurable map, and ϕ and ψ are measurable real-valued functions. For S and ψ given, $\phi(\theta) = \log \int e^{\langle S(x)|\theta\rangle - \psi(x)} d\nu(x)$, and the parameter space of the model is defined as the subset $\Omega_\theta \subset \mathbb{R}^d$ for which $\phi(\theta) < \infty$. The parameter space is always convex (application of Hölder's inequality to the integral ϕ gives $\phi(a\theta_1 + (1-a)\theta_2) \leq a\phi(\theta_1) + (1-a)\phi(\theta_2)$). By Th. 1 above, the statistic S is classically sufficient. In other words, an exponential family representation always implies the existence of a sufficient statistic. Under mild regularity conditions, the converse is also true. This is a result known (amongst other names) as the *Pitman-Koopman lemma*, and it exists in a variety of flavors which differ in the choice of smoothness assumptions. References include Darmois (1935); Pitman (1936); Koopman (1936); Barankin and Maitra (1963); Hipp (1974); and many others. The following version is due to Harold Jeffreys.

Theorem 2 (Pitman-Koopman Lemma; Jeffreys (1961))**.** *Let the random quantities X_1, $X_2, \ldots$ be conditionally i.i.d. given the value of some random quantity Θ, and assume that the conditional distribution $P_X(X_i|\Theta)$ is dominated by a measure ν. Let $p(x|\theta)$ be the corresponding conditional density. Assume further that the support of $f_{X|\theta}$ is independent of the value of θ:*

$$\forall \theta_1, \theta_2 \in \Omega_\theta : \qquad \mathit{supp}\ p(\,.\,|\theta_1) = \mathit{supp}\ p(\,.\,|\theta_2) \qquad \nu\text{-a.e.} \tag{6}$$

Then if there is a sufficient statistic $S_n : \Omega_x^n \to \Omega_s$ for each sample size $n \in \mathbb{N}$, and if Ω_s has finite dimension, $P_X(X|\Theta)$ is an exponential family model.

For data drawn conditionally i.i.d. (given θ) from an exponential family model, the data-dependent terms in the exponent decompose additively over samples. This in turn implies that, for any sample size n, the sufficient statistic S_n of the joint data distribution can be represented as a sum involving only observation-wise application of the statistic $S := S_1$:

$$S_n(x_1, \ldots, x_n) = \sum_{i=1}^{n} S(x_i)\ . \tag{7}$$

It is well-known that exponential family models have conjugate priors of a generic form: Let $p(x|\theta)$ be an exponential family density as above. Then a generic conjugate prior is given by the density (w.r.t. Lebesgue measure on $\mathbb{R}^d$)

$$p(\theta|\alpha, y) := \frac{1}{K(\alpha, y)} \exp\Big(\langle\theta|y\rangle - \alpha\phi(\theta)\Big)\ , \tag{8}$$

with hyperparameters $\alpha \in \mathbb{R}_+$ and $y \in \Omega_\theta$, and partition function $K(\alpha, y) = \int e^{\langle\theta|y\rangle - \alpha\phi(\theta)} d\lambda^d(\theta)$. The prior in (8) is often called the *natural* or *canonical* conjugate prior of $p(x|\theta)$.

The canonical conjugate prior can be derived by means of the Neyman factorization theorem, as shown by DeGroot (1970). Assume that the model admits a sufficient statistic of fixed

dimension w.r.t. sample size (which by the Pitman-Koopman lemma implies, up to regularity conditions, that the model is of exponential family form). Now regard the functions $g_n(\,.\,,\theta)$ as a single function $g(n,\,.\,,\theta)$ and use it as the shape function of a probability model on Θ: Let $K(n,y) := \int g(n,y,\theta)d\lambda^d(\theta)$, and define the prior density as

$$p(\theta|n,y) := \frac{1}{K(n,y)} g(n,y,\theta)\ . \tag{9}$$

Since the model is of exponential family form (5), the joint density of n observations $x_1,\ldots,x_n$ is $\exp(\langle\sum_{i=1}^n S(x_i)|\theta\rangle - n\phi(\theta) - \sum_{i=1}^n \psi(x_i))$, hence $g(n,y,\theta) = \exp(\langle y|\theta\rangle - n\phi(\theta))$, and so $\frac{1}{K(n,y)}g(n,y,\theta)$ takes the form given in (8) with $\alpha = n$. It is straightforward to verify that the posterior corresponding to $p(\theta|\alpha,y)$ under observations $x_1,\ldots,x_n$ has density

