The University of British Columbia
Department of Statistics
Technical Report #229

Correlation parameterization in random function models to improve normal approximation of the likelihood or posterior

Béla Nagy, Jason L. Loeppky, and William J. Welch

Department of Statistics
The University of British Columbia
333-6356 Agricultural Road
Vancouver, BC, V6T 1Z2, Canada

June 14, 2007

Abstract

Transformations can help small sample likelihood/Bayesian inference by improving the approximate normality of the likelihood/posterior. In this article we investigate when one can expect an improvement for a one-dimensional random function (Gaussian process) model. The log transformation of the range parameter is compared with an alternative (the logexp) for the family of Power Exponential correlations. Formulas are developed for measuring non-normality based on Sprott (1973). The effect of transformations on non-normality is evaluated analytically and by simulations. Results show that, on average, the log transformation improves approximate normality for the Squared Exponential (Gaussian) correlation function, but this is not always the case for the other members of the Power Exponential family.

1 Introduction

Asymptotic normality results for likelihood and Bayesian inference are considered highly valuable and a wide range of statistical models have been thoroughly explored from this perspective. Small sample normality, on the other hand, has received relatively less attention. But in practice, small sample results are often more relevant than large sample results.

This is especially true for the random function models used in computer experiments (Sacks, Welch, Mitchell and Wynn 1989), where sample sizes are routinely small relative to the dimensionality of the problem because of the excessive computational cost of obtaining data. Hence, the lack of small sample focus is even more puzzling in this research area. Theory is lagging behind current practice, since from the practitioners' point of view the crucial question is how to make the most of a limited number of data points. But theoretical arguments are usually based on asymptotics (Stein 1999; Zhang and Zimmerman 2005), providing little guidance for small samples.

The goal of this paper is to begin addressing this gap by a systematic exploration of one of the simplest cases when essentially everything boils down to just one parameter. This is only the first step toward exploring higher-dimensional problems and most of our results are only applicable to the one-dimensional case (except Section 4). The inspiration for this work came from the thesis of Karuri (2005), who observed that the log transformation improved posterior approximate normality for one- and two-dimensional examples and demonstrated its usefulness for integration and prediction. Our focus is on the likelihood, which can also be interpreted as an unnormalized posterior for a uniform prior.

The main contribution of this paper is the application of the theory in Sprott (1973) to quantify the effect of transformations on approximate normality of the likelihood/posterior for small sample sizes without resorting to asymptotics. The main finding is that in the one-dimensional case with the Squared Exponential correlation, the previously noticed usefulness of the log transformation for some data sets by Karuri (2005) holds in general for the class of data described by the model: on average, the log transformation improves approximate small sample normality of the likelihood/posterior.

This is useful for both likelihood and Bayesian inference. Likelihood inference often uses the observed or expected Fisher information as a measure of standard error; hence, its validity depends entirely on the approximate normality of the likelihood. Bayesian inference is based on the posterior distribution. If that can be approximated well by a normal distribution, that can greatly simplify the implementation, as shown in Nagy, Loeppky and Welch (2007). The reader is referred to that paper for understanding how the model can be used for prediction in the context of computer experiments. Here we just briefly outline the prediction formulas for the model to give some intuition for the roles of the different model parameters.

We consider a simple one-dimensional special case of the statistical formulation in Sacks et al. (1989) to model a deterministic response Y as a function of some variable x:

$$Y(x) \;=\; Z(x),$$

where x is real and $Z(x)$ is a real-valued random function (Gaussian stochastic process) with

$$\mathrm{E}(Z(x)) \;=\; 0 \quad \text{and} \quad \mathrm{Cov}(Z(w),\, Z(x)) \;=\; \sigma^2\, R(w,\, x),$$

where σ^2 is the process variance and $R(w,\, x)$ is the correlation. Our procedure can be applied to any single parameter correlation function, as long as it is three times differentiable with respect to the correlation parameter, denoted θ. In this paper, we use the Power Exponential family of correlations:

$$R(w,\, x) \;=\; \exp\left\{-\theta|w-x|^p\right\}, \tag{1}$$

where $\theta \in (0,\ \infty)$ and $p \in (0,\ 2]$. For simplicity, we treat p as fixed and known, leaving the range parameter θ the only unknown in the correlation function. Of special interest to us is the Squared Exponential (Gaussian) correlation function with $p = 2$ that is used to model smooth functions.

The likelihood is a function of the two unknowns:

$$L(\sigma^2,\ \theta)\ \propto\ \frac{1}{(\sigma^2)^{\frac{n}{2}}\ |R|^{\frac{1}{2}}}\ \exp\left\{-\frac{y^T R^{-1} y}{2\sigma^2}\right\}, \tag{2}$$

where n is the sample size, y is the data, and R is the $n \times n$ correlation matrix that is a function of θ.

The primary use of the model is to predict the response at a new, untried x_0. If in addition to p, θ and σ^2 are also known, then the Best Linear Unbiased Predictor (BLUP) is given by

$$\hat{y}_0(\theta)\ =\ r(x_0)^T R^{-1} y,$$

with Mean Squared Error

$$\text{MSE}_{\hat{y}_0}(\sigma^2,\ \theta)\ =\ \sigma^2\ \left(1\ -\ r(x_0)^T R^{-1} r(x_0)\right),$$

where $r(x_0)$ is a vector of correlations between the new x_0 and the original design points (that is a function of the range parameter θ). Thus θ exerts its influence on the BLUP and its Mean Squared Error through the correlation vector $r(x_0)$ and the correlation matrix R.

In contrast, the dependence on σ^2 is much simpler. It is a factor in the MSE formula, but the BLUP itself is independent of σ^2. This has important implications when the parameters are unknown. It is easier to deal with uncertainty in σ^2 than in θ because the predictor is not affected by σ^2 and the MSE is simply proportional to σ^2.

This suggests a convenient simplification: if we could eliminate σ^2 analytically, then we could focus on studying the dependence on θ, which is the "interesting" variable in these models. Fortunately, this is easy to do by either "profiling" or "integrating", as shown in Section 4. Either way, the result is the one-parameter likelihood function $L(\theta)$, that is used for subsequent calculations. Having eliminated σ^2, we can call $L(\theta)$ the "profile likelihood" or the "integrated likelihood" or just simply the "likelihood" for short, and the log of this function the "log-likelihood": $l(\theta) = \log L(\theta)$.

This paper is about the shape of $l(\theta)$ and $L(\theta)$. Specifically, we are interested in quantifying how well $l(\theta)$ can be approximated by a quadratic function, which is equivalent to measuring how close $L(\theta)$ is to a normal density function (up to a scale). We also want to know whether a transformation of the parameter θ can bring the likelihood closer to normality.

For example, Figure 1 illustrates how much the log transformation can improve approximate normality, especially when the sample size is small ($n = 3$ in this case). On the original scale (left), the contour plot of the two-parameter likelihood (2) is highly non-normal, having a banana-shaped peak around the Maximum Likelihood Estimate (MLE) and a sharp ridge along the axes, marked by the dashed line. Below the contour plot, the one-parameter version of this dashed line is also highly non-normal. This is the profile likelihood $L(\theta)$ that can be obtained by maximizing (2) over all σ^2 given θ. We can see that the normal approximation of $L(\theta)$ based on the mode at the MLE of θ (dotted curve) is a poor approximation of the profile likelihood (dashed curve).

