The Canadian Journal of Statistics
Vol. 22, No. 1, 1994, Pages 163–173
La Revue Canadienne de Statistique

Some aspects of the reparametrization of statistical models

Robert TIBSHIRANI and Larry WASSERMAN

University of Toronto and *Carnegie Mellon University*

Key words and phrases: Orthogonal parameters, reparametrizations.
AMS 1985 subject classifications: 62F15, 62F35.

ABSTRACT

Definitions are given for orthogonal parameters in the context of Bayesian inference and likelihood inference. The exact orthogonalizing transformations are derived for both cases, and the connection between the two settings is made precise. These parametrizations simplify the interpretation of likelihood functions and posterior distributions. Further, they make numerical maximization and integration procedures easier to apply. Several applications are studied.

RÉSUMÉ

Nous présentons des définitions pour des paramètres orthogonaux dans le contexte de l'inférence de Bayes et de l'inférence de vraisemblance. Les transformations d'orthogonalisation exactes sont obtenues dans les deux cas et le lien entre les deux approches est précisé. Ces paramétrisations simplifient l'interprétation des fonctions de vraisemblance et des distributions a posteriori. En outre, elles rendent l'application des procédures de maximisation numérique et d'intégration plus facile. Quelques applications sont étudiées.

1. INTRODUCTION

In this paper we discuss the reparametrization of statistical models. Loosely speaking, we seek parametrizations in which inference about the parameters of interest is independent of the nuisance parameters. A precise definition of this notion depends on the inferential context. In a Bayesian analysis, the posterior distribution of the parameters is of primary interest. For ease of interpretation, we may be interested in having the parameters *a posteriori* independent. To implement numerical methods reliably we may require more — we may want the posterior to be approximately standard Gaussian. Similarly, in a non-Bayesian analysis one might view a convenient parametrization as one for which the joint likelihood factors. Alternatively, we might like the restricted maximum-likelihood estimate for the parameter of interest to be in some sense independent of any nuisance parameters. This simplifies interpretation of the likelihood function and is likely to make numerical maximization techniques more successful. In the likelihood context, this has become known as an "orthogonal" parametrization. For more discussion on the advantages of reparametrization, see Hills and Smith (1992, 1993), Kass and Slate (1992, 1993).

Consider for example the Fieller-Creasy problem (Fieller 1954, Creasy 1954). Let X_1 and X_2 be independent Gaussian random variables with means θ_1 and θ_2 and unit variances. Suppose that $\psi = \theta_2/\theta_1$ is of specific interest. This problem is considered by Efron (1985) and Fraser and Reid (1989), among others. Figure 1(a) shows a contour plot of the likelihood as a function of ψ and θ_2, having observed $X_1 = 8$ and $X_2 = 4$. The contours show a dependence of ψ on θ_2. Figure 1(b) shows the likelihood as a function of ψ and $\beta = (\theta_1^2 + \theta_2)^{\frac{1}{2}}$. There is little dependence of ψ on β.

We study various definitions of independent and orthogonal parameters and give some automatic methods for their construction. The transformations that we derive are data-dependent, an important fact to be kept in mind when considering their application. In Section 2 we discuss the definitions and their relationships, along with previous work on this subject. Section 3 describes a method for constructing orthogonal parameters in the Bayesian and non-Bayesian cases. Section 4 illustrates these methods in a variety of examples. In Section 5 we consider briefly the relationship between data-dependent and data-independent transformations. Finally, Section 6 contains a discussion of the results.

2. INDEPENDENT AND ORTHOGONAL PARAMETERS

Let $\mathcal{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ be a random sample of d-dimensional vectors, each $\mathbf{Y}_i$ independently and identically distributed with density $f_{\mathbf{Y}}(\mathbf{y}, \boldsymbol{\theta})$, where the unknown parameter $\boldsymbol{\theta}$ lies in a subset of $\mathbb{R}^{p+1}$. Denote the components of $\boldsymbol{\theta}$ by $(\psi, \gamma_1, \ldots, \gamma_p)$. We assume that ψ is the parameter of interest and that $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)$ are nuisance parameters. The likelihood function will be denoted by $L(\boldsymbol{\theta})$, and we assume that $L(\boldsymbol{\theta})$ is bounded. Without loss of generality, we further assume that the supremum of $L(\boldsymbol{\theta})$ is one. A prior density (if available) is denoted by $\pi(\boldsymbol{\theta})$, and the corresponding posterior density is denoted by $\pi(\boldsymbol{\theta} \mid \mathcal{Y})$. The maximum-likelihood estimate is $\hat{\boldsymbol{\theta}} = (\hat{\psi}, \hat{\gamma}_1, \ldots, \hat{\gamma}_p)$.