$$p(\theta|x_1,\ldots,x_n) \propto \exp\Big(\Big\langle\theta\Big|y + \sum_{i=1}^n S(x_i)\Big\rangle - (\alpha+n)\phi(\theta)\Big)\ , \tag{10}$$

and so $p(\theta|x_1,\ldots,x_n) = p(\theta|\alpha+n, y+\sum S(x_i))$. The generic form

$$\lambda \mapsto \lambda + n \qquad \text{and} \qquad y \mapsto \sum_i S(x_i) \tag{11}$$

of the parameter updates is, in terms of calculus, of course due to the sample-wise application (7) of the sufficient statistic, and the log-linearity of the model in the value of S. The consequence is a linear geometry in parameter space: The images of observations under the sufficient statistic, their averages, the parameters, and the hyperparameters all constitute points in the space Ω_θ. Posteriors are obtained by linear interpolation. The convexity of Ω_θ guarantees its closure under such interpolations. A result of Diaconis and Ylvisaker (1979) uses this linear arithmetic to characterize the set of all conjugate priors in exponential families. They show, for the continuous case, that conjugate priors are those for which the expectation of the sample mean with respect to the posterior is linear.

Theorem 3 (Diaconis-Ylvisaker characterization of conjugate priors)**.** *Let $P_X(\,.\,|\Theta)$ be a natural exponential family model dominated by Lebesgue measure, with open parameter space $\Omega_\theta \subset \mathbb{R}^d$. Let P_Θ be a prior on Θ which does not concentrate on a singleton. Then P_Θ has a density of the form (8) w.r.t. Lebesgue measure on $\mathbb{R}^d$ if and only if*

$$\mathbb{E}_{P_\Theta(\Theta|X_1=x_1,\ldots,X_n=x_n)}\left[\mathbb{E}_{P_X(X|\Theta=\theta)}\left[X\right]\right] = \frac{y+n\hat{x}}{a+n}\ . \tag{12}$$

That is, given observations $x_1,\ldots,x_n$, the expected value of a new draw x under unknown value of the parameter is linear in the sample average $\hat{x} = \frac{1}{n}\sum x_i$.

3 Functional Conjugacy

By the concept of functional conjugacy, we formalize the idea that posterior updates in Bayesian inference can be computed can be represented as updates of the posterior parameters. Therefore, (i) the set of posteriors must be representable as a parameterized family of models and (ii) there must be a mapping which, given the prior, takes the data to the corresponding value of the posterior parameter.

Formal definition. Assume that the sample space Ω_x, the parameter space Ω_θ and the space Ω_y of prior hyperparameters are Polish and equipped with their respective Borel algebras. Let the sampling model $P_{\mathrm{X}}(X|\Theta)$ and the prior family $P_\Theta(\Theta|Y)$ be regular conditional probabilities on Ω_x and Ω_θ, respectively. Write $P_\Theta(\Theta|X,Y)$ for the posterior of $P_{\mathrm{X}}(X|\Theta)$ under prior $P_\Theta(\Theta|Y)$. If we assume all three conditional models to be dominated, with densities $f(x|\theta)$ (likelihood), $g(\theta|y)$ (prior) and $p(\theta|x,y)$ (posterior) respectively, functional conjugacy can be formalized as follows: There is function $q(\theta|t)$, parameterized by t in some set Ω_t, and a mapping $T(x,y)$ with values in Ω_t, such that

$$p(\theta|x,y) = q(\theta|T(x,y)) \tag{13}$$

for all possible values of x, y and θ. For a definition that does not rely on densities, we have to substitute q by a parameterized probability measure π_t. The Bayesian assumption that parameters are random variables has direct consequences for the formalization of both requirements (i) and (ii): The parameterized measure π_t representing the posterior must be a conditional probability. Since Ω_θ is Polish, this conditional probability has a regular version. Since the value of T substitutes for a random variable, and is conditioned upon, it must itself be interpretable as a random variable, and so the mapping T must be measurable. In summary, we obtain the following general definition.