On the log scale (right), the contours are more ellipsoidal, suggesting less non-normality. Below that, the difference is even more striking for the profile likelihood (dashed) that is virtually

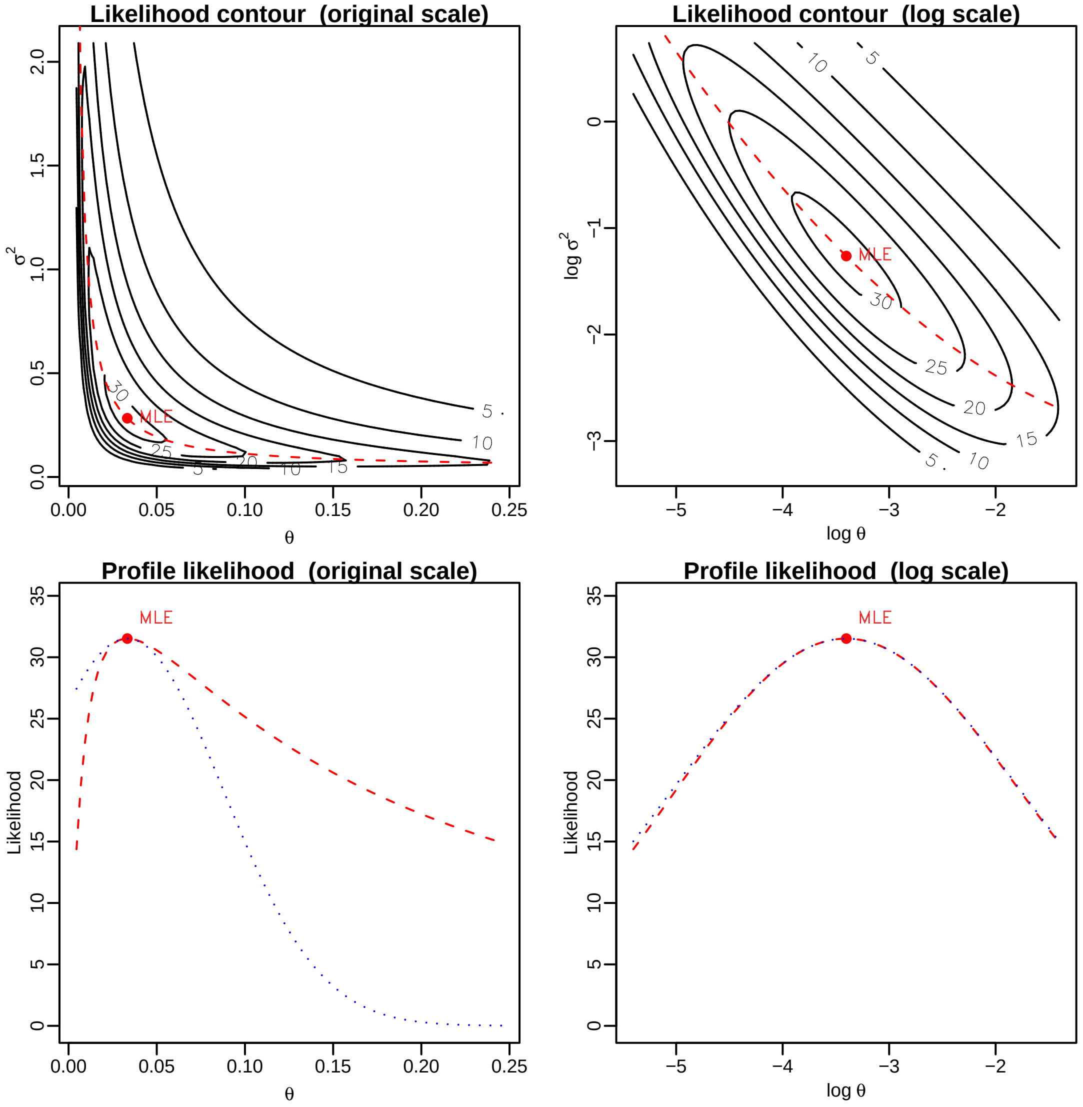

Figure 1: The log transformation improved approximate normality of the likelihood for this example with $p = 2$ and only three data points ($n = 3$). The top two plots are for the two-parameter likelihood (2) and the bottom two for the one-parameter profile likelihood (with σ^2 eliminated). The ridges of the contours are marked by the dashed lines, reaching their apex at the MLE. Below the contour plots, these dashed lines are plotted as functions of the range parameter, representing the profile likelihood function (that is the likelihood maximized over all σ^2 given θ). In addition to the profile likelihoods (dashed curves), their normal approximation is also shown for comparison (dotted curves). These are unnormalized normal density functions centered on the MLE of the range parameter with variance set to the negative inverse of the second derivative of the profile likelihood at the MLE. On the log scale, the profile likelihood (dashed) and its normal approximation (dotted) are so close that the difference is difficult to notice.

indistinguishable from its normal approximation (dotted) over the domain of $\log\theta$ shown (corresponding to the domain of θ on the left). At the first look it may not be apparent that there are two separate lines in this plot (one dashed and one dotted) that overlap almost perfectly.

This example is for the $p = 2$ special case when the log transformation is expected to reduce non-normality. This is demonstrated both empirically in Section 2 and theoretically in Section 3. (Other values of p between 0 and 2 are also explored in Section 3, but the results are less clear-cut).

The log transformation is an intuitively natural choice for mapping a positive parameter to the whole real line when small values not too far from zero predominate. An extra benefit for the Power Exponential family is that it can unify the treatment of slightly different forms: although we use form (1) exclusively, our results readily translate to other forms, such as

$$\exp\left\{-(\theta|w-x|)^p\right\} \quad \text{or} \quad \exp\left\{-\left(\frac{|w-x|}{\theta}\right)^p\right\}$$

since the only difference on the log scale is just a constant scaling factor of p or $-p$ that does not affect normality/non-normality.

As an alternative to the log, we also explore the logexp transformation that is defined as $\log(\exp(\theta) - 1)$. This is inspired by the $\rho = e^{-\theta}$ parameterization used in computer experiments, for example Linkletter, Bingham, Hengartner, Higdon and Ye (2006). One way to transform $\rho \in (0,\ 1)$ to facilitate approximate normality is by using the logit function to map it to the real line. That leads to the logexp transformation for the original range parameter θ:

$$\text{logit}(1-\rho) = \log\left(\frac{1-\rho}{\rho}\right) = \log\left(\frac{1}{\rho} - 1\right) = \log(\exp(\theta) - 1).$$

The following two sections describe two measures of non-normality based on Sprott (1973). Section 2 includes a simulation study for $p = 2$. Section 3 also explores three other choices in addition to the $p = 2$ case: $p = 0.5,\ 1$, and 1.5. Since p is given, θ is the only unknown parameter in the model to transform. The process variance σ^2 is treated as a nuisance parameter and two options are presented for its elimination in Section 4. Finally, Section 5 outlines upcoming follow-up work and future research.