We consider a transformation of the parameters $(\psi, \gamma_1, \ldots, \gamma_p)$ to $(\alpha, \beta_1, \ldots, \beta_p)$, where α is a strictly monotonic function of ψ and each β_j is an invertible function of γ_j when all the remaining parameters are held fixed. We restrict α to be a function of ψ only, so that inferential statements about α may be directly transferred to statements about ψ, the parameter of interest. In practice, we may leave ψ untransformed. We introduce the transformation α mainly for theoretical convenience.

Before proceeding we need some notation. Let $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$, and let $i_{\alpha,\beta_j} = \mathcal{E}\{-\partial^2 \log L(\alpha, \boldsymbol{\beta})/\partial\alpha\, \partial\beta_j\}$ be the elements of the expected Fisher information matrix. Similarly, let $j_{\alpha,\beta_j} = -\partial^2 \log L(\alpha, \boldsymbol{\beta})/\partial\alpha\, \partial\beta_j$ be the elements of the observed information matrix. Finally, let $\tilde{\boldsymbol{\theta}} = (\alpha, \beta_1, \beta_2, \ldots, \beta_p)$.

The following are several possible definitions for orthogonality of α and β:

D1. *Independence.* α and $\boldsymbol{\beta}$ are *a posteriori* independent.

D2. *Likelihood factorization.* $L(\tilde{\boldsymbol{\theta}}) = L_1(\alpha)L_2(\boldsymbol{\beta})$ for some L_1 and L_2.

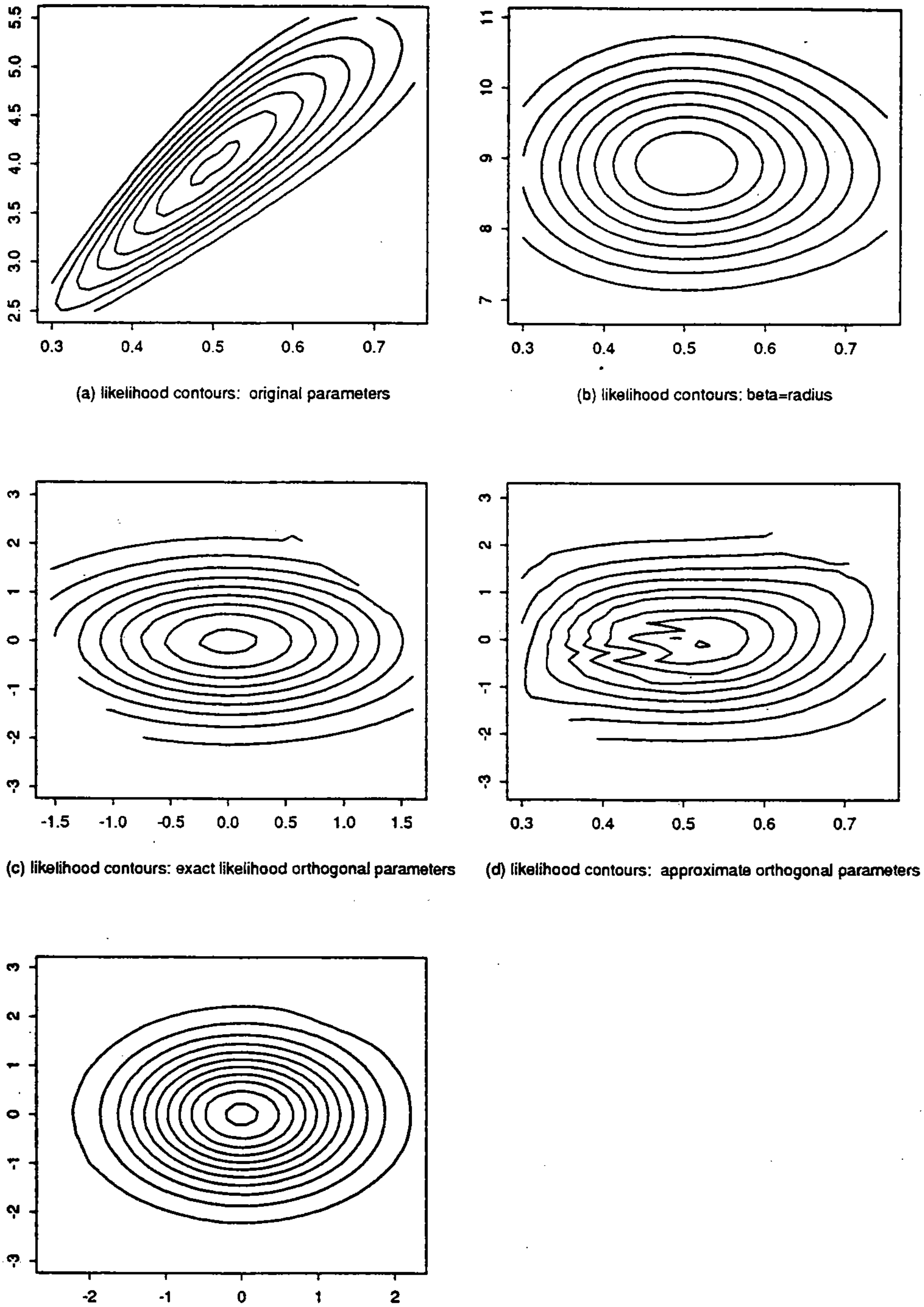
D3a. *Global observed orthogonality.* $j_{\alpha,\beta_j}(\alpha, \boldsymbol{\beta}) = 0$ for all $(\alpha, \boldsymbol{\beta})$, $j = 1, \ldots, p$.