Definition 4. Let $P_{\mathrm{X}}(X|\Theta)$ and $P_\Theta(\Theta|Y)$ be regular conditional probabilities. Let $P_\Theta(\Theta|X,Y)$ be the posterior of $P_{\mathrm{X}}(X|\Theta)$ under prior $P_\Theta(\Theta|Y)$. The two models will be called *functionally conjugate* if there exists a regular conditional probability $\pi : \mathcal{B}_\theta \times \Omega_t \to [0,1]$, parameterized on a measurable Polish space $(\Omega_t, \mathcal{B}_t)$, and a measurable map $T : \Omega_x \times \Omega_y \to \Omega_t$, such that

$$P_\Theta(A|X=x, Y=y) = \pi(A, T(x,y)) \qquad \text{for all } A \in \mathcal{B}_\theta \ . \tag{14}$$

We have already discussed the natural conjugate priors of exponential family models in the previous section. Such models with their natural conjugate priors are examples of both types of conjugacy: Algebraic, because the posterior is in the same model class as the prior, and functional, because they admit a closed-form mapping to the posterior parameters. In this particular case, the posterior index T is given by the mapping

$$T(x,(\lambda,y)) = (\lambda+1, y+S(x)) \ . \tag{15}$$

Relation to sufficiency. Obviously, the definition is closely related to that of classical sufficiency (Def. 2), though the conditional on the left of Eq. (14) is the posterior rather than the sampling model. The definition admits a sampling interpretation similar to that of classical sufficiency: Given the value of T (but not the data), we can sample from the posterior as if we knew the data. From a technical point of view, the definition expresses this requirement – that the posterior only resolves information in the data that is resolved by T – by demanding the posterior to be measurable w.r.t. the σ-algebra generated by the mapping T: Fix one particular prior, i.e. one value $y_0 \in \Omega_y$ of the hyperparameter, and write $T_{y_0}(x) := T(x,y_0)$ for the resulting mapping from data to posterior parameters. The posterior of a given event $A \in \mathcal{B}_\theta$ under observation $X = x$ is the conditional expectation $\mathbb{E}[\mathbb{I}_A|X=x, Y=y_0]$. The definition requires this to be equivalently expressible as

$$\mathbb{E}[\mathbb{I}_A|X=x, Y=y_0] = \mathbb{E}[\mathbb{I}_A|T=T_{y_0}(x)] = \mathbb{E}[\mathbb{I}_A|T=\ .\,] \circ T_{y_0} \circ X(\omega) \ , \tag{16}$$

where ω is an element of the abstract probability space Ω. Since $\mathbb{E}[\mathbb{I}_A|T = \,.\,]$ is a $\mathcal{B}_t$-measurable function, this makes the overall posterior $\sigma(X \circ T_{y_0})$-measurable. An arbitrary (non-conjugate) posterior is measurable w.r.t. $\sigma(X)$. Since in $\sigma(X \circ T_{y_0}) \subset \sigma(X)$, this can be regarded as a coarsening of the conditional information available to determine the posterior. In the case of algebraic conjugacy, the posterior is an element of the class $P_\Theta(\Theta|Y)$ of priors, and can be completely determined by specifying an appropriate value y for the hyperparameter Y. Algebraic conjugacy therefore makes the posterior $\sigma(Y)$-measurable.

Whereas functional conjugacy somewhat resembles classical sufficiency, there is a more direct relationship between functional conjugacy and the Bayesian notion of sufficiency (cf Def. 3). The key difference between the two concepts is that Bayesian sufficiency requires the posterior to be determined by the statistic under *any* prior. In contrast, the definition above only guarantees a complete determination of those posteriors which are derived from the prior family $P_\Theta(\Theta|Y)$. Interestingly, under mild regularity conditions, this already implies the complete determination of the posterior under any prior. The following theorem states that, for dominated families, conjugacy implies sufficiency if the prior densities do not vanish anywhere on parameter space.

Theorem 4 (Conjugate models admit sufficient statistics)**.** *Let $P_X(X|\Theta)$ and $P_\Theta(\Theta|Y)$ be two dominated parametric families on Borel spaces $(\Omega_x, \mathcal{B}_x)$ and $(\Omega_\theta, \mathcal{B}_\theta)$, and functionally conjugate with posterior index T. If $P_\Theta(\Theta|Y = y)$ has a strictly positive density for all $y \in \Omega_y$ with respect to some dominating measure on Ω_θ, then the function $T_{y_0}(\,.\,) := T(\,.\,, y_0)$ is a sufficient statistic for $P_X(X|\Theta)$ for all $y_0 \in \Omega_y$.*