2 Observed Non-Normality

Let $l(\theta)$ denote the (one-parameter) log-likelihood, $\hat{\theta}$ the Maximum Likelihood Estimate (MLE) of θ, $l'''(\hat{\theta})$ the third-derivative of the log-likelihood at the MLE, and $-l''(\hat{\theta})$ the "observed" information. Then the Observed Non-Normality (ONN) measure (Sprott 1973) is defined as follows:

$$\text{ONN for } \theta \;=\; |\, l'''(\hat{\theta})\, (-l''(\hat{\theta}))^{-\frac{3}{2}}\, |.$$

If the likelihood $L(\theta)$ is proportional to a normal density function, then $l(\theta)$ is quadratic and $l'''(\hat{\theta})$ is zero, making the ONN zero as well. Otherwise, the magnitude of the third derivative (standardized by the second) measures the deviation from normality.

Although Sprott (1973) originally proposed the measures for likelihoods, the extension to posteriors is immediate by employing a uniform prior: $l(\theta)$ becomes the log-posterior (up to an additive constant) and $\hat{\theta}$ becomes the Maximum Posterior Likelihood Estimate (MPLE) or maximum

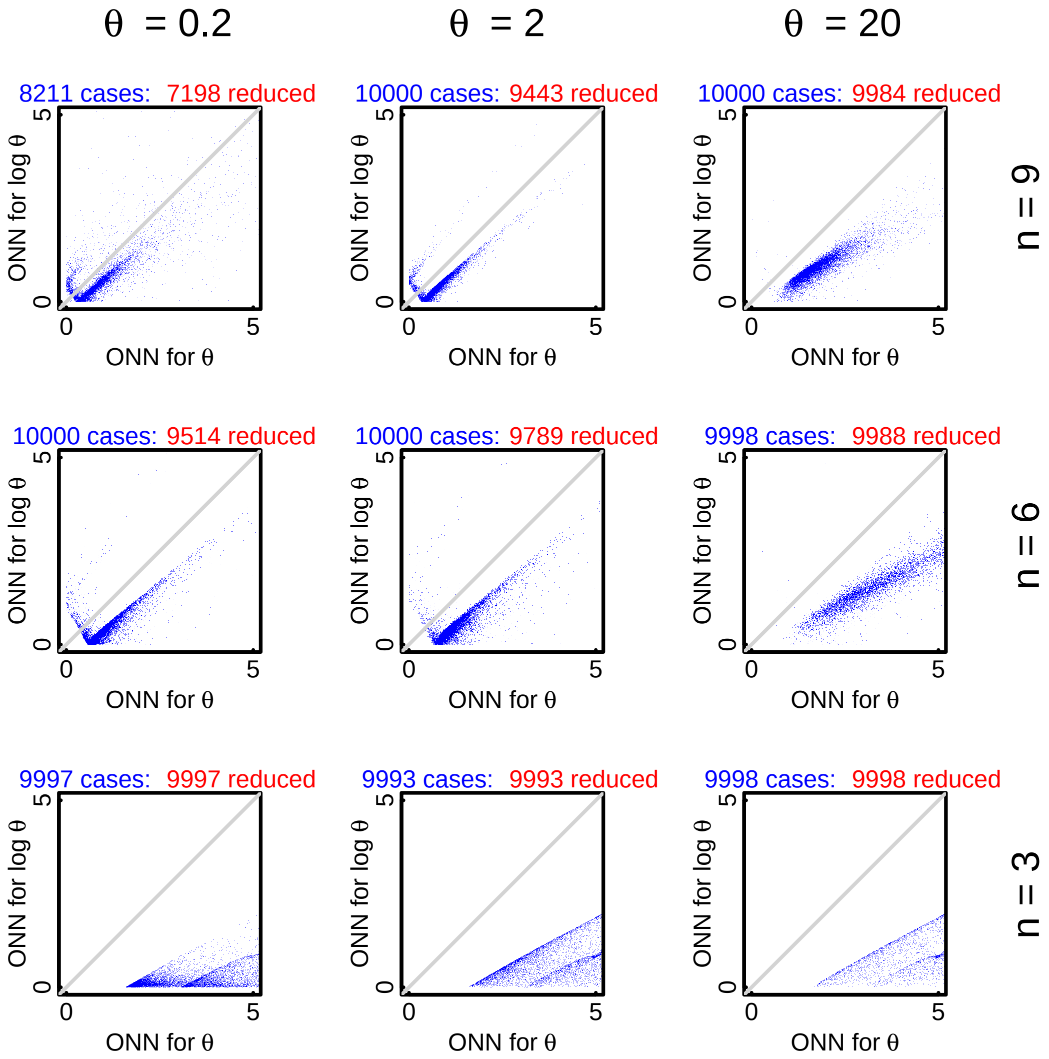

Figure 2: The log transformation reduced Observed Non-Normality in most cases for $p = 2$ for all nine combinations of θ and n. The horizontal and vertical coordinates are the ONN on the original scale and on the log scale, respectively. Each individual dot represents a successful realization. The total number of successful realizations are indicated on the top of each plot, followed by how many of them had their ONN reduced by the log transformation. The gray diagonal is the dividing line between those cases that had their ONN reduced and those that did not. Most are under the diagonal, which means reduced Observed Non-Normality: ONN for $\log \theta$ < ONN for θ. (It is evident that $n = 9$ data points can already be too many for $\theta = 0.2$, since out of the 10,000 simulated data sets, computations succeeded in only 8,211 cases. It is well-known that there is a certain limit on the sample size when $p = 2$, especially for such a small θ. Failure rates were much lower for the other combinations of θ and n). The design was equally spaced on the $[0,\ 1]$ interval: $\{\, i/(n-1) \mid i = 0,\ \ldots,\ n-1 \,\}$.

a posteriori (MAP). A Bayesian interpretation is provided in Section 4, after discussing how to eliminate σ^2 to get a likelihood/posterior that is a function of only the range parameter.

A simulation study was conducted to evaluate the effect of the log transformation on non-normality for $p = 2$ under a wide variety of settings. Figure 2 shows how many times non-normality was reduced by the log transformation for $10,000$ simulated data sets from the random function model (realizations of the Gaussian process) for nine possible combinations of the true θ and the sample size n. The layout of the figures follows the 3×3 factorial design for the three levels of the range parameter from an extremely high correlation ($\theta = 0.2$) to an extremely low ($\theta = 20$) and for the three levels of the sample size from small ($n = 3$) to large ($n = 9$).

For all nine combinations, Observed Non-Normality was reduced in most cases, since most realizations lie below the diagonal, where the ONN on the log scale (vertical axis) is less than the ONN on the original scale (horizontal axis). The counts of reduced ONN (out of the total number of cases that could be computed for that particular combination of θ and n) are shown on the top of each plot. This suggests that the log transformation of θ is much more likely to decrease non-normality of the original $L(\theta)$ than to increase it, according to this measure. Furthermore, the smaller the sample size, the greater and the more likely the reduction.

Going back to Figure 1, now we can quantify the non-normality of the profile likelihood before and after the log transformation. Before, the ONN for θ is 4.97, but after, the ONN for $\log \theta$ is only 0.04. This is consistent with the reductions seen for $n = 3$ in Figure 2.