D3b. *Local observed orthogonality.* $j_{\alpha,\beta_j}(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = 0$, $j = 1, \ldots, p$.

D4. *Expected orthogonality.* $i_{\alpha,\beta_j}(\alpha, \boldsymbol{\beta}) = 0$ for all $(\alpha, \boldsymbol{\beta})$, $j = 1, \ldots, p$.

Some relationships may be seen among these definitions. If the prior $\pi(\alpha, \boldsymbol{\beta})$ is constant, then D1 is equivalent to D2. Also, D2 implies both D3a and D3b. The reverse implications do not hold in general. Property D4 makes sense only if the reparametrizations don't involve the data $\mathcal{Y}$. In that case, if we add the proviso that D2, D3a, and D3b hold for all $\mathcal{Y}$, we have that D2 implies D4 and D3a implies D4. The condition D4 is the standard definition as employed by Huzurbazar (1950), Jeffreys (1961), and Cox and Reid (1987).

In general, none of these definitions leads to parametrizations that are unique. Hence, it is convenient to introduce a notion that is stronger than orthogonality. We shall say that α and $\boldsymbol{\beta}$ are *Bayesian-orthogonal with respect to* π if, *a posteriori*, α and $\boldsymbol{\beta}$ possess a standard $p+1$-dimensional multivariate Gaussian distribution. And we shall say that α and $\boldsymbol{\beta}$ are *likelihood-orthogonal* if $L(\alpha, \boldsymbol{\beta})$ is proportional to a standard $p+1$-dimensional



(a) likelihood contours: original parameters

(b) likelihood contours: beta=radius

(c) likelihood contours: exact likelihood orthogonal parameters

(d) likelihood contours: approximate orthogonal parameters

(e) posterior contours: Bayesian orthogoanl parameters

FIGURE 1: Contours for the Fieller-Creasy example. Unless labelled otherwise, in all contour plots the parameter of interest varies horizontally and the nuisance parameter varies vertically.

multivariate Gaussian density function. Note that Bayesian orthogonality implies D1 and likelihood orthogonality implies D2, D3a, and D3b.

Several researchers have studied reparametrizations. The problem is considered from the Bayesian point of view in Huzurbazar (1950), Jeffreys (1961, Section 4.31), Dagenais and Liem (1981), Naylor and Smith (1982), Smith *et al.* (1985, 1987), and Albert (1988). Dagenais and Liem (1981) use a parametric family of transformations to achieve approximate normality of the posterior. Albert uses data-analytic techniques to find appropriate transformations. This paper provides a rigorous, automatic version of these methods. Smith *et al.* (1985, 1987) provide excellent overviews of the numerical problems in Bayesian inference, and they discuss the need to reparametrize. Hills and Smith (1992, 1993) discuss graphical methods for choice and assessment of model parametrizations.