Discussion of assumptions. The regularity condition (strict positivity of the prior densities) in particular requires that the prior densities have identical support. This is a similar requirement to that of the Pitman-Koopman lemma, though the former is imposed on the sample space Ω_x and the latter on parameter space Ω_θ. Without a suitable regularity condition, it is possible to construct pathological examples of conjugate models that do not yield a sufficient statistic: Consider for instance the a set of priors consisting of all Dirac measures on the parameter space Ω_θ, for some smooth sampling distribution. Then for every Dirac concentrated at some $\theta \in \Omega_\theta$, the posterior is again the same Dirac measure, regardless of the observations. The class is therefore conjugate. Since the Dirac measures are parameterized by their position, the identity mapping $\mathrm{Id}_{\Omega_\theta}$ completely determines the posterior parameter. Apparently, this does not imply that $\mathrm{Id}_{\Omega_\theta}$ (which does not even depend on the data) is a sufficient statistic for the sampling distribution. However, such examples only exist in cases where the prior is in some way degenerate. To illustrate the condition of strict positivity of the prior density, note that if, on the other hand, the parameter space Ω_θ were to consist of only a single value, the mapping T would still be trivial – but so would be the sampling model, which consists only of a single measure. Roughly speaking, the size of the parameter space bounds both the possible complexity of the sampling model and of the mapping T. In order to ensure that T carries sufficient information about the sampling distribution, we have to ensure that no part of the parameter space is "blotted out" by prior assumption. The theorem could, in this regard, be formulated the other way around: For a parametric family of priors with arbitrary (possibly vanishing) densities, T_{y_0} is sufficient for the model obtained from $P_{\mathrm{X}}(X|\Theta)$ by restricting the set of parameters to the common support $\cap_{y\in\Omega_y}\mathrm{supp}\left(g(\,.\,|y)\right)$ of all prior densities, and restricting the priors accordingly.

Consequences for finite-dimensional models. Combination of Th. 4 with the Pitman-Koopman lemma (Th. 2) implies that – if the regularity conditions of both theorems are satisfied – functional conjugacy occurs only in exponential families.

Corollary 1. *Let $P_X(X|\Theta)$ and $P_\Theta(\Theta|Y)$ be two dominated parametric models with strictly positive conditional densities. Assume that the space Ω_t of posterior parameters has finite dimension. Then if $P_X(X|\Theta)$ and $P_\Theta(\Theta|Y)$ are functionally conjugate, $P_X(X|\Theta)$ is an exponential family model.*

4 Proof of Theorem 4

Considering the close connection between functional conjugacy and Bayesian sufficiency, it will come as no surprise that the same technique can be used in the proof, though the technical details differ. The idea is to establish classical sufficiency for the observation model by constructing the parameter-independent regular conditional probability required by the definition, and then express it in terms of the functionally conjugate posterior to show that it depends on the data only through the posterior index T. A similar construction can be used to proof that Bayesian sufficiency implies classical sufficiency (Blackwell and Ramamoorthi, 1982; Schervish, 1995). The proof idea is originally derived from Halmos and Savage (1949, proof of Theorem 1). The same is true for the representation (18) below, which is due to the proof establishing that a dominated set of measures has an equivalent countable subset (Halmos and Savage, 1949, Lemma 7).

Proof (Theorem 4). By definition, sufficiency of T_{y_0} for $P_{\mathrm{X}}(X|\Theta)$ requires the existence of a Markov kernel $k : \mathcal{B}_x \times \Omega_y \to [0,1]$ such that

$$P_{\mathrm{X}}(A|\Theta = \theta, T_{y_0} = t) =_{\text{a.e.}} k(A,t) \; . \tag{17}$$

We will construct a candidate kernel k and then show that it satisfies Eq. (17). To define k, we note that the family $P_{\mathrm{X}}(X|\Theta)$ is dominated by assumption. According to Halmos and Savage (1949), this implies existence of a measure ρ on $(\Omega_x, \mathcal{B}_x)$ such that (i) $\mu_{X|\theta} \ll \rho \ll \nu_X$ and (ii) ρ has a representation as a countable convex combination of measures in the family. That is,

$$\rho = \sum_{i\in\mathbb{N}} c_i P_{\mathrm{X}}(\,.\,|\Theta = \theta_i) \tag{18}$$