For this simulation, two ONNs were computed separately for each data set (before and after the log transformation). However, this is not necessary, since each can be obtained from the other. This relationship is explained in the next section and then used for Sprott's second measure that is more convenient than the ONN, because it can be calculated analytically without any simulations of actual data sets.

3 Expected Non-Normality

Sprott's second measure is the Expected Non-Normality (ENN) that follows from the first by replacing the third and second derivatives with their expectations, so that $El'''(\hat{\theta})$ is standardized by the "expected" Fisher information:

$$\text{ENN for } \theta \;=\; |\, El'''(\hat{\theta})\,(-El''(\hat{\theta}))^{-\frac{3}{2}}\,|.$$

This measure is more appropriate when one wishes to consider a family of possible likelihoods without conditioning on any particular data set. Hence, the rest of our results are based on the ENN instead of the ONN. Sprott (1973) also provided a formula that quantifies the effect of a transformation ϕ on non-normality, where ϕ is a twice differentiable function of θ. After applying the ϕ transformation, the ENN becomes

$$\text{ENN for } \phi(\theta) \;=\; \left| El'''(\hat{\theta})\,(-El''(\hat{\theta}))^{-\frac{3}{2}} \;+\; \frac{3\,\phi''(\hat{\theta})}{\phi'(\hat{\theta})\,(-El''(\hat{\theta}))^{\frac{1}{2}}} \right|,$$

where the first term inside the absolute value is the same as before in the definition of the ENN for θ and the second term is the effect of the transformation ϕ. An analogous relationship holds for the ONN (without the expectations). Note that the presence of the second derivative in the numerator implies invariance with respect to linear transformations.

Given a function ϕ and a value for $\hat{\theta}$, this formula enables one to visualize the difference in the Expected Non-Normality due to the transformation ϕ. Comparisons are shown in Figures 3 and 4 for the log and the logexp, respectively. The same equally spaced design was used again, identical to the one used for the simulations. Also, the sample sizes were 3, 6, and 9, as before, to facilitate comparisons. An extra large sample size $n = 12$ was also added.

The simplest of these plots is the log transformation for $p = 2$ in Figure 3. For each of the four sample sizes, there is a line segment representing the ENN for $\hat{\theta} \in [0.2,\ 20]$ (i.e. over the same range that was used for the true θ in the simulation study). This shows increasing ENN as a function of $\hat{\theta}$ both on the original scale (horizontally) and on the log scale (vertically). For $n = 12$, we can see a short fat segment close to the origin, indicating relatively low ENN on both scales. As the sample size falls, we can observe a shift to the right (growing ENN on the original scale) and also increasing segment length (growing ENN on the log scale), so that the end of the thinnest segment for $n = 3$ is off the plot for large values of $\hat{\theta}$ close to 20.

What is unique about this plot, compared to all the others in Figures 3 and 4, is that it shows a clear advantage of the log transformation with respect to the ENN measure: the ENN for $\log \theta$ is always less than the ENN for θ for all $\hat{\theta}$ values computed between 0.2 and 20. Moreover, the difference is always substantial, since none of the lines come close to the gray diagonal in the middle that marks the line where the ENN is the same on both axes. This is what makes the log transformation special for the $p = 2$ case. One way to interpret this result in words is to say that the log transformation is “expected” to reduce non-normality for $p = 2$. This is consistent with the results of the simulation study in the previous section that used the ONN. Figure 2 for the ONN and the plot for $p = 2$ in Figure 3 for the ENN can be compared directly since the scales (from 0 to 5) are the same on both axes.

Continuing the $p = 2$ case, Figure 4 suggests that the logexp is also expected to reduce non-normality, but not as much as the log. The major disadvantage of the logexp is that as $\hat{\theta}$ increases, the curves approach the diagonal, meaning vanishing reductions. Figure 5 makes the comparison between the log and the logexp more explicit by plotting one directly against the other. The $p = 2$ case is again very clear: the ENN for the log is less than the ENN for the logexp, since the area above the diagonal is never breached.

In summary, the ENN-based analysis of the $p = 2$ case shows that both transformations are expected to reduce non-normality and that the expected reductions of the log are greater than that of the logexp. However, these simple conclusions cannot be extended to other values of p between 0 and 2. To illustrate, Figures 3, 4, and 5 also have plots for $p = 0.5$, 1, and 1.5. Looking at these cases in each figure leads to the following observations:

- Figure 3: The log transformation is expected to reduce non-normality for small sample sizes. However, this is not necessarily true for large samples, as shown by the $n = 12$ curve breaching the area above the diagonal for $p = 0.5$, 1, and 1.5.
- Figure 4: The logexp transformation is expected to reduce non-normality with no exceptions. But reductions are negligible for large $\hat{\theta}$ when the curves lie close to the diagonal.
- Figure 5: For large $\hat{\theta}$, the expected reductions of the log are greater than that of the logexp. However, for small $\hat{\theta}$, the logexp can achieve smaller ENN than the log, as indicated by portions of the curves above the diagonal for $p \neq 2$.