From the likelihood point of view, orthogonal parameters were studied by Cox and Reid (1987). They show that the expected orthogonalizing transformations are given by solving a set of differential equations. In general, these equations are difficult to solve. Furthermore, they produce a pleasant reparametrization only in some average sense, rather than in the observed likelihood. In the context of sampling properties of likelihood functions, the Cox-Reid approach is appropriate. However, our motivation is to transform the observed likelihood into a more convenient form. Holland (1973) showed that for a vector parameter it is not possible in general to variance-stabilize, that is, make the expected information matrix equal to the identity matrix; hence it is impossible in general to make all off-diagonal elements equal to zero. This latter fact was also shown by Cox and Reid (1987). Property D4 is a weaker requirement, involving only the (α, β_j) components.

The remainder of this paper is devoted to the construction of orthogonalizing transformations. A point that has not been emphasized in the literature is that orthogonalizing with respect to posterior distributions is, in general, in conflict with the idea of orthogonalizing with respect to the likelihood function, even if constant priors are employed. This will be made precise in Section 3, where we give the exact orthogonalizing transformations in both cases. These transformations do coincide for a particular prior that, generally, is not constant.

3. METHODS FOR REPARAMETRIZATION

3.1. The Bayesian Case.

For simplicity we begin with the two-parameter case $\boldsymbol{\theta} = (\psi, \gamma)$. It is easy to see that a transformation to Bayesian orthogonality always exists. We set $\beta(\psi, \gamma) = \Phi^{-1}\big(F_{\gamma|\psi}(\gamma)\big)$ and $\alpha(\psi) = \Phi^{-1}\big(F_{\psi}(\psi)\big)$, where Φ is the standard Gaussian distribution function, $F_{\gamma|\psi}$ is the conditional posterior distribution function for γ given ψ, and F_{ψ} is the marginal posterior distribution function for ψ. To see that this is the orthogonalizing transformation, note that by construction, α is marginally standard Gaussian and β is standard Gaussian conditional on each fixed value of α. Hence, α and β are jointly standard Gaussian.

Note that for each fixed value of ψ, the cumulative distribution transform of γ is uniform, and hence independent of ψ. This fact can be exploited to derive independent posterior parameters if independence is all that is required.

Now we give an algorithm for approximating α and β. Let $(\psi_1, \psi_2, \ldots, \psi_k)$ be a grid of ψ values, and let $(\gamma_1, \gamma_2, \ldots, \gamma_k)$ be a grid of γ values. Estimate $F_{\gamma|\psi}$ and F_{ψ} by

$$\bar{F}_{\gamma|\psi_j}(\gamma_r) = \frac{\sum_{i=1}^{r} L(\psi_j,\, \gamma_i)\pi(\psi_j,\, \gamma_i)}{\sum_{i=1}^{k} L(\psi_j,\, \gamma_i)\pi(\psi_j,\, \gamma_i)}, \qquad r, j = 1, 2, \ldots, k,$$

and

$$\bar{F}_\psi(\psi_r) = \frac{\sum_{j=1}^{r}\sum_{i=1}^{k} L(\psi_j, \gamma_i)\pi(\psi_j, \gamma_i)}{\sum_{j=1}^{k}\sum_{i=1}^{k} L(\psi_j, \gamma_i)\pi(\psi_j, \gamma_i)}, \qquad r = 1, 2, \ldots, k.$$

Clearly, if L and π are reasonably well behaved, then the estimates will converge to the true distributions as the grid size increases. Plugging the estimates into the expressions for α and β yields transformations $\bar{\beta}$ and $\bar{\alpha}$, say. These are approximations to the exact orthogonalizing transformations.

It may not be necessary to carry out the orthogonalization in this way. A useful approximate Bayesian orthogonalization can easily be derived. First, note that if the distribution of γ is Gaussian for each fixed ψ, but with mean and variance possibly depending on ψ, then the transformation β is

$$\beta(\psi, \gamma) = \frac{\gamma - m_\gamma(\psi)}{s_\gamma(\psi)},$$

where $m_\gamma(\psi) = \mathcal{E}(\gamma|\psi)$ and $s_\gamma(\psi) = \{\mathcal{V}ar(\gamma|\psi)\}^{\frac{1}{2}}$. In general, this form of β may be regarded as an approximation to the exact transformation. As long as γ is approximately Gaussian for each fixed ψ, then this approximate transformation will suffice. Note that this is a much weaker assumption than assuming that the joint distribution is approximately Gaussian.

