# A Representation of the Posterior Mean for a Location Model

Nicholas G. Polson

*Biometrika*, Volume 78, Issue 2 (Jun., 1991), 426-430.

*Biometrika* (1991), **78**, 2, *pp.* 426–30
*Printed in Great Britain*

# A representation of the posterior mean for a location model

BY NICHOLAS G. POLSON
*Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, U.S.A.*

SUMMARY

An exact representation of the posterior mean is developed for a location model and the class of priors that are normal scale mixtures. The result makes use of the conditional distribution of the maximum likelihood estimator and Masreliez's theorem. Directions for future development are indicated.

*Some key words*: Bayesian inference; Conditional inference; Robustness; Score function.

## 1. INTRODUCTION

An exact representation for the posterior mean, $E(\theta|y)$, is given where $y$ is a $1\times n$ vector of observations from a location model, $f(x-\theta)$, and $\theta$ has a prior density, $p(\theta)$, that is a normal scale mixture. Let $L(\theta)$ denote the likelihood function and let $y=(\hat{\theta}, a)$ where $\hat{\theta}$ is the maximum likelihood estimator and $a$ is the maximal ancillary. The representation makes use of two results: the conditional distribution of the maximum likelihood estimator, $p(\hat{\theta}|\theta, a)$ (Barndorff-Nielsen, 1983), and a result of Masreliez (1975). It is shown that, under a normal prior, $E(\theta|y)$ can be represented as a linear transformation of the score function of $p(\hat{\theta}|a)$, where $p(\hat{\theta}|a)=\int p(\hat{\theta}|\theta, a)p(\theta)\, d\theta$. The representation can be viewed as a generalization of Masreliez's result that deals with the model, $X=\theta+\varepsilon$, $\theta\sim N(m, \tau^2)$ and represents the posterior mean in terms of the derivative of the logarithm of the marginal density for $X$. The Appendix gives a sufficient regularity condition and a rigorous proof of an interchange of derivative and integral sign required for Masreliez's theorem. A representation is also developed in the case where the prior density is a normal scale mixture. The results have a quantitative robustness appeal. For example, under any model and a normal prior the sensitivity of $E(\theta|y)$ with respect to aberrant observations is quantified by the behaviour of $d\log p(\hat{\theta}|a)/d\hat{\theta}$. For other Bayesian robustness results involving score functions or nonnormality, see Ericson (1969), Finucan (1971), Andrews, Arnold & Krutchkoff (1972), Dawid (1973), Box & Tiao (1973), O'Hagan (1979), Ramsey & Novick (1980), Smith (1983).

## 2. A REPRESENTATION OF THE POSTERIOR MEAN

### 2·1. *General discussion*

The following theorem is due to Masreliez (1975) and is the basis for our investigation. The Appendix gives a sufficient regularity condition and a rigorous justification for the required interchange of derivative and integral in the proof that is given here for completeness.

THEOREM. *Suppose that* $X=\theta+\varepsilon$ *with model* $f(X-\theta)$ *and* $\theta\sim N(m, \tau^2)$. *Assume that the regularity condition of the Appendix holds. Then*

$$E(\theta|X)=m-\tau^2\frac{d}{dX}\log p(X), \qquad (1)$$

*where* $p(X)$ *denotes the marginal beliefs about* $X$.

*Proof.* For simplicity take $p(\theta) \propto \exp(-\frac{1}{2}\theta^2)$ and note the identities

$$\theta p(\theta) = -\frac{d}{d\theta} p(\theta), \quad \frac{\partial}{\partial X} f(X-\theta) = -\frac{\partial}{\partial \theta} f(X-\theta).$$

Hence, integration by parts yields

$$\int \theta f(X-\theta) p(\theta)\, d\theta = -\int \frac{\partial}{\partial X} f(X-\theta) p(\theta)\, d\theta.$$