for some countable sequences $(\theta_i)_{i\in\mathbb{N}}$ of parameters in Ω_θ and $(c_i)_{i\in\mathbb{N}}$ of mixture weights, where $\sum_{i\in\mathbb{N}} c_i = 1$. We define k as the conditional probability given T w.r.t. the measure ρ:

$$k(A,t) := \mathbb{E}_\rho\left[\mathbb{I}_A | T = t\right] \; . \tag{19}$$

Formally, this is the conditional expectation $\mathbb{E}[X_\rho|T=t]$ of a random variable X_ρ which generates the measure ρ, that is $\rho = X_\rho(\mathbb{P})$. To show that k satisfies (17), we have to show that k behaves like the conditional probability $P_{\mathrm{X}}(A|\Theta=\theta, T_{y_0}=t)$, i.e. that it integrates as $P_{\mathrm{X}}(A|\Theta=\theta, T_{y_0}=t)$ does w.r.t. the measures $P_{\mathrm{X}}(\,.\,|\Theta=\theta)$.

To show this, we will have to manipulate an integral over k, using the properties of conditional expectations. Since k is defined w.r.t. ρ, the integral measure $P_{\mathrm{X}}(\,.\,|\Theta=\theta)$ has to be expressed in terms of ρ as $dP_{\mathrm{X}}(x|\Theta=\theta) = h(x,\theta)d\rho(x)$. For the density h to be compatible with $\mathbb{E}_\rho\left[\mathbb{I}_A|T_{y_0}\right]$ under the integral, the function $h(\,.\,|\theta)$ must be measurable w.r.t. the σ-algebra generated by $T(\,.\,,y)$, and so we will have to show that h can be represented as a function which depends on x only through T.

Step 1: Expressing h as a function of T. By assumption, both families $P_{\mathrm{X}}(\,.\,|\Theta = \theta)$ and $P_\Theta(\,.\,|Y = y)$ are dominated. Let ν_X and ν_Θ be dominating measures, and denote the respective conditional densities as

$$f(\,.\,|\theta) := \frac{dP_{\mathrm{X}}(\,.\,|\Theta = \theta)}{d\nu_X} \qquad \text{and} \qquad g(\,.\,|y) := \frac{dP_\Theta(\,.\,|Y = y)}{d\nu_\Theta} \,. \tag{20}$$

By functional conjugacy,

$$\pi(A, T(x, y)) =_{\text{a.e.}} P_\Theta(A|X = x, Y = y) \,. \tag{21}$$

We first have to argue that the conditional probability $\pi(A, T(x, y))$ has a density w.r.t. to ν_Θ. Without loss of generality, we can assume that $\text{range}(T) = \Omega_t$, i.e. the function value $\pi(A, t)$ for any t can be expressed as the posterior for some values x and y. (Otherwise, restrict the function $\pi(A, \,.\,)$ to $T(\Omega_x \times \Omega_y)$, equipped with the trace σ-algebra $\mathcal{B}_t \cap T(\Omega_x \times \Omega_y)$, which makes the restriction measurable.) The posterior is dominated by the prior and the prior by ν_Θ, such that π is dominated by ν_Θ according to (21). The density of the measure $\pi(\,.\,, t)$ w.r.t. to ν_Θ will be denoted as $q(x|t)$ in the following.

By the chain rule, the density of $P_{\mathrm{X}}(\,.\,|\Theta = \theta)$ (for any θ) with respect to ρ is

$$\frac{dP_{\mathrm{X}}(\,.\,|\Theta = \theta)}{d\rho} = \frac{f(\,.\,|\theta)}{\sum_{i\in\mathbb{N}} c_i f(\,.\,|\theta_i)} \,. \tag{22}$$

Since both families are dominated, the Bayes' theorem is applicable, and the density of the posterior is

$$\frac{dP_\Theta}{d\nu_\Theta}(\theta|X = x, Y = y) =_{\text{a.e.}} \frac{f(x|\theta)g(\theta|y)}{F(x, y)} \,, \tag{23}$$

where $F(x, y) = \int_{\Omega_\theta} f(x|\theta)g(\theta|y)d\nu_\Theta(\theta)$. The "almost everywhere" in (23) is due to the fact that the integral $F(\,.\,, y)$ in the Bayes theorem may take values $\{0, \infty\}$ on a null set. Since the model is conjugate, the posterior density can be expressed as

$$\frac{dP_\Theta(\theta|X = x, Y = y)}{d\nu_\Theta} = q(\theta|T(x, y)) \,. \tag{24}$$