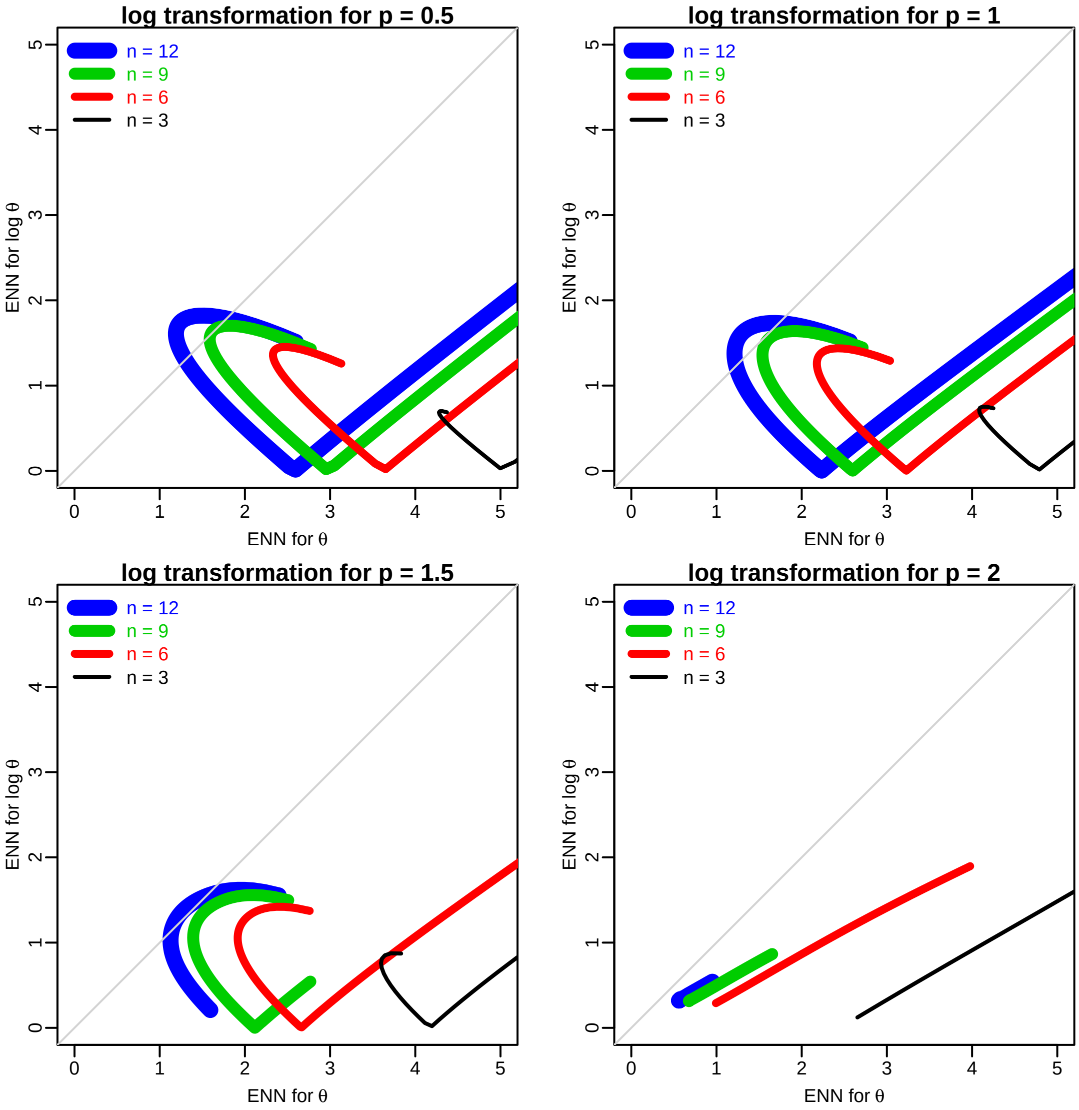

Figure 3: The effect of the log transformation on Expected Non-Normality: The horizontal and vertical coordinates are the ENN on the original scale and on the log scale, respectively. The gray diagonal is the dividing line between those cases that had their ENN reduced by the log transformation and those that did not. For each of the four sample sizes $n = 3, 6, 9, 12$, both ENN measures were calculated for selected $\hat{\theta}$ values between 0.2 and 20 (some of which could not be computed because of numerical issues for small $\hat{\theta}$ and large n). The resulting curves are plotted with their thickness proportional to the sample size. Expected Non-Normality was reduced by the log transformation for $p = 2$. However, for $p = 0.5$, 1, and 1.5, the thickest curve for $n = 12$ crosses over the diagonal for some $\hat{\theta}$ values somewhere between 0.2 and 20 (which means that for those cases, the ENN for $\log\theta$ became greater than the original ENN for θ).

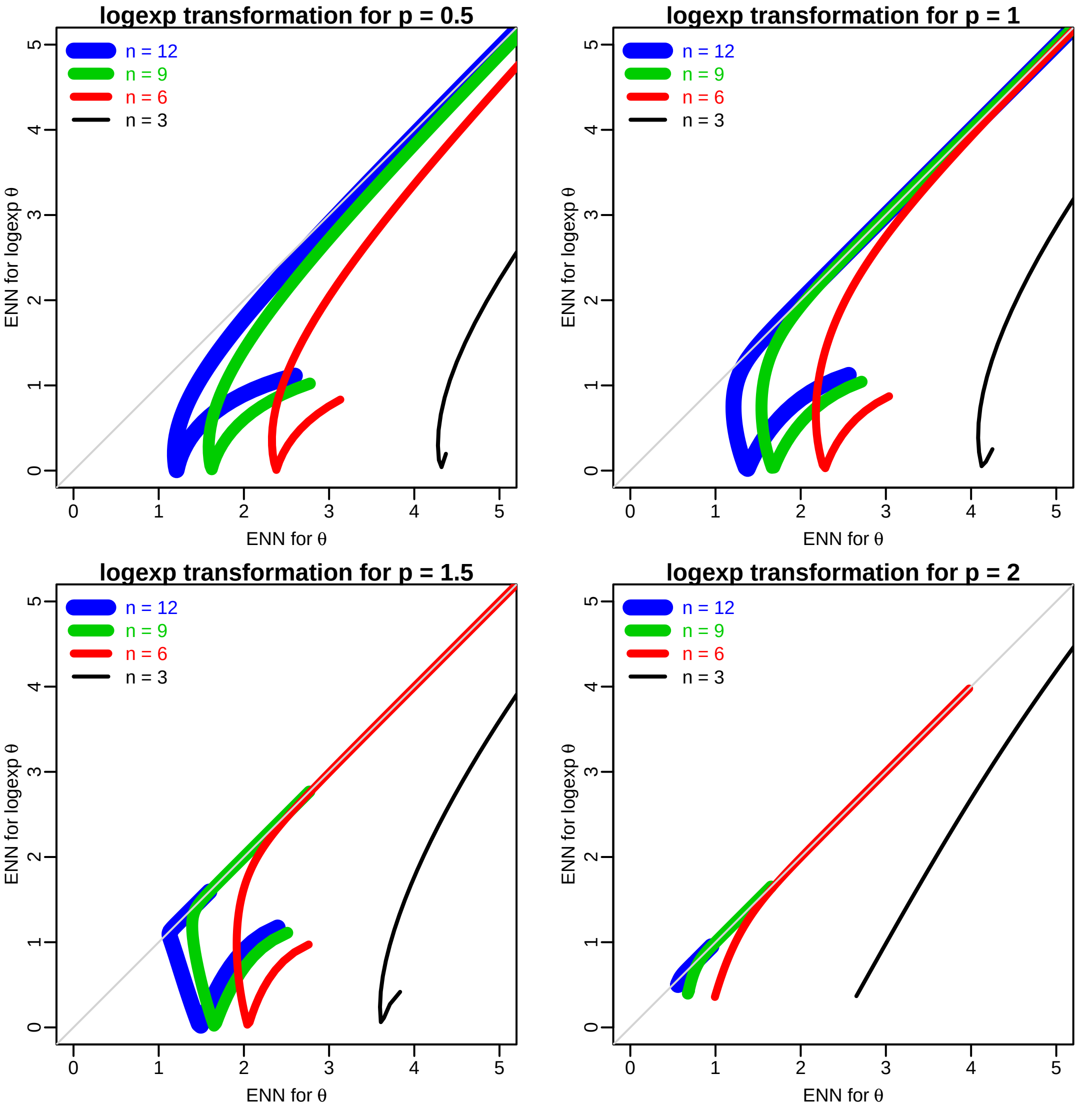

Figure 4: The effect of the logexp transformation on Expected Non-Normality: The horizontal and vertical coordinates are the ENN on the original scale and on the logexp scale, respectively. For each of the four sample sizes $n = 3,\ 6,\ 9,\ 12$, both ENN measures were calculated for selected $\hat{\theta}$ values between 0.2 and 20 (some of which could not be computed because of numerical issues for small $\hat{\theta}$ and large n). The resulting curves are plotted with their thickness proportional to the sample size. The area above the gray diagonal is completely empty for all four values of p, which means that the logexp transformation helps normality. However, as $\hat{\theta}$ increases, all curves approach the diagonal, which means that differences quickly become negligible and for large $\hat{\theta}$ the ENN for logexp $\theta \approx$ ENN for θ.