Our algorithm may be simplified by using this approximate transformation. We simply set $\bar{\beta} = \{\gamma - \bar{m}_\gamma(\psi)\}/\bar{s}_\gamma(\psi)$, where $\bar{m}_\gamma(\psi)$ and $\bar{s}_\gamma(\psi)$ are the estimates of $m_\gamma(\psi)$ and $s_\gamma(\psi)$ obtained from $\bar{F}_{\gamma|\psi}$ and $\bar{F}_\psi$. In case this transformation is unsuccessful, we can always resort to the exact form.

EXAMPLE 1 (The Fieller-Creasy problem). For the problem described in the introduction, Figure 1(e) shows the posterior contours corresponding to the exact Bayesian-orthogonalizing transformations, while Figure 1(d) displays the likelihood contours after this Bayesian transformation has been carried out. An important point is that the Jacobian of the approximate transformation is a function of ψ only. Thus a plot of the posterior contours corresponding to the approximate transformations would differ from Figure 1(d) only by some stretching factor in the ψ-direction.

So far we have restricted ourselves to the two-parameter case. With more than two parameters, the exact reparametrization is

$$\begin{aligned} \beta_1 &= \Phi^{-1}\{F_{\gamma_1|\psi}(\gamma_1)\}, \\ \beta_2 &= \Phi^{-1}\{F_{\gamma_2|\psi,\gamma_1}(\gamma_2)\}, \\ &\vdots \end{aligned} \tag{1}$$

where the subscripted F denotes the conditional distribution given the subscripted random variables. The transformation α is as before. The approximate version is the same as that given in the two-parameter case with the mean and standard deviation of the conditional distribution of γ_j given ψ, $\gamma_1, \ldots, \gamma_{j-1}$.

3.2. The Non-Bayesian Case.

In general, the reparametrization given in the previous section does not orthogonalize the likelihood, even if the prior $\pi(\psi, \gamma)$ is constant. To see this, note that

$$\pi(\alpha, \beta | \mathcal{Y}) = L(\alpha(\psi), \beta(\theta))\pi(\alpha(\psi), \beta(\theta))J(\psi, \gamma : \alpha, \beta)$$

In a sense, (5) is doing too much, since it (approximately) produces a set of mutually orthogonal parameters. If only ψ is of interest, the approximation (6) may be adequate. In Tibshirani and Wasserman (1989), we examine this approximation in the Box-Cox transformation model and find that it works reasonably well.

4. FURTHER EXAMPLES

EXAMPLE 2 (Exponential model). It is instructive to consider a one-parameter example. Here the goal is to transform to normality. Let $Y_1, \ldots, Y_n$ be exponential with mean ψ. Then the likelihood-normalizing transformation $\alpha(\psi) = \operatorname{sign}(\psi - \hat{\psi})\{\bar{Y}/\psi - 1 - \log(\bar{Y}/\psi)\}^{\frac{1}{2}}$. Note that, to first order, $\mathcal{E}\alpha^2 = 0$, so the parametrization is only weakly data-dependent. This reparametrization agrees with the Bayesian-normalizing transformation for the prior

$$\pi_0(\psi) \propto \operatorname{sign}(\psi - \hat{\psi}) \frac{ne^{n\bar{Y}/\psi}(\psi - \bar{Y})}{\{\bar{Y}/\psi - 1 - \log(\bar{Y}/\psi)\}^{\frac{1}{2}}}.$$

EXAMPLE 3 (Gamma model). Albert (1987) considered the gamma model

$$L(\psi, \gamma) = \frac{\gamma^{-n\psi}}{\Gamma^n(\psi)} p^{\psi} e^{-s/\gamma}$$