Now, if the regularity conditions in the Appendix hold for the interchange of derivative with respect to $X$ and integral sign then

$$E(\theta | X) = -\frac{p'(X)}{p(X)}. \tag{2}$$

For general $m$ and $\tau$ the result becomes (1). □

We now extend Masreliez's theorem to an independent and identically distributed sample of size $n$ from a location model. Note that the conditional distribution of $\hat{\theta}$, denoted by $p(\hat{\theta}|\theta, a)$ is of location form, denoted by $f(\hat{\theta}-\theta|a)$, and is given by (Barndorff-Nielsen, 1983)

$$p(\hat{\theta}|\theta, a) = c j(\hat{\theta})^{\frac{1}{2}} \frac{L(\theta)}{L(\hat{\theta})}, \tag{3}$$

where $j(\hat{\theta})$ is the observed Fisher's information matrix and $c$ is a normalizing constant. For a location model $j(\hat{\theta})$ is constant. The distribution given in (3) can be exact for problems outside the location case and is approximately true for a wide class of models.

The following theorem shows how Masreliez's theorem and (3) can be used to give us a representation of the posterior mean, $E(\theta|y)$, for the location problem under an arbitrary likelihood and a normal prior, $\theta \sim N(m, \tau^2)$.

THEOREM. *The posterior mean can be represented in terms of the score function of* $p(\hat{\theta}|a)$ *by*

$$E(\theta|y) = m - \tau^2 \frac{d}{d\hat{\theta}} \log p(\hat{\theta}|a), \tag{4}$$

*where* $y = (\hat{\theta}, a)$ *and* $p(\hat{\theta}|a) = \int p(\hat{\theta}|\theta, a) p(\theta)\, d\theta$.

*Proof.* By Bayes's theorem,

$$p(\theta|\hat{\theta}, a) = p(\hat{\theta}|\theta, a) p(\theta|a)/p(\hat{\theta}|a).$$

But as $a$ is ancillary $p(\theta|a) = p(\theta)$ and so

$$p(\theta|y) = f(\hat{\theta}-\theta|a) p(\theta)/p(\hat{\theta}|a).$$

Therefore, we can act as if the sample size is one with observation $\hat{\theta}$ with sampling distribution $f(\hat{\theta}-\theta|a)$. Hence, Masreliez's theorem applies to yield

$$E(\theta|y) = m - \tau^2 \frac{d}{d\hat{\theta}} \log p(\hat{\theta}|a),$$

where $p(\hat{\theta}|a) = \int p(\hat{\theta}|\theta, a) p(\theta)\, d\theta$ and $p(\hat{\theta}|\theta, a)$ is given by (3). □

For illustration, suppose that $y_i \sim N(\theta, 1)$ and $\theta \sim N(0, 1)$. Then $\hat{\theta} \sim N(\theta, n^{-1})$ and the score function of $p(\hat{\theta}|a)$ is given by

$$\frac{d}{d\hat{\theta}} \log p(\hat{\theta}|a) = -\frac{n}{n+1}\hat{\theta}.$$

But $\hat{\theta} = \bar{y}$ and (4) yields $E(\theta|y) = n\bar{x}/(n+1)$.

Section 2·2 considers the case where the prior is a normal scale mixture.

### 2·2. *Normal scale mixtures*

Suppose now that $\tau$ has a prior $p(\tau)$ and so $p(\theta)$ is a normal scale mixture, for example, the student, logistic or exponential power family (Andrews & Mallows, 1974; West, 1987). The following result extends Masreliez's theorem and the representation (4) to the class of priors that are normal scale mixtures. A similar result holds for normal location mixtures.