The regularity assumption on g (that g does not vanish anywhere on Ω_θ) guarantees that the quotient (22) can be expressed in terms of the posterior and the prior: Since g is non-zero everywhere, the Bayes equation (23) can be solved for f, and substitution into (22) gives

$$\frac{dP_{\mathrm{X}}(x|\Theta = \theta)}{d\rho} = \frac{q(\theta|T(x, y))\frac{F(x,y)}{g(\theta|y)}}{\sum_i c_i q(\theta_i|T(x, y))\frac{F(x,y)}{g(\theta_i|y)}} = \frac{q(\theta|T(x, y))}{g(\theta|y)\sum_i c_i \frac{q(\theta_i|T(x,y))}{g(\theta_i|y)}} =: h(y, \theta, T(x, y)) \,. \tag{25}$$

Zero values of the denominator can occur, but only on a ρ-null set: Let M_0 be the set of all x for which the denominator vanishes. Since g is strictly positive, $x \in M_0$ implies $\sum_i c_i q(\theta_i|T(x, y)) = 0$. Hence, as $q(\theta|T(x, y))$ only vanishes if $f(x|\theta) = 0$,

$$M_0 = \{x|f(x|\theta_i) = 0 \text{ for all } \theta_i\} \,. \tag{26}$$

Since f is measurable and $(c_i)_{i\in\mathbb{N}}$ countable, M_0 is measurable, and with (18),

$$\rho(M_0) = \int_{M_0} \frac{d\rho}{d\nu_X}(x)d\nu_X(x) = \int_{M_0} \sum_i c_i f(x|\theta_i)d\nu_X(x) = 0 \,. \tag{27}$$

As a density, h can be modified on a ρ-null set, and we set $h(y, \theta, T(x, y)) = 0$ whenever $x \in M_0$. *Step 2: The kernel k satisfies Eq. (17).* What remains to be shown is that $k(A, t)$ is a version of the conditional probability $P_{\mathrm{X}}(A|\Theta = \theta, T = t)$ for all $\theta \in \Omega_\theta$. Formulated on the abstract probability space $(\Omega, \mathcal{A}, \mathbb{P})$, that means proving that for any $A \in \mathcal{A}$,

$$\int_B P_{\mathrm{X}}(A|\Theta, T_{y_0})(\omega) d\mathbb{P}(\omega) = \int_B k(A, \,.\,) \circ T_{y_0} \circ X d\mathbb{P}(\omega) \qquad \text{for all } B \in \sigma(T_{y_0} \circ X) \,. \tag{28}$$

This is equivalent to showing, on Ω_x rather than Ω, that

$$\int_{T_{y_0}^{-1}(C)} P_{\mathrm{X}}(A|\Theta = \theta, T_{y_0} = T_{y_0}(x)) dP_{\mathrm{X}}(x|\Theta = \theta) = \int_{T_{y_0}^{-1}(C)} k(A, \,.\,) \circ T_{y_0}(x) dP_{\mathrm{X}}(x|\Theta = \theta) \tag{29}$$

for all $A \in \mathcal{B}_x$, $\theta \in \Omega_\theta$ and $C \in \mathcal{B}_y$. The integral on the left-hand side is $P_{\mathrm{X}}(A \cap T_{y_0}^{-1}(C)|\Theta = \theta)$, for which in turn

$$P_{\mathrm{X}}(A \cap T_{y_0}^{-1}(C)|\Theta = \theta) = \int_{A \cap T_{y_0}^{-1}(C)} dP_{\mathrm{X}}(x|\Theta = \theta) \,. \tag{30}$$

According to (25)

$$dP_{\mathrm{X}}(x|\Theta = \theta) = h(y_0, \theta, \,.\,) \circ T_{y_0} d\rho(x) \,, \tag{31}$$

so the integral can be rewritten as

$$\int_{A \cap T_{y_0}^{-1}(C)} dP_{\mathrm{X}}(x|\Theta = \theta) = \int_{T_{y_0}^{-1}(C)} \mathbb{I}_A(x) h(y_0, \theta, \,.\,) \circ T_{y_0} d\rho(x) \,. \tag{32}$$