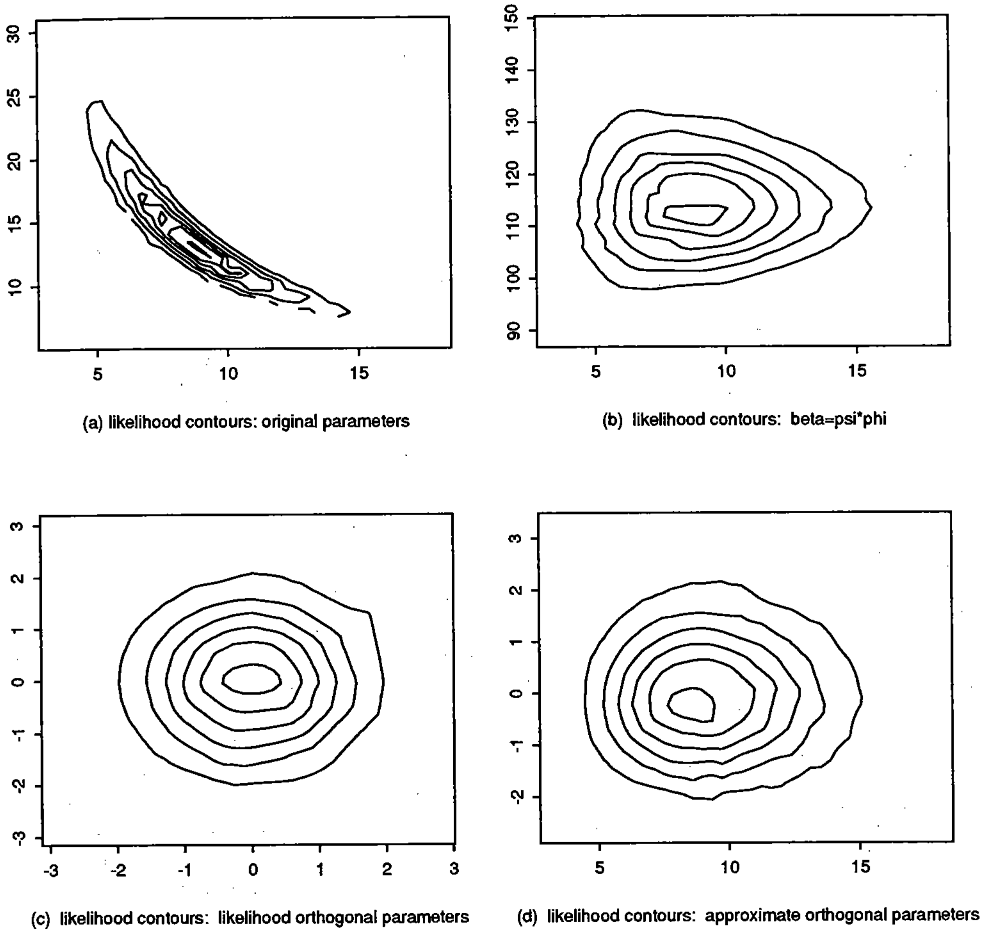
with data $n = 20$, $s = 2269$, $\log p = 93.47$. The expected orthogonalizing transformation is $\beta = \psi\gamma$, as derived by Cox and Reid (1987); Albert also deduced this reparametrization from a graphical method. Figure 2 shows the likelihood contours corresponding to the original parameters, to the Cox-Reid parametrization, and to the exact and approximate likelihood orthogonalization. The virtues of reparametrizing this problem are discussed by Albert (1987).

The components of the approximate reparametrization, $m(\psi)$ and $s(\psi)$, are shown in Figure 3(a) and (b). In the important range for ψ (5 to 15), they both are approximately of the form $1/\psi$. Now $(\gamma - \psi^{-1})/\psi^{-1} = \psi\gamma - 1$, which is equivalent to the expected orthogonalizing transformation. Figure 3(c) shows a plot of the approximate β versus $\psi\gamma$ for each of the grid points. If the two transformations were equivalent, then the parametrizations would be linear functions of each other. This is almost the case.

5. DATA-DEPENDENT AND DATA-INDEPENDENT TRANSFORMATIONS

The orthogonalizing transformations used in this paper are data-dependent. Here, we briefly investigate this dependence. We simulate datasets from a gamma model (Example 3), with parameters set equal to the maximum-likelihood estimates from the original fit. For each dataset, we computed $m(\psi)$ and $s(\psi)$. The results of 10 such simulations are shown in Figure 3(d) and (e). There is little change in the transformations, and this was confirmed with a larger number of simulations.

This suggests the following method for finding data-independent transformations that will produce approximate orthogonalization. Generate $\theta_1, \theta_2, \ldots$ from the prior π. For each θ_i simulate n observations from $f(Y, \theta_i)$. Compute the Bayesian-orthogonalizing transformations each time, and average them. This transformation may be fitted to a curve from a convenient class of functions (for example, polynomials). The resulting transformation is then used as an approximate, data-independent orthogonalizing transformation. We have not investigated this technique here, but see Slate (1991) for a recent discussion.



