The Scale Invariant Prior and Its Generalizations

Stephen P. Smith

Abstract. The scale invariant prior is revisited, for a single variance parameter and for a variance-covariance matrix. These results are generalized to develop different scale invariant priors where probability measure is assigned through the sum of variance components that represent partitions of total variance, or through a sum of variance-covariance matrices representing partitions of a total variance-covariance matrix.

1. Introduction

The attribute of "invariance" implies both the presence of information when making a measurement, and also the complete lack of information. That is, to assert that a measurement looks the same from all points of view implies the certainty of information, but it also implies that the measurement cannot be distinguished by alternative reference frames that may rest beyond simple empiricism. Invariance of this sort typifies relativity that is found depending on a frame of reference, making a circularity in the meaning of measurement[1].

Translation invariance applies to parameters that are free to vary on the real line, in all directions. This invariance can be described in geometric terms as points in space that are relative to a frame of reference that implies an origin and a coordinate system that propagates out from the origin. Its always the visible difference between a position and the origin that carries invariant information, whereas an unreferenced point in absolute space carries no visible information. The density function that treats all unreferenced locations equally is the flat prior, or a constant, and it represents no information. To measure distance, however, also requires a scale and if the scale is also arbitrary then scale invariance is also indicated when there is no information to prefer one scale over another.

The scale invariant prior can be justified purely on the arbitrariness offered by the units of measurement, absent a more abstract construct of invariance. It's the measurements themselves that are found relative to the standard of measurement or the yard stick that's implied by the frame of reference, and therefore statistical inferences can be made invariant to the selected standard by the appropriate selection of a vague prior that is impartial to all possible standards.

For example, what is known about a sample of observations, denoted by the vector $\mathbf{y}$, is represented by the transformed vector $c \times \mathbf{y}$ where c is an arbitrary constant that can be anything. Statistical inferences are made invariant to the constant c, but note this is far

[1]In the subject of physics (i.e., special and general relativity), the only measurements that are considered meaningful are invariants.

less demanding than if the vector **y** was transformed, element by element, by an arbitrary monotone transformation, where useful information reduces to rank information which is a result of a more ambitions demand for invariance. The simplicity offered by the present investigation is a direct result of limiting consideration to scale invariance.

The scale invariant prior attaches directly to the parameter σ that measure dispersion, like the standard deviation represented by the elements of **y**, but the statical model can be more complicated than the simple case that involves point estimates of location and variation. Nevertheless, transforming **y** to c×**y** results in changing σ to c×σ, and the appropriate prior that is impartial to the choice of c is the well known scale invariant prior:

$$\pi(\sigma) = \frac{1}{\sigma}$$

While this prior is improper, because it integrates to infinity, we may compare probabilities that are evaluated over bounded regions to demonstrate scale invariance as Berger (1980, pages 70-71) does. Define the following probability:

$$P(\sigma \in A) = \int_A \frac{1}{\sigma} d\sigma$$

Scale invariance is implied by the identity $P(\sigma \in A) = P(\sigma \in c^{-1}A)$, following Berger, which carries a transformation where the Jacobian cancels because of the form of the prior.

It is also useful to reexpress the same scale invariant prior but in terms of σ^2 (a variance component) by going through a standard change of variables, and this produces the result:

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2} \qquad (1)$$

For a simple linear model that describes **y**, containing only fixed effects and only one random effect representing a set of residuals, the flat prior can be used to treat all the fixed effects, and the above scale invariant prior can be used to treat the dispersion parameter that's attached to the random residues (see Tanner 1993, page 12). Then the Bayesian posterior distribution can be sampled directly whereby the fixed effects are sampled from a multivariate normal distribution (conditional on σ^2), and the dispersion parameter (in the form of σ^2) is sampled from an inverted Chi-square distribution. The simulation can now be extended by performing Gibb's updates, thus sampling from the posterior distribution for all the fixed effects and the dispersion parameter, together.