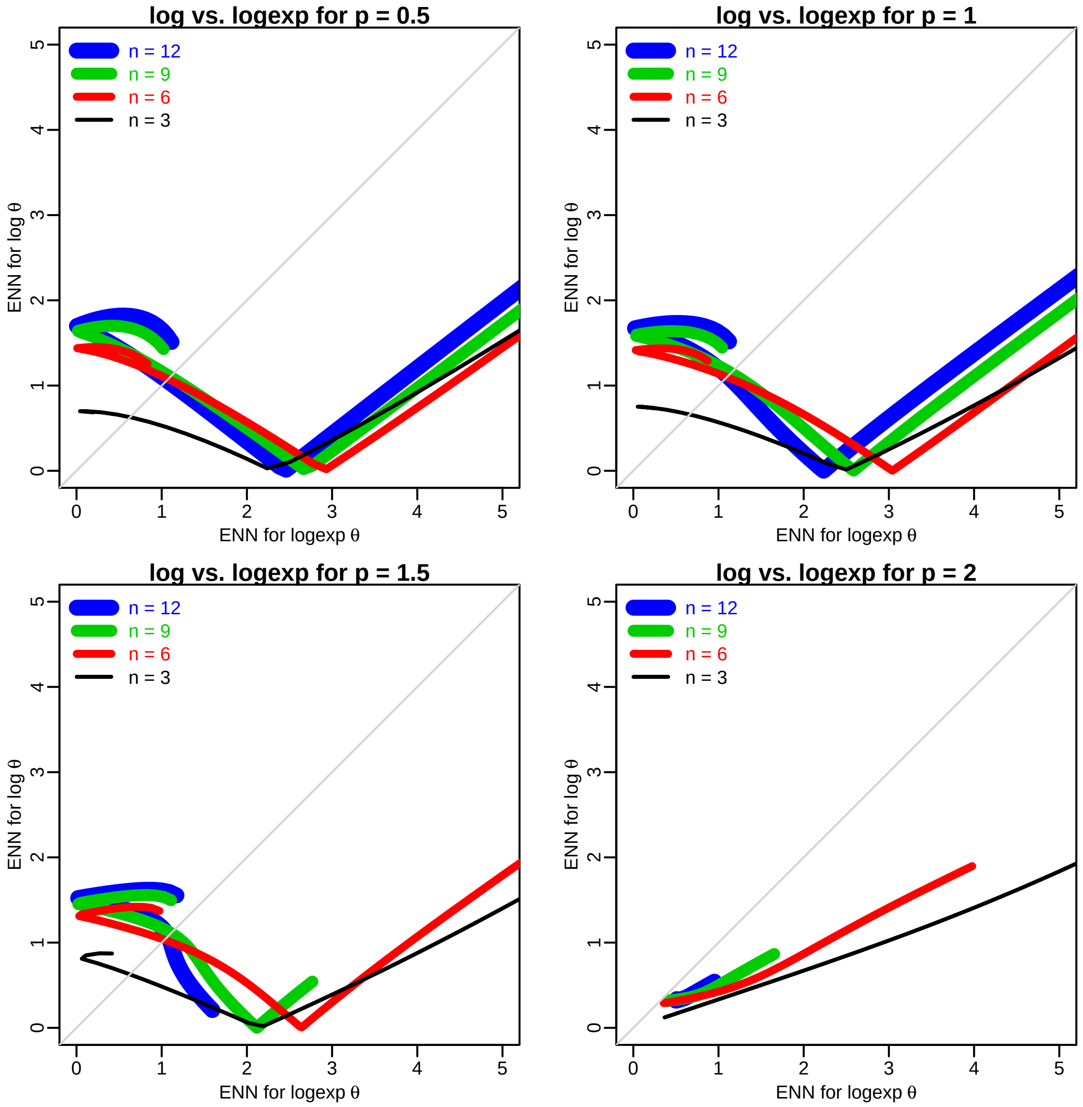

Figure 5: The Expected Non-Normality of the log vs. the logexp transformation: The horizontal and vertical coordinates are the ENN on the logexp scale and on the log scale, respectively. For each of the four sample sizes $n = 3, 6, 9, 12$, both ENN measures were calculated for selected $\hat{\theta}$ values between 0.2 and 20 (some of which could not be computed because of numerical issues for small $\hat{\theta}$ and large n). The resulting curves are plotted with their thickness proportional to the sample size. Expected Non-Normality was reduced more by the log transformation than the logexp in all cases for $p = 2$. But for $p = 0.5$, 1, and 1.5, curves starting above the diagonal at $\hat{\theta} = 0.2$ indicate that the logexp leads to lower ENN than the log for small $\hat{\theta}$ up to some threshold where the curves cross over the diagonal and the relationship reverses and then the logexp leads to higher ENN than the log for all $\hat{\theta}$ above that threshold.

4 Eliminating the process variance σ^2

This section formally defines the one-parameter likelihood function $L(\theta)$. Here we can be more general than the rest of the paper, since the only assumption needed about θ is that it is a parameter of the correlation matrix R. (For example, this enables generalizations to higher dimensions when θ is a vector).

Two possible ways are presented to deal with the nuisance parameter σ^2: "maximizing out" to get the profile likelihood and "integrating out" to get the integrated likelihood (see Berger, Liseo and Wolpert (1999) for a general discussion of these methods). While profiling is common in likelihood-based settings, Bayesians are usually more comfortable with integrating. Although in this case the same $L(\theta)$ function is obtained both ways, interpretations can still differ depending on the underlying framework.

4.1 Profile likelihood

For a fixed θ, $L(\sigma^2,\, \theta)$ in equation (2) has a unique maximum at

$$\hat{\sigma}^2(\theta) \;=\; \frac{y^T R^{-1} y}{n}.$$

This is easily obtained by differentiating $L(\sigma^2,\, \theta)$ with respect to σ^2 or by observing that

$$\sigma^2 \mid \theta,\, y \;\sim\; IG\left(\frac{n}{2} - 1,\, \frac{y^T R^{-1} y}{2}\right)$$

and using the $\beta/(\alpha+1)$ formula for the mode of an Inverse Gamma distribution $IG(\alpha,\, \beta)$ with density function

$$f(\, x \mid \alpha,\, \beta\,) \;=\; \frac{\beta^\alpha\, \exp\left\{-\frac{\beta}{x}\right\}}{\Gamma(\alpha)\, x^{\alpha+1}}.$$

Plugging in $\hat{\sigma}^2(\theta)$ into (2) yields the profile likelihood:

$$L(\theta) \;=\; L(\hat{\sigma}^2(\theta),\, \theta) \;\propto\; \frac{1}{(\hat{\sigma}^2(\theta))^{\frac{n}{2}}\, |R|^{\frac{1}{2}}}\; \exp\left\{-\frac{y^T R^{-1} y}{2\hat{\sigma}^2(\theta)}\right\} \;\propto\; (y^T R^{-1} y)^{-\frac{n}{2}}\, |R|^{-\frac{1}{2}}.$$

Now the maximum likelihood estimation can be done using $L(\theta)$ instead of the original $L(\sigma^2,\, \theta)$, reducing the dimensionality of the required numerical optimization by one.

4.2 Integrated likelihood

Bayesians prefer to put a prior distribution on σ^2 before eliminating it. According to Berger, De Oliveira and Sansó (2001), the most common choice is that of Handcock and Stein (1993), who used the improper prior $1/\sigma^2$ for $\sigma^2 > 0$. This can be interpreted as a relative weight function giving prior weights inversely proportional to the magnitude, encouraging σ^2 to be close to zero. (On the log scale, this becomes the uniform prior for $\log \sigma^2$).

The $1/\sigma^2$ prior can also be used as a joint prior for all the model parameters, by putting a uniform prior on some parameterization of θ. Note that the only role of θ in this section is that the correlation matrix R depends on it. Hence, all arguments in this section are independent of the actual correlation structure encapsulated in R and the actual parameterization represented by (the possibly multivariate) θ, including any transformations of θ.

For example, in the one-dimensional case using the original parameterization for a positive θ, the $1/\sigma^2$ joint prior is obtained by multiplying together the $1/\sigma^2$ prior for σ^2 and a constant 1 prior for θ (assuming independence of σ^2 and θ). Since both priors are improper, their product is improper, too. A well-known disadvantage of such simple priors is that they can lead to improper posteriors (Berger et al. 2001).

We update the prior by multiplying with the likelihood (2) to get the posterior:

$$\frac{1}{\sigma^2}\ L(\sigma^2,\,\theta)\ \propto\ \frac{1}{(\sigma^2)^{\frac{n}{2}+1}\,|R|^{\frac{1}{2}}}\ \exp\left\{-\frac{y^TR^{-1}y}{2\sigma^2}\right\}$$

and notice that

$$\sigma^2\mid\theta,\,y\ \sim\ IG\left(\frac{n}{2},\,\frac{y^TR^{-1}y}{2}\right)$$

which means that σ^2 can be integrated out from the posterior to get the marginal posterior of θ:

$$\int_0^\infty \frac{1}{(\sigma^2)^{\frac{n}{2}+1}\,|R|^{\frac{1}{2}}}\ \exp\left\{-\frac{y^TR^{-1}y}{2\sigma^2}\right\}\,d\sigma^2\ =\ \frac{\Gamma\left(\frac{n}{2}\right)}{\left(\frac{y^TR^{-1}y}{2}\right)^{\frac{n}{2}}\ |R|^{\frac{1}{2}}}\ \propto\ (y^TR^{-1}y)^{-\frac{n}{2}}\,|R|^{-\frac{1}{2}}.$$

Berger et al. (2001) refer to this as the integrated likelihood. Note that up to a multiplicative constant, this is the same as the profile likelihood function $L(\theta)$ above. This is an interesting property of this model; in general, the two different procedures lead to different results.

5 Discussion

The main result of the paper is that on average, the log transformation improves approximate normality for the Squared Exponential (Gaussian) correlation function, but this is not necessarily the case for other members of the Power Exponential family. We have also provided general procedures for measuring non-normality and for evaluating the effect of transformations on non-normality in the one-dimensional case. The only requirement was the differentiability of the correlations and the transformations.

Finding the optimal (or even a satisfactory) transformation for normality for a particular covariance structure is still an open problem. We are working on approximate methods inspired by the “vanishing third derivative” of Anscombe (1964) that is optimal for the criteria of Sprott (1973).

From a practical standpoint, one of the crucial research questions is how this will scale up to higher dimensions. In a two-dimensional setting, Karuri (2005) presented examples of the log transformation reducing skewness and making posteriors more ellipsoidal. We are currently investigating a multivariate generalization of Sprott’s measures by Kass and Slate (1994), using the same profile/integrated likelihood as in the previous section (that has the same form, independently of the dimensionality of θ).

It is important to reiterate that these are small sample approximate normality results, not large sample or asymptotic results. The distinction is essential, since that is exactly what makes them relevant in practice. For example, typically, we have only a limited number of runs from a computationally expensive computer model (it may take days or even weeks to obtain each data point). On the other hand, even if we had large samples, they would often turn out to be uncomputable because of the ill-conditioning of the correlation matrix (especially for the Squared Exponential).

But that does not mean that we cannot make use of asymptotic methods for inference, since validity depends on normality, not on the sample size. For example, Wald confidence intervals are based on a quadratic approximation of the profile log-likelihood. Hence, the more a transformation reduces non-normality, the more accurate the Wald approximation becomes. A nearly-quadratic log-likelihood can also make the numerical optimization easier. For example, Newton's method and related quasi-Newton algorithms work best on surfaces that are well approximated by a quadratic.

Markov chain Monte Carlo (MCMC) methods are also helped by approximately normal posteriors. Better yet, if we can replace the posterior with its normal approximation, then we can sample from that directly (avoiding MCMC). This Fast Bayesian Inference (FBI) is the subject of Nagy et al. (2007), demonstrating the usefulness of the log transformation for achieving accurate prediction uncertainty assessments.

Acknowledgments

This research was funded by the Natural Sciences and Engineering Research Council of Canada and the National Program on Complex Data Structures of Canada. Computations were made using WestGrid, which is funded in part by the Canada Foundation for Innovation, Alberta Innovation and Science, BC Advanced Education, and the participating research institutions. WestGrid equipment is provided by IBM, Hewlett Packard and SGI.

References

ANSCOMBE, F. J. (1964). Normal likelihood functions. *Annals of the Institute of Statistical Mathematics*, **16** 1–19.

BERGER, J. O., DE OLIVEIRA, V. and SANSÓ, B. (2001). Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, **96** 1361–1374.

BERGER, J. O., LISEO, B. and WOLPERT, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, **14** 1–28.

CONNIFFE, D. and SPENCER, J. E. (2001). When moments of ratios are ratios of moments. *Journal of the Royal Statistical Society, Series D: The Statistician*, **50** 161–168.

GEARY, R. C. (1933). A general expression for the moments of certain symmetrical functions of normal samples. *Biometrika*, **25** 184–186.

HANDCOCK, M. S. and STEIN, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, **35** 403–410.

KARURI, S. W. (2005). *Integration in Computer Experiments and Bayesian Analysis*. Ph.D. thesis, University of Waterloo.

KASS, R. E. and SLATE, E. H. (1994). Some diagnostics of maximum likelihood and posterior nonnormality. *The Annals of Statistics*, **22** 668–695.

LINKLETTER, C., BINGHAM, D., HENGARTNER, N., HIGDON, D. and YE, K. Q. (2006). Variable selection for Gaussian process models in computer experiments. *Technometrics*, **48** 478–490.

NAGY, B., LOEPPKY, J. L. and WELCH, W. J. (2007). Fast Bayesian Inference for Gaussian Process Models. Tech. Rep. 230, Department of Statistics, The University of British Columbia. URL `http://www.stat.ubc.ca/Research/TechReports/techreports/230.pdf`.

SACKS, J., WELCH, W. J., MITCHELL, T. J. and WYNN, H. P. (1989). Design and analysis of computer experiments (C/R: P423-435). *Statistical Science*, **4** 409–423.