(a) likelihood contours: original parameters

(b) likelihood contours: beta=psi*phi

(c) likelihood contours: likelihood orthogonal parameters

(d) likelihood contours: approximate orthogonal parameters

FIGURE 2: Likelihood contours for the gamma example.

6. DISCUSSION

The use of numerical techniques and the need for interpretability make reparametrizations of statistical models useful. We have made the notion of orthogonalizing transformations precise in both the Bayesian and non-Bayesian cases. Further, we have emphasized the fact that orthogonalization with respect to the likelihood function is different from orthogonalization with respect to the posterior.

Some interesting questions that still need to be addressed are:

(1) What are the sampling properties of data-dependent orthogonalizing transformations? This is important for understanding the relationship between data-dependent and data-independent transformations.

(2) What is the theoretical significance of the prior that links the likelihood- and Bayesian-orthogonalizing transformations? Since this function is constant if and only if the observed likelihood is Gaussian, might this be used as a diagnostic to evaluate the nonnormality of the likelihood function? See Hills and Smith (1992, 1993) and Kass and Slate (1992, 1993) for discussion on this point.

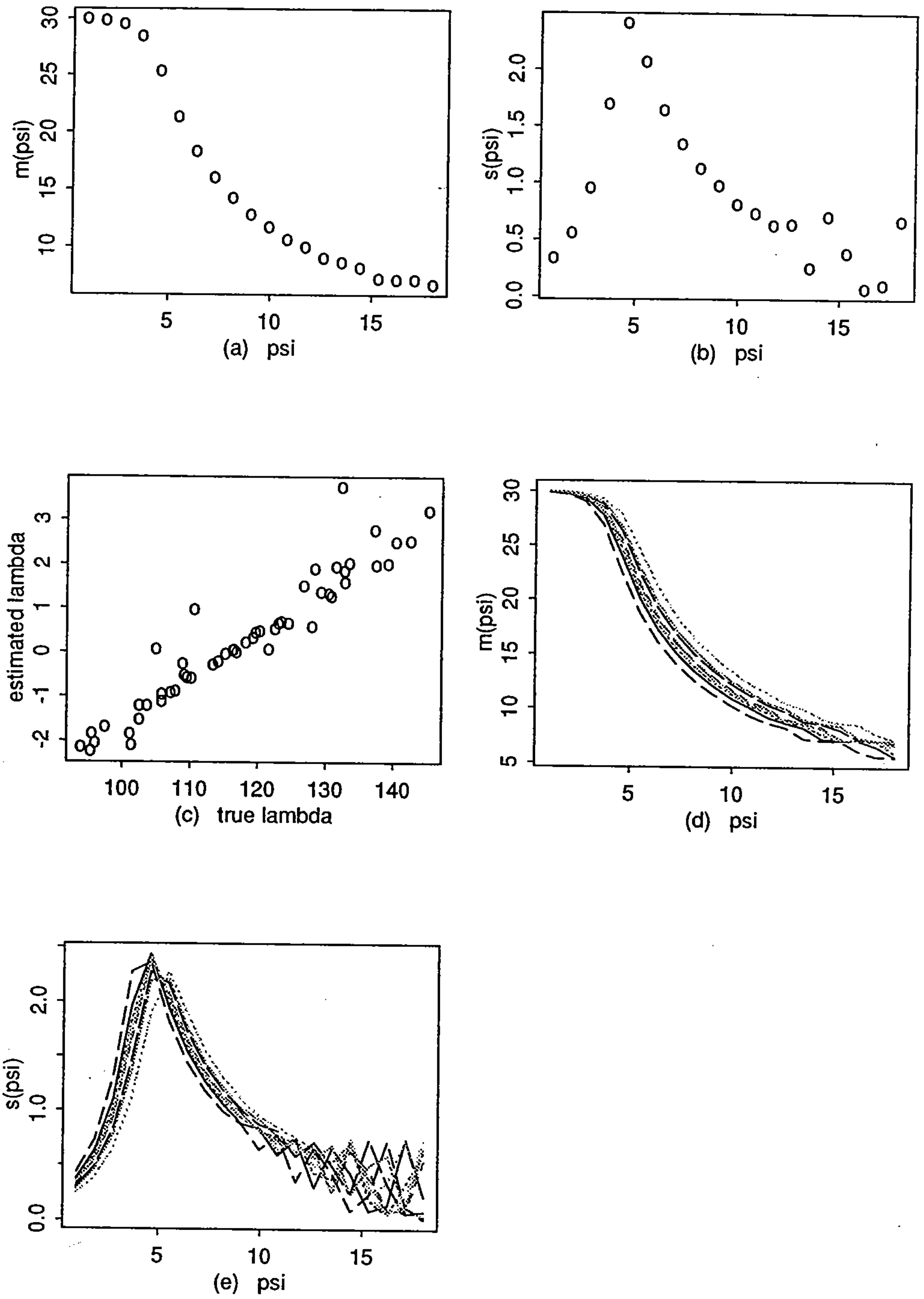


FIGURE 3: Plots for the gamma example.

(3) What are efficient methods to carry out orthogonalization in high dimensions?

ACKNOWLEDGEMENTS

We would like to thank Nancy Reid and Tom DiCiccio for helpful discussions, J. Kadane for bringing Albert's paper to our attention, and a referee for helpful comments.

REFERENCES

Albert, J.H. (1987). Nuisance parameters and the use of exploratory graphical methods in a Bayesian analysis. Technical report, Bowling Green State University.

Bates, D.M., and Watts, D.G. (1988). *Nonlinear Regression Analysis and Its Application.* Wiley, New York.

Cox, D.R., and Reid, N. (1987). Orthogonal parameters and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. Ser. B*, 49(1), pp. 1–39.

Creasy, M.A. (1954). Limits for the ratio of means. *J. Roy. Statist. Soc. Ser. B*, 16, 186–194.

Dagenais, M.G., and Liem, T.C. (1981). Numerical approximations of marginals from "well-behaved" joint posteriors. *J. Statist. Comput. Simulation*, 12, 157–173.