For the general variance component model, where σ^2 can be partitioned into several variance components, the seductive approach is to assign the above scale invariant prior to each variance component, and then continue with a Bayesian simulation that follows the obvious path of sampling from the multivariate normal distribution, followed by sampling from inverted Chi-square distributions in turn for each variance component,

and putting the whole thing together by making Gibb's update cycles (e.g., Gelfand et al, 1990; Wang, Rutledge and Gianola, 1993)[2]. While this simple adaptation looks attractive on first impression, this is actually very misleading. The application of scale invariance must be reconnected to the yard stick that makes measurements on actual observations, **y**, and so far there is nothing that extends the discovered nicety found for σ^2 to the rest of the variance components. A bigger problem is that its possible to over-weigh the posterior distribution with improper priors that end up making the posterior distribution improper (e.g., Hobert and Casella 1996; Daniels 1999). Therefore, a fix is sought to find the appropriate prior for multiple variance components under the sought goal of scale invariance that's applied consistently, perhaps at the expense of having to abandon the convenient draws from the inverted Chi-square distribution that fitted well as Gibb's updates.

Daniels (1999) provides an adequate list of alternative non-informative priors that may be adapted for multiple variance components in hierarchical models, even multivariate variance-covariance matrices. Understand that this makes my present paper much less original in a broad sense, but my paper has a very narrow focus. My particular goal is to extend scale invariance in a coherent way from first principles, thus generalizing the results to multiple variance components (in Section 2) and variance-covariance matrices (in Section 3).

2. Scale Invariant Prior for Multiple Variance Components

For the purpose of illustration, the total variance will be assumed to be composed of three variance components: $\sigma^2 = x^2 + y^2 + z^2$. If the data had only unique representations of the random factors in combination, then all that can be estimated is the total variance, σ^2. In any regard, the improper prior given by (1) still applies. But this prior represents a marginal distribution, involving one parameter, not three. Two parameters, suitably transformed, need to be integrated out of the joint distribution to recover (1). It's the joint distribution that is sought. Because the variance components add to make the total variance[3], because changing **y** to $c\times\mathbf{y}$ impacts the sum in the same way as the components, i.e., $c^2\times\sigma^2 = c^2\times x^2 + c^2\times y^2 + c^2\times z^2$, the natural joint distribution to seek is of the form $\pi(\sigma^2)=\pi(x^2 + y^2 + z^2)$ where probability measure is assigned democratically through the total variance. There are other less sensible possibilities that may single out the variance components while maintaining the resemblance of symmetry, but the choice may also complicate the subsequent integration that is implied. The adopted form, $\pi(x^2 + y^2 + z^2)$, works by giving a tractable integration.

---

[2]For cases where the degree of belief parameter was set to zero.

[3]By comparison, the square-root of the total variance and each of the individual variance components have a different relationship, albeit one implied by this additive relationship involving the variances.

Now make the following change of variables, coming with the specified ranges of integration.

$x^2=x^2$	$0 < x^2 < t^2$
$t^2=x^2+y^2$	$0 < t^2 < \sigma^2$
$\sigma^2 = x^2+y^2+z^2$	$0 < \sigma^2 < \infty$

Substituting the changed variables into the joint distribution simply generates $\pi(\sigma^2)$, and the Jacobian is 1. The integration that reproduces (1) is given below.

$$\frac{1}{\sigma^2} = \int_0^{\sigma^2} \int_0^{t^2} \pi(\sigma^2) dx^2 dt^2 = \frac{1}{2}\sigma^4 \pi(\sigma^2)$$

Therefore, the joint prior is given by the following,

$$\pi(x^2 + y^2 + z^2) \propto (x^2 + y^2 + z^2)^{-3} \qquad (2)$$

and while this prior is improper, its not as improper or as badly behaved as the prior $\pi_1(x^2, y^2, z^2) = x^{-2} \times y^{-2} \times z^{-2}$ would be. Now gone is an easy draw from an inverted Chi-square distribution during a Bayesian simulation. Nevertheless, the Metropolis algorithm or one of its variants are available to substitute for some of the Gibb's updates corresponding to the variance components. Because the form of the posterior distribution is not complicated much by (2), the development of custom software is not impeded.