Let $\sigma(T_{y_0}) = T_{y_0}^{-1}(\mathcal{B}_y) \subset \mathcal{B}_x$. By the basic properties of conditional expectations, if Z is any $\sigma(T_{y_0})$-measurable, non-negative numerical random variable, then

$$\int_D \mathbb{I}_A(x) Z(x) d\rho(x) = \int_D \mathbb{E}_\rho \left[\mathbb{I}_A(x)|\sigma(T_{y_0})\right] Z(x) d\rho(x) \qquad \text{for all } D \in \sigma(T_{y_0}) \,. \tag{33}$$

(Note: This is where the fact that h depends on x only through $T_{y_0}(x)$ becomes relevant.) Since $h(y, \theta, \,.\,)$ is $\mathcal{B}_y$-measurable, $h(y, \theta, \,.\,) \circ T_{y_0}$ is measurable w.r.t. $\sigma(T_{y_0})$, and non-negative as h is a density. Hence

$$\begin{aligned}
\int_{T_{y_0}^{-1}(C)} \mathbb{I}_A(x) h(y_0, \theta, \,.\,) \circ T_{y_0} d\rho(x) &= \int_{T_{y_0}^{-1}(C)} \mathbb{E}_\rho \left[\mathbb{I}_A|\sigma(T_{y_0})\right](x) h(y_0, \theta, \,.\,) \circ T_{y_0} d\rho(x) \\
&= \int_{T_{y_0}^{-1}(C)} \mathbb{E}_\rho \left[\mathbb{I}_A|T_{y_0} = \,.\,\right] \circ T_{y_0}(x) h(y_0, \theta, \,.\,) \circ T_{y_0} d\rho(x) \\
&= \int_{T_{y_0}^{-1}(C)} k(A, \,.\,) \circ T_{y_0} dP_{\mathrm{X}}(x|\Theta = \theta) \,.
\end{aligned} \tag{34}$$

Thus, the Markov kernel k satisfies Eq. (17), and $T_{y_0} = T(\,.\,, y_0)$ is a sufficient statistic for the parametric model $P_{\mathrm{X}}(X|\Theta)$. □

The intuitive notion that the posterior is completely determined by the mapping T_{y_0} finds its technical expression in the measurability assumption used in Eq. (33): Dependence on T_{y_0} rather than X corresponds to measurability (in the abstract probability space Ω) w.r.t. the coarser σ-algebra $\sigma(T_{y_0} \circ C)$ rather than $\sigma(X)$. The fact that h can be expressed as a function which depends only on T_{y_0}, and hence is $\sigma(T_{y_0} \circ X)$-measurable, means that we can achieve equality in (33) by conditioning $\mathbb{I}_A$ only on $\sigma(T_{y_0} \circ X)$ (which yields the kernel k), rather than conditioning on all of $\sigma(X)$ (which would require the original posterior).

References

Barankin, E. M. and Maitra, A. P. (1963). Generalization of the Fisher-Darmois-Koopman-Pitman theorem on sufficient statistics. *Sankhya*, **25**, 217–244.

Blackwell, D. and Ramamoorthi, R. V. (1982). A Bayes but not classically sufficient statistic. *Annals of Statistics*, **10**(3), 1025–1026.

Darmois, G. (1935). Sur les lois de probabilite a estimation exhaustive. *C. R. Acad. Sci. Paris*, **260**, 1265–1266.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. John Wiley & Sons.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Annals of Statistics*, **7**(2), 269–281.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. A*, **222**, 309–368.

Halmos, P. R. and Savage, L. J. (1949). Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Ann. Math. Stat.*, **20**, 225–241.

Hipp, C. (1974). Sufficient statistics and exponential families. *Annals of Statistics*, **2**(6), 1283–1292.

Jeffreys, H. (1961). *Theory of probability*. Oxford University Press.

Kolmogorov, A. N. (1942). Sur l'estimation statistique des paramètres de la loi de Gauss. *Izv. Akad. Nauk SSSR Ser. Mat.*, **6**, 3–32.

Koopman, B. (1936). On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, **39**, 399–409.

Lindley, D. V. (1972). *Bayesian Statistics. A Review.* SIAM.

Neyman, J. (1935). Su un teorema concernente le cosiddette statistiche sufficienti. *Inst. Ital. Atti. Giorn.*, **6**, 320–334.

Pitman, E. J. G. (1936). Sufficient statistics and intrinsic accuracy. *Proceedings of the Cambridge Philosophical Society*, **32**, 567–579.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Harvard University Press.

Robert, C. P. (1994). *The Bayesian Choice*. Springer.

Schervish, M. J. (1995). *Theory of Statistics*. Springer.