Efron, B. (1985). Bootstrap confidence intervals for parametric problems. *Biometrika*, 72, 45–58.

Fieller, E.C. (1954). Some problems in interval estimation. *J. Roy. Statist. Soc. Ser. B*, 16, 175–183.

Fraser; D.A., and Reid, N. (1989). Adjustments to profile likelihood. *Biometrika*, 76, 477–488.

Hills, S.E., and Smith, A.F.M. (1992). Parameterization issues in Bayesian inference. *Bayesian Statistics 4* (J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, *eds.*), Clarendon Press, Oxford, 227–246.

Hills, S.E., and Smith, A.F.M. (1993). Diagnostic plots in Bayesian inference. *Biometrika*, 80, 61–74.

Holland, P. (1973). Covariance stabilizing transformations. *Ann. Statist.* 1(1), 84–92.

Huzurbazar,

Jeffreys, H. (1961). *Theory of Probability.* Third Edition. Oxford Univ. Press, London.

Kass, R.E., and Slate, E.H. (1992). Reparameterization and diagnostics of posterior non-normality. *Bayesian Statistics 4* (J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, *eds.*), Clarendon Press, Oxford, 289–305.

Kass, R.E., and Slate, E.H. (1993). Some diagnostics for likelihood and posterior non-normality. *Ann. Statist.*, to appear.

Naylor, J.C., and Smith, A.F.M., (1982). Applications of a methods for the efficient computation of posterior distributions. *Appl. Statist.*, 31(3), 214–225.

Slate, E. (1991). Reparameterization of statistical models. Ph.D. Thesis, Department of Statistics, Carnegie Mellon University.

Smith, A.F.M., Skene, A.M., Shaw, J.E.H., Naylor, J.C., and Dransfield, M. (1985). The implementation of the Bayesian paradigm. *Comm. Statist. A*, 14, 1079–1102.

Smith, A.F.M., Skene, A.M., Shaw, J.E.H., and Naylor, J.C. (1987). Progress with numerical and graphical methods for practical Bayesian statistics. *The Statistician*, 36, 75–82.

Tibshirani, R.J., and Wasserman, L. (1989). Some aspects of the reparameterization of statistical models. Technical Report 449, Carnegie-Mellon University.

Tierney, L., and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.*, 81, 82–86.

Received 8 April 1991
Revised 18 May 1993
Accepted 14 October 1993

Department of Preventative Medicine and Biostatistics
University of Toronto
Toronto, Ontario
Canada M5S 1A8

Department of Statistics
Carnegie Mellon University
Pittsburgh, PA
U.S.A. 15213-3890