THEOREM. *Suppose that* $\theta|\tau \sim N(m, \tau^2)$ *and that* $\tau$ *has density* $p(\tau)$ *such that* $A = \int \tau^2 p(\tau)\, d\tau$ *is finite. Then*

$$E(\theta|y) = m - D(\hat{\theta}) \frac{d}{d\hat{\theta}} \log p(\hat{\theta}|a),$$

*where*

$$D(\hat{\theta}) = A \frac{d}{d\hat{\theta}} p^*(\hat{\theta}|a) \Big/ \frac{d}{d\hat{\theta}} p(\hat{\theta}|a), \quad p^*(\tau) = \frac{\tau^2 p(\tau)}{A}, \quad p^*(\hat{\theta}|a) = \int p(\hat{\theta}|a, \tau) p^*(\tau)\, d\tau.$$

*Proof.* By Fubini's theorem

$$E(\theta|y) = m + \frac{1}{p(\hat{\theta}|a)} \int \left( \int (\theta - m) f(\hat{\theta} - \theta|a) p(\theta|\tau)\, d\theta \right) p(\tau)\, d\tau.$$

The inner integral can be rewritten, using normality of $p(\theta|\tau)$, integration by parts and interchanging integral and derivative, as

$$\int (\theta - m) f(\hat{\theta} - \theta|a) p(\theta|\tau)\, d\theta = -\tau^2 \frac{d}{d\hat{\theta}} p(\hat{\theta}|a, \tau)$$

so that

$$E(\theta|y) = m - \frac{1}{p(\hat{\theta}|a)} \int \frac{d}{d\hat{\theta}} p(\hat{\theta}|a, \tau) \tau^2 p(\tau)\, d\tau.$$

By definition,

$$p^*(\tau) = \frac{\tau^2 p(\tau)}{A}, \quad p^*(\hat{\theta}|a) = \int p(\hat{\theta}|a, \tau) p^*(\tau)\, d\tau$$

so

$$E(\theta|y) = m - \frac{A}{p(\hat{\theta}|a)} \frac{d}{d\hat{\theta}} p^*(\hat{\theta}|a). \tag{5}$$

Therefore, by definition of $D(\hat{\theta})$,

$$E(\theta|y) = m - D(\hat{\theta}) \frac{d}{d\hat{\theta}} \log p(\hat{\theta}|a),$$

as required. □

## 3. DISCUSSION

An exact representation for the posterior mean for a location model under a nonnormal likelihood was developed. Analytical Bayesian computations for models outside the normal family are notoriously difficult. A notable exception is Spiegelhalter (1977, 1987) where some results for the double exponential and Cauchy likelihoods are given. The representation makes use of the conditional distribution of the maximum likelihood estimator and Masreliez's theorem. It establishes a link between Bayes and conditional inference by functionally relating the posterior mean and the score function of $p(\hat{\theta}|a)$. For other links, see Polson (1987) and Davison (1986). Similar representations hold for scale parameters and for the posterior variance. Further research lies in describing the behaviour of the posterior mean, in terms of a score function, for models outside the location family. An approximation to the score function of $p(\hat{\theta}|a)$, for example, via a Laplace-type approximation (Tierney & Kadane, 1986), would be of use for exploring this and other extensions, for example, in the multivariate and predictive cases.


## Acknowledgements

I would like to thank my supervisor Professor A. F. M. Smith for his suggestions and encouragement. I would also like to thank Nancy Reid and Larry Wasserman and the referee for their helpful comments.


## Appendix

### *Interchange of derivative and integral for* (2)

Masreliez's theorem requires that we can interchange derivative with respect to $X$ and the integral sign for the functional $\int f(X-\theta)p(\theta)\,d\theta$, where $p(\theta)$ is a normal density. By a linear transformation, the integral is equivalent to $\int f(\theta)p(X-\theta)\,d\theta$. Due to the fact that $p(.)$ is normal we have to check that the function $I(X)$ has derivatives of all orders, where

$$I(X)=\int \psi(\theta)\exp(X\theta-\theta^2)\,d\theta$$

and $\psi(\theta)=L(\theta)\exp(-\frac{1}{2}\theta^2)$. A sufficient regularity condition is that $\psi(\theta)$ is dominated by a polynomial. Let