Its immediate how to generalize this result for a different number of variance components; its always the total variance raised to the power that's the negative number of variance components.

3. Scale Invariant Prior for Multivariate Variance-Covariance Matrices

The developments in Section 2 are not entirely satisfactory, given that restricting the prior to be of the form $\pi(x^2, y^2, z^2) = \pi(x^2 + y^2 + z^2)$ carries some arbitrariness that does not define the prior uniquely. As previous noted, an alternative scale invariant prior $\pi_1(x^2, y^2, z^2)$ exists. In this case, the result of a change of scale transformation leads to a Jacobian that cancels with the form of the prior, leading to the declaration that π_1 is a scale invariant prior. Therefore, there is an alternative way to identify scale invariant priors based on the Jacobian, and not needing the integration used in Section 2. Integration can still be attempted, but introducing the same change of variables, i.e., turning x^2, y^2 and z^2 into x^2, t^2 and σ^2, causes the integration of $\pi_1(x^2, t^2-x^2, \sigma^2-t^2)$ to diverge to infinity as soon as x^2 is integrated between 0 and t^2 in the very first integral, indicating that this improper prior is more improper than (2). Despite this weakness, π_1

is interesting because in relates to a variance matrix, **V**, and to the vague prior density (Press, 2003, Section 5.4.2) that is given by $\pi_2(\mathbf{V}, k) = |\mathbf{V}|^{-k}$ where $|\cdot|$ is the matrix determinant and when k=1 for a diagonal matrix **V**. Precisely, $\pi_1=\pi_2(\mathbf{V}, 1)$ when:

$$\mathbf{V} = \begin{bmatrix} x^2 & & \\ & y^2 & \\ & & x^2 \end{bmatrix}$$

In this particular case, all the off-diagonal covariances found in **V** are zero, but the result generalizes for a general variance-covariance matrix. Setting k=½(p+1) in $\pi_2(\mathbf{V}, k)$ where p is the order of that matrix **V** returns the Jeffreys invariant prior[4], which is also the vague prior first introduced by Geisser and Cornfield (1963). To demonstrate scale invariance introduce the diagonal matrix $\mathbf{W}_{p\times p} = \text{diag}\{ w_i \}$, representing units of measurement w_i for the i-th variate, i=1, 2, ...p. A change of variables is given by the quadratic form, $\mathbf{V}^* = \mathbf{WVW}^T$, representing an arbitrary scale change on each variate in matrix notation. The Jacobian, **J**, is a diagonal matrix of order ½p(p+1), and moreover, $|\mathbf{J}|=|\mathbf{W}|^{p+1}$. Therefore, scale invariance corresponds to finding k such that, $\pi_2(\mathbf{V}, k) = \pi_2(\mathbf{WVW}^T, k)\times|\mathbf{J}|$. This reduces to solving k where $|\mathbf{W}|^{-2k}\times|\mathbf{W}|^{p+1}=1$, and therefore k=½(p+1). This demonstrates that the integration described in Section 2 is not needed for finding a scale invariant prior, with $|\mathbf{J}|$ available.

Multivariate models may have several random effects, beyond the random residuals. This permits multiple variance matrices, just as multiple variance components were introduced in Section 2. For the sake of illustration, consider the introduction of two p×p variance matrices, **G** and **R**. Each of these matrices can be reparameterized by the scale transformations: $\mathbf{G}^* = \mathbf{WGW}^T$ and $\mathbf{R}^* = \mathbf{WRW}^T$, again with reference to the diagonal matrix **W** representing p units of measure. The Jacobian matrix doubles in order, and its determinant is given by $|\mathbf{J}|=|\mathbf{W}|^{2p+2}$. The naive next step is to double the order of matrix **V** to contain both **G** and **R**, and to define a 2p×2p matrix **U** as indicated below.

$$\mathbf{V} = \begin{bmatrix} \mathbf{G} & \\ & \mathbf{R} \end{bmatrix} \qquad \mathbf{U} = \begin{bmatrix} \mathbf{W} & \\ & \mathbf{W} \end{bmatrix}$$

Finally, solve for k where $\pi_2(\mathbf{V}, k) = \pi_2(\mathbf{UVU}^T, k)\times|\mathbf{W}|^{2p+2}$. This reduces to $|\mathbf{W}|^{-4k}\times|\mathbf{W}|^{2p+2}=1$, with k=½(p+1). This value of k is unchanged from what was found before, despite the fact that **V** is now a 2p×2p matrix.

