Supplementary Appendices to: "Size-correction and principal components for interspecific comparative studies."

Appendix S1

Simulation code: Phylogenetic size-correction simulations and analysis written for MATLAB. Function calls simdata( ), below, and phyl_resid( ), provided in the Appendix to the main text.

```
% function [npB pB]=resid_analysis(filename,nC) reads C matrices from
% 'filename,' simulates data with expected residual covariance matrix V,
% performs phylogenetic and non-phylogenetic size correction, and computes
% residual betas (in npB & pB).  Repeats nC times.
function [npB pB]=resid_analysis(filename,V,nC)

    % set global (n=num taxa)
    n=100;

    % open file with Cs
    fid=fopen(filename,'r');

    % set type I errors / power to 0.0
    p_err=0.0; np_err=0.0;

    for i=1:nC

        % get C
        C=fscanf(fid,'%f',[n n]);

        % simulate data for x (size) and Y
        [x,Y]=simdata(V,C);

        % size-correct (phylogenetically)
        R=phyl_resid(C,x,Y);

        % put residuals in new X and y
        X(:,1)=ones(n,1);
        X(:,2)=R(:,1);
        y=R(:,2);

        % calculate beta (phylogenetically)
        pB(i,:)=((X'*C^-1*X)^-1*X'*C^-1*y)';

        % now estimate the SE of B
        s2=(1/(n-2))*(y-X*pB(i,:)')'*C^-1*(y-X*pB(i,:)');
        sB=s2*(X'*C^-1*X)^-1;
        if(((pB(i,2)-1.96*sqrt(sB(2,2)))>0)||((pB(i,2)+1.96*sqrt(sB(2,2)))<0))
            p_err=p_err+1/nC;
        end

        % or
```

```
        % set identity matrix
        I=diag(ones(n,1),0);

        % size-correct (non-phylogenetically)
        R=phyl_resid(I,x,Y);

        % put residuals in new X and y
        X(:,1)=ones(n,1);
        X(:,2)=R(:,1);
        y=R(:,2);

        % calculate beta
        npB(i,:)=((X'*C^-1*X)^-1*X'*C^-1*y)';

        % now estimate the SE of B
        s2=(1/(n-2))*(y-X*npB(i,:)')'*C^-1*(y-X*npB(i,:)');
        sB=s2*(X'*C^-1*X)^-1;
        if(((npB(i,2)-…1.96*sqrt(sB(2,2)))>0)||((npB(i,2)+1.96*sqrt(sB(2,2)))<0))
            np_err=np_err+1/nC;
        end

    end

    fprintf('Type I error (phylogenetic) = %f\n',p_err);
    fprintf('Type I error (non-phylogenetic) = %f\n',np_err);

    fclose(fid);

% done

% function [x,Y]=simdata(V,C) simulates three characters on a phylogenetic
% tree, given by the matrix C. x is size. The other two, in Y, are each
% correlated with size r=0.95, and residually correlated with each other
% according the covariance matrix V.
function [x,Y]=simdata(V,C)

    % set global variables
    n=max(size(C));

    % set R1 (residual covariance matrix)
    R1=V;

    % create the first two characters, non phylogenetically
    Y=randn(n,2)*chol(R1);

    % now create non-phylogenetic size
    s=randn(n,1);

    % make each of Y correlated with size
    R2=[1.0 0.95; 0.95 1.0];
    temp=[s Y(:,1)]*chol(R2); Y(:,1)=temp(:,2);
    temp=[s Y(:,2)]*chol(R2); Y(:,2)=temp(:,2);

    % make each phylogenetic
    x=(s'*chol(C))'; Y=(Y'*chol(C))';

% done
```