$$f(X,\theta)=\psi(\theta)\exp(X\theta-\theta^2),$$

so that

$$\frac{d}{dX}f(X,\theta)=\theta\psi(\theta)\exp(X\theta-\theta^2).$$

In order for differentiation under the integral sign to hold, we require a dominating function, independent of $X$, for $|df(X,\theta)/dX|$ (Weir, 1973, p. 118). Unfortunately, there does not exist a dominating function that works for all $X\in R$. However, every point in $R$ lies in an open interval $(-x,x)$; so it will be enough to prove that there exists a local dominating function, that is for each positive real $x$, $df(X,\theta)/dX$ is dominated by a function in $L^1$, the space of Lebesgue integrable functions. Now, for $|X|<x$,

$$\left|\frac{d}{dX}f(X,\theta)\right|\leqslant|\theta\psi(\theta)|\exp(-\theta^2+x|\theta|). \tag{A1}$$

Let $g(\theta)$ equal the right-hand side of (A1). Clearly, $g(\theta)$ is continuous, therefore integrable on compact intervals and hence on $(-1-x,1+x)$. For $|\theta|\geqslant 1+x$, $\theta^2\geqslant(1+x)|\theta|$. This implies that in this region $g(\theta)$ is dominated by a polynomial times $e^{-|\theta|}$. Therefore $g(\theta)\in L^1$. Hence $I(X)$ possesses a first derivative and by reapplication has derivatives of all orders.

## References


ANDREWS, D. F., ARNOLD, J. C. & KRUTCHKOFF, R. G. (1972). Shrinkage of the posterior measure in the normal case. *Biometrika* **59**, 693–5.

ANDREWS, D. F. & MALLOWS, C. L. (1974). Scale mixtures of normality. *J. R. Statist. Soc.* B **36**, 99–102.

BARNDORFF-NIELSEN, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343–65.

BOX, G. E. P. & TIAO, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Reading, Mass: Addison-Wesley.

DAVISON, A. C. (1986). Approximate predictive likelihood. *Biometrika* **73**, 323–33.
DAWID, A. P. (1973). Posterior expectations for large observations. *Biometrika* **60**, 664–6.
ERICSON, W. A. (1969). A note on the posterior mean of a population mean. *J. R. Statist. Soc.* B **31**, 332–4.
FINUCAN, H. M. (1971). Posterior precision for non-normal distributions. *J. R. Statist. Soc* B **33**, 95–7.
MASRELIEZ, C. J. (1975). Approximate non-Gaussian filtering with linear state and observation relations. *IEEE. Trans. Autom. Control* **AC-20**, 107–10.
O'HAGAN, A. (1979). On outlier rejection phenomena in Bayes inference. *J. R. Statist. Soc.* B **41**, 358–67.
POLSON, N. G. (1987). Discussion of paper by D. R. Cox and N. Reid. *J. R. Statist. Soc.* B **36**, 24.
RAMSEY, J. O. & NOVICK, M. R. (1980). PLU robust Bayesian decision theory. *J. Am. Statist. Assoc.* **75**, 901–7.
SMITH, A. F. M. (1983). Bayesian approaches to outliers and robustness. In *Specifying Statistical Models*, Ed. J. P. Florens, M. Mouchart, J. P. Raoult, L. Simar and A. F. M. Smith, pp. 13–35. New York: Springer-Verlag.
SPIEGELHALTER, D. J. (1977). A test for normality against symmetric alternatives. *Biometrika* **64**, 415–9.
SPIEGELHALTER, D. J. (1987). Exact Bayesian inference on the parameter of a Cauchy distribution with vague prior information. In *Bayesian Statistics* **2**, Ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, pp. 743–51. Amsterdam: North-Holland.
TIERNEY, L. & KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Am. Statist. Assoc.* **81**, 82–6.
WEIR, A. J. (1973). *Lebesgue Integration and Measure.* Cambridge University Press.
WEST, M. (1987). On scale mixtures of normality. *Biometrika* **74**, 694–7.

[*Received April* 1990. *Revised December* 1990]