The naive approach can be criticized for the same reason that π_1 was criticized in Section 2, because it may be too improper and might even make the posterior distribution improper. Fortunately, there is now an immediate fix for this short coming, and no integration is needed. Note that variance-covariance matrices can add into a

[4] This prior is defined as the determinant of the Fisher information matrix raised to the power -½, see Press (2005, Section 3.6.2).

total variance-covariances matrix: $\mathbf{V}=\mathbf{G}+\mathbf{R}$ and $\mathbf{WVW}^T = \mathbf{W}[\mathbf{G}+\mathbf{R}]\mathbf{W}^T = \mathbf{WGW}^T + \mathbf{WRW}^T$. Now the matrix $\mathbf{V}$ does not double in order, while the determinant of the Jacobian is still $|\mathbf{J}|=|\mathbf{W}|^{2p+2}$. Lastly, solve for k in $\pi_2(\mathbf{G}+\mathbf{R}, k) = \pi_2(\mathbf{W}[\mathbf{G}+\mathbf{R}]\mathbf{W}^T, k)\times|\mathbf{W}|^{2p+2}$. This reduces to $|\mathbf{W}|^{-2k}\times|\mathbf{W}|^{2p+2}=1$, and k=p+1 which is now doubled. The sought scale invariant prior for two variance matrices is given by the following.

$$\boldsymbol{\pi}_3 \propto \frac{1}{|\mathbf{G}+\mathbf{R}|^{p+1}} \qquad (3)$$

Prior (3) generalizes when there is a total variance matrix $\mathbf{T}$ that can be partitioned into r variance matrices that sum to $\mathbf{T}$:

$$\boldsymbol{\pi}_3 \propto \frac{1}{|\mathbf{T}|^{\frac{r}{2}(p+1)}}$$

As a check, the prior (3) is derived a second time below using the method based on the integration from Section 2. Start with prior given by Geisser and Cornfield (1963), and represent this prior as a marginal distribution that can be recovered by integrating out one of the variance matrices from the joint prior that is sought. Restrict the joint prior, π_4, to assign probability measure based on the total variance-covariance matrix $\mathbf{T}=\mathbf{G}+\mathbf{R}$. Change the variables of integration from $\mathbf{G}$ and $\mathbf{R}$ to $\mathbf{T}$ and $\mathbf{R}$, noting that the determinant of the Jacobian is 1. Define the matrix inequality $\mathbf{T}>\mathbf{R}$ to mean that for all $\mathbf{x}\in R^p$, $\mathbf{x}^T(\mathbf{T}-\mathbf{R})\mathbf{x}>0$. Then $\mathbf{R}$ varies between $\mathbf{0}$ (the null matrix) and $\mathbf{T}$. The matrix $\mathbf{T}$ varies between $\mathbf{0}$ and "infinity," but this last integration is never employed. The first integration and the mathematical assignments are represented below.

$$\int_{\mathbf{0}}^{\mathbf{T}} \boldsymbol{\pi}_4(\mathbf{T})d\mathbf{R} = \boldsymbol{\pi}_4(\mathbf{T})\int_{\mathbf{0}}^{\mathbf{T}} d\mathbf{R} \propto \frac{1}{|\mathbf{T}|^{\frac{p+1}{2}}} \qquad (4)$$

The integral is actually a multiple integral, and would be very difficult to evaluate in part because a better parameterization is needed to map out the range of integration and this comes with a determinant of the Jacobian that can be very complicated. However, there is a transformation of $\mathbf{R}$ that leads to a simplification $\mathbf{H}=\mathbf{L}^{-1}\mathbf{R}\mathbf{L}^{-1\,T}$, where $\mathbf{L}$ is the Cholesky decomposition of $\mathbf{T}$, i.e., $\mathbf{L}\mathbf{L}^T=\mathbf{T}$. Now $\mathbf{H}$ varies between $\mathbf{0}$ and $\mathbf{I}$, and the Jacobian reveals[5] that $d\mathbf{R}=|\mathbf{T}|^{½(p+1)}d\mathbf{H}$. Therefore, the integral becomes the following.

[5]This derivation depends on the Vech operator, and the key result listed by Harville (1997, bottom of page 366).

$$\boldsymbol{\pi}_4(\mathbf{T})\int_{\mathbf{0}}^{\mathbf{T}} d\mathbf{R} = \boldsymbol{\pi}_4(\mathbf{T})|\mathbf{T}|^{\frac{p+1}{2}}\int_{\mathbf{0}}^{\mathbf{I}} d\mathbf{H}$$

Now the integral on the right can be ignored because it is a constant. Combining this result with the right side of (4) shows that $\pi_4(\mathbf{T}) \propto |\mathbf{T}|^{-p-1}$, and this agrees with (3).

4. Conclusion

When there are multiple scale parameters it was found that the scale invariant prior is not uniquely defined. Far from being a negative result, this observation can be turned into something useful by constructing new scale invariant priors from multiplicative combinations of such priors. All is permitted as long as the Jacobian cancels in the construction. This basically moves from particular priors (2) and (3) that were found to be barely improper to new constructions that are now more improper. For example, as an alternative to (2) the following asymmetrical prior might find utility:

$$\boldsymbol{\pi}_5\left(x^2, y^2, z^2\right) \propto \frac{1}{x^2\left(x^2+y^2\right)\left(x^2+y^2+z^2\right)}$$

In some applications involving hierarchical models, the natural partitioning of what's known about variance might better justify the use of π_5 rather than (2). Nevertheless, be warned not to make the prior too improper (Hobert and Casella, 1996).

Alternatively, maybe (2) and (3) are preferred among all scale invariant alternatives because these priors are the most symmetrical and assign probability measure the most democratically through the total variance (or the total variance-covariance matrix), and are the least improper? Some criterion is needed to brake the stalemate, otherwise different units of measurement can have a non-relative impact on information content when scale invariance is abandoned. It is certainly true that when total variance can be partitioned into individual variance components then an arbitrary change in the standard of measurement impacts all the components in the same way, and that comes close to settling the issue because probability measure must be assigned consistently with that fact.

References

Berger, J.O., 1980, *Statistical Decision Theory: Foundations, Concepts, and Methods*. Sringer-Verlag.

Daniels, M., 1999, A Prior for Variance in Hierarchical Models. *Canadian Journal of Statistics*, 27, 3, 567-578.

Geisser, S., and J. Cornfield, 1963, Posterior Distributions for Multivariate Normal Parameters. *Journal of the Royal Statistical Society, Serried B*, 25, 368-376.

Gelfand, A.E., S.E. Hills, A. Racine-Poon, A.F.M. Smith, 1990, Illustration of Bayesian Inference in Normal Data Models using Gibbs Sampling. *Journal of the American Statistical Association*, 85, 412, 972-985.

Harville, D.A., 1997, *Matrix Algebra From A Statistician's Perspective*, Springer.

Hobert, J.P., and G. Casella, 1996, The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models. *Journal of the American Statistical Association*, 91. 436, 1461-1473.

Press, S.J., 2003, *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*, 2nd Edition. John Wiley and Sons.

Press, S.J., 2005, *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference, 2nd Edition. Dover Publications*, Inc.

Tanner, M.A., 1993, *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 2nd Edition. Springer-Verlag.

Wang, C.S., J.J. Rutledge and D. Gianola, 1993, Marginal Inferences about Variance Components in a Mixed Linear Model using Gibbs Sampling. *Genetics Selection Evolution*, 25, 1, 41-62.