Appendix S2

Simulation code: Phylogenetic principal components simulations and analysis written for MATLAB. Function calls sorteig( ) and phyl_pca( ), both provided in the Appendix to the main text.

```
% function [r_vp,r_dp,r_vnp,r_dnp]=pca_analysis(filename,nC) reads
% 'filename' containing nC tree matrices, generates random data for m=4
% characters, performs phylogenetic & non-phylogenetic PCA, and returns
% four vectors containing the correlations of estimates with generating
% eigenvectors (r_vp, rvnp) and the correlations of estimates with
% generating eigenvalues (r_dp, r_dnp).
function [r_vp,r_dp,r_vnp,r_dnp]=pca_analysis(filename,nC)

    fid=fopen(filename,'r');

    % set number of traits, taxa
    m=4; n=100;

    for i=1:nC

        % get C
        C=fscanf(fid,'%f',[n n]);

        % random covariance structure
        D=diag([1 0.5 0.25 0.125]); V=orth(randn(m));
        [V,D]=sorteig(V,D); R=V*D*V^-1;

        % generate non-phylogenetic data
        X=randn(n,m)*chol(R);

        % transform into the phylogenetic space
        X=(X'*chol(C))';

        % perform phylogenetic PCA
        [Sp,Evalp,Evecp,Lp]=phyl_pca(C,X);

        % calculate the correlations between generating and estimated eigenvectors
        r_vp(i)=mean(diag(abs(V'*Evecp)));

        % calculate the correlation between the generating and estimated
        % eigenvalues
        r_dp(i)=...
         (diag(D)./sqrt(sum(diag(D).^2)))'*(diag(Evalp)./sqrt(sum(diag(Evalp).^2)));

        % generate identity matrix, I
        I=diag(ones(n,1),0);

        % perform non-phylogenetic PCA
        [Snp,Evalnp,Evecnp,Lnp]=phyl_pca(I,X);

        % calculate the correlations between generating and estimated eigenvectors
        r_vnp(i)=mean(diag(abs(V'*Evecnp)));

        % calculate the correlation between the generating and estimated
        % eigenvalues
        r_dnp(i)=...
         (diag(D)./sqrt(sum(diag(D).^2)))'*(diag(Evalnp)./sqrt(sum(diag(Evalnp).^2)));
```

```
    end
    fclose(fid);
% done
```

Appendix S3

Given data generated for size (x) and a size-dependent morphological character (y), in which the residual variance-covariance matrix for y is a direct (but not precisely known) function of the phylogeny, we might use Pagel's (1999) λ, which transforms the off-diagonals of the matrix **C**, defined in the text. I'll call this matrix $\mathbf{C}_\lambda$. Note that this is just one of many possible evolutionary models.

According to our model, in which y is a function of x with residual error structure proportional to $\mathbf{C}_\lambda$, **y** is an $n \times 1$ vector distributed as $MVN(\mathbf{X}\beta, \sigma^2\mathbf{C}_\lambda)$. The likelihood of **β**, σ^2, and λ, given **X** (a matrix containing a vector of 1.0s and size) and **y** is thus given by the multivariate normal equation:

$$L(\beta, \sigma^2, \lambda) = \frac{\exp[-(\mathbf{y} - \mathbf{X}\beta)'(\sigma^2\mathbf{C}_\lambda)^{-1}(\mathbf{y} - \mathbf{X}\beta)/2]}{(2\pi)^{n/2}\left|\sigma^2\mathbf{C}_\lambda\right|^{1/2}}, \quad \text{(S3.1)}$$

or, alternatively:

$$\log(L) = -(\mathbf{y} - \mathbf{X}\beta)'(\sigma^2\mathbf{C}_\lambda)^{-1}(\mathbf{y} - \mathbf{X}\beta)/2 - n \cdot \log(2\pi)/2 - \log(|\sigma^2\mathbf{C}_\lambda|)/2. \quad \text{(S3.2)}$$

It should be noted that this is not the same equation as the likelihood equation for multivariate λ provided by Freckleton et al. (2002; equation 5). This is because in this case we are maximizing the likelihood of our phylogenetic model only for the residual error in y (Rohlf 2001; Rencher and Schaalje 2008).

We can use analytic solutions derived elsewhere for the maximum likelihood values of **β** and σ^2, given λ (Rencher and Schaalje 2008). These are as follows:

$$\mathbf{b} = (\mathbf{X}'\mathbf{C}_\lambda^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}_\lambda^{-1}\mathbf{y}, \text{ and} \quad \text{(S3.3)}$$

$$\sigma^2 = (\mathbf{y} - \mathbf{X}\mathbf{b})'\mathbf{C}_\lambda^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b})/n$$

where **b** represents our estimate of **β**. We must estimate λ numerically.