SPROTT, D. A. (1973). Normal likelihoods and their relation to large sample theory of estimation. *Biometrika*, **60** 457–465.

STEIN, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag Inc.

ZHANG, H. and ZIMMERMAN, D. L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika*, **92** 921–936.

Appendix

This Appendix contains all formulas necessary to compute the ONN and the ENN. After introducing some notation and definitions, first, second, and third derivatives are calculated for relevant functions and matrices. Then two lemmas are given for expectations, followed by two theorems for the ONN and the ENN.

Notation and Definitions

Let y denote the response vector having length n, mean zero, and covariance matrix $\sigma^2 R$, where σ^2 is the process variance and R is the symmetric, positive definite $n \times n$ design correlation matrix (that is a function of the parameter θ). Let G denote the inverse matrix of R and define the matrices $F = GR'$, $S = GR''$, and $T = GR'''$, where R', R'', and R''' are the first, second, and third derivatives of R, respectively (with respect to θ). The trace of a matrix is denoted by $tr(\cdot)$. For concise notation, we also define $t(\cdot) = tr(\cdot)/n$.

Taking the log of $L(\theta)$ in Section 4, the log-likelihood is (up to an additive constant):

$$l(\theta) = -\frac{n}{2}\log\frac{y^T R^{-1} y}{n} - \frac{1}{2}\log|R|.$$

The functions g and h are used to simplify calculations:

$$g(\theta) = \frac{y^T R^{-1} y}{n} \quad \text{and} \quad h(\theta) = -\frac{\log|R|}{n}.$$

Suppressing θ from $l(\theta)$, $g(\theta)$, and $h(\theta)$ gives the following equation for the log-likelihood:

$$l \;=\; \frac{n}{2}\,(h \;-\; \log g).$$

Formulas for Derivatives

Differentiating with respect to θ leads to the following formulas for the first, second, and third derivatives of the log-likelihood:

$$l' \;=\; \frac{n}{2}\left(h' \;-\; \frac{g'}{g}\right),$$

$$l'' \;=\; \frac{n}{2}\left(h'' \;+\; \left(\frac{g'}{g}\right)^2 \;-\; \frac{g''}{g}\right),$$

$$l''' \;=\; \frac{n}{2}\left(h''' \;-\; 2\,\left(\frac{g'}{g}\right)^3 \;+\; 3\,\frac{g'g''}{g^2} \;-\; \frac{g'''}{g}\right),$$

where $h' = -t(F), \quad h'' = t(F^2 - S), \quad h''' = -t(2F^3 - 3FS + T)$,

and $g' = y^T G' y/n,\; g'' = y^T G'' y/n,\; g''' = y^T G''' y/n$,

where $G' = -FG, \quad G'' = (2F^2 - S)G, \quad G''' = -(6F^3 - 3FS - 3SF + T)G$.

Lemmas for Expectations

Lemma 1. For any symmetric $n \times n$ matrix Q

$$E\, y^T Q y \;=\; \sigma^2\, tr(QR).$$

Proof: $E\, y^T Q y \,=\, E\, tr(y^T Q y) \,=\, E\, tr(Q y y^T) \,=\, tr(Q\, E\, y y^T) \,=\, tr(Q\, \sigma^2 R) \,=\, \sigma^2\, tr(QR)$, where we used the fact that y has mean zero and covariance matrix $\sigma^2 R$.

Lemma 2. For any symmetric $n \times n$ matrix Q

$$E\, \frac{y^T Q y}{y^T G y} \;=\; t(QR).$$

Proof: Let $z = C^{-1} y$ for $y \;\sim\; N(0,\; \sigma^2 R)$, where C is the lower-triangular Cholesky-factor of the covariance matrix $\sigma^2 R$. Then $z \;\sim\; N(0,\; I_n)$ and

$$CC^T \;=\; \sigma^2 R \quad\Rightarrow\quad R \;=\; CC^T/\sigma^2 \quad\Rightarrow\quad R^{-1} \;=\; \sigma^2 \big(C^T\big)^{-1} C^{-1}.$$

Substituting $y = Cz$ and $G = \sigma^2 \left(C^T\right)^{-1} C^{-1}$ we get:

$$E \, \frac{y^T Q y}{y^T G y} \;=\; E \, \frac{z^T C^T Q C z}{z^T C^T \sigma^2 (C^T)^{-1} C^{-1} C z} \;=\; \frac{1}{\sigma^2} \, E \, \frac{z^T (C^T Q C) z}{z^T z}.$$

Conniffe and Spencer (2001) state that the expectation of a ratio of this form is the ratio of the expectations for any quadratic form in the numerator. This is a consequence of the fact that the ratio is independent of its denominator, a result attributed to Geary (1933). Hence we can apply Lemma 1 separately to the numerator and the denominator:

$$E \, \frac{y^T Q y}{y^T G y} \;=\; \frac{E \, y^T Q y}{E \, y^T G y} \;=\; \frac{\sigma^2 \, tr(QR)}{\sigma^2 \, tr(GR)} \;=\; \frac{tr(QR)}{tr(I_n)} \;=\; \frac{tr(QR)}{n} \;=\; t(QR).$$

Theorems for the ONN and the ENN

Theorem 1. $l''(\hat{\theta})$ and $l'''(\hat{\theta})$ for the Observed Non-Normality measure are:

$$l'' \;=\; \frac{n}{2} \left(h'' \,+\, (h')^2 \,-\, \frac{g''}{g} \right) \qquad \text{and} \qquad l''' \;=\; \frac{n}{2} \left(h''' \,-\, 2\,(h')^3 \,+\, 3\,h' \, \frac{g''}{g} \,-\, \frac{g'''}{g} \right).$$

Proof: When $\theta = \hat{\theta}$ (the MLE of θ), then $l' = 0$ and that implies that $g'/g = h'$. Replacing g'/g with h' in the second and third derivative formulas for l gives the result.

Theorem 2. $El''(\hat{\theta})$ and $El'''(\hat{\theta})$ for the Expected Non-Normality measure are:

$$El'' = \frac{n}{2} \left(t^2(F) \,-\, t(F^2) \right) \text{ and } El''' = \frac{n}{2} \left(2t^3(F) \,-\, 6t(F)t(F^2) \,+\, 3t(F)t(S) \,-\, 3t(FS) \,+\, 4t(F^3) \right).$$

Proof: By the results of the previous theorem, the expectations are:

$$El'' \;=\; \frac{n}{2} \left(h'' \,+\, (h')^2 \,-\, E \, \frac{g''}{g} \right) \quad \text{and} \quad El''' \;=\; \frac{n}{2} \left(h''' \,-\, 2\,(h')^3 \,+\, 3\,h' \, E \, \frac{g''}{g} \,-\, E \, \frac{g'''}{g} \right).$$

Now Lemma 2 can be applied to the expectations of the ratios:

$$E \, \frac{g''}{g} \;=\; E \, \frac{y^T G'' y}{y^T G y} \;=\; t(G''R) \qquad \text{and} \qquad E \, \frac{g'''}{g} \;=\; E \, \frac{y^T G''' y}{y^T G y} \;=\; t(G'''R).$$

Substituting the formulas for G'', G''', and h', h'', h''' completes the proof.