Once we have obtained **b**, we simply calculate the residuals as in the main text, as follows:

$$\mathbf{r} = \mathbf{y} - \mathbf{Xb} . \quad \text{(S3.4)}$$

Figure S3.1A shows a tree and data for x (size) and y generated using $\lambda = 0.5$, whereas Figure S3.1B shows the corresponding likelihood surface for λ. In this case the maximum likelihood estimate for λ is $\hat{\lambda} = 0.477$.

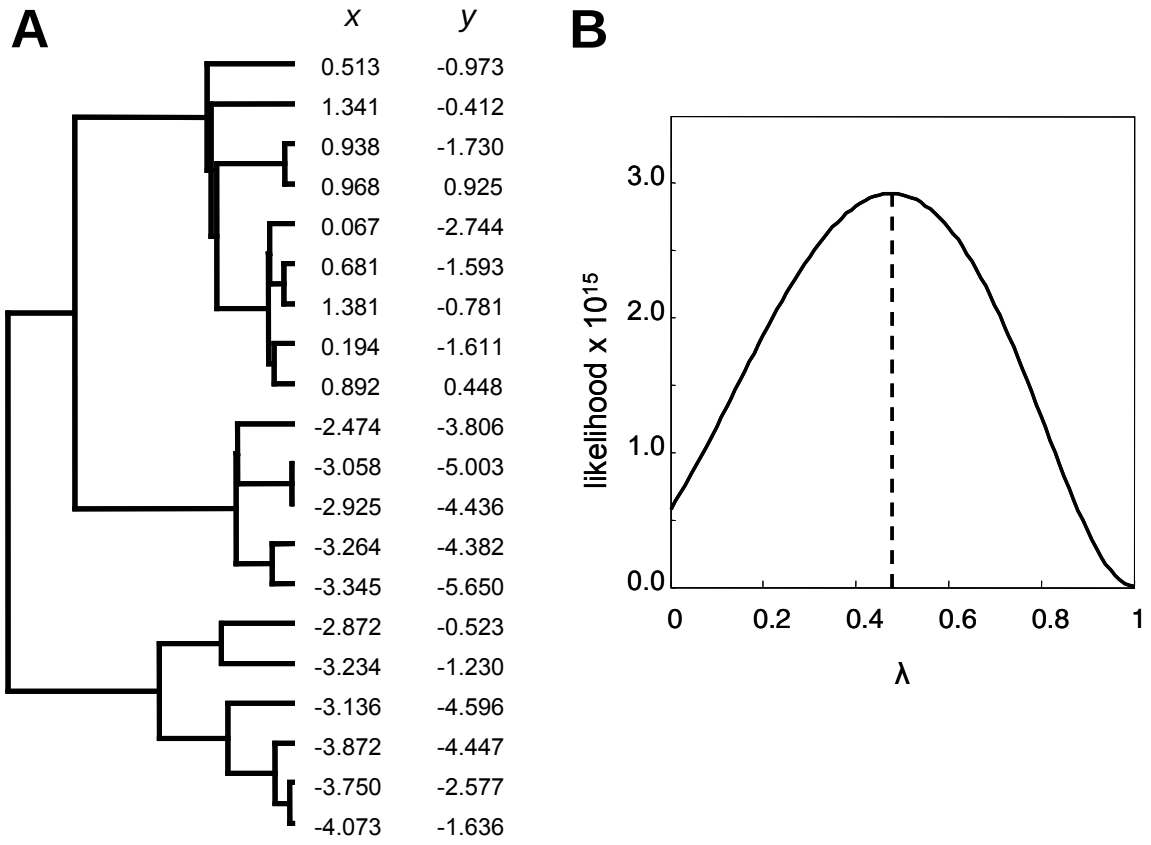

Figure S3.1. A) A stochastic, 20 taxon phylogeny with data generated for size (x) and a size-correlated morphological character (y) under the generating conditions of $\beta = 0.75$, and $\lambda = 0.50$. B) Likelihood surface for λ obtained as described in the text.

Literature Cited in the Appendices

Freckleton, R. P., P. H. Harvey, and M. Pagel. 2002. Phylogenetic analysis and comparative data: A test and review of evidence. Am. Nat. 160:712-726.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. Nature 401:877-884.

Rencher, A. C., and G. B. Schaalje. 2008. Linear Models in Statistics, Second Edition. John Wiley & Sons, Hoboken, NJ.

Rohlf, F. J. 2001. Comparative methods for the analysis of continuous variables: Geometric interpretations. Evolution 55:2